XI Conference –European Congress of Methodology

Tuesday 22 July 2025 - Friday 25 July 2025

EAM2025

Book of Abstracts

Contents

1 Wednesday, 23 July 2025

1.1 Session 10 : ”Statistical Applications in Social Sciences and Sports”

Title

Measuring Distance to University in Germany: How Accurate is the Straight-Line Approach?

Author(s)
Daniel Hein [1]

[1] German Centre for Higher Education Research and Science Studies

Abstract
Many studies in the field of higher education use distance as a simple measure of accessibility, commuting, or moving. They often define distance as the straight-line distance, while only a few studies measure distance using the actual travel distance. This approach is supposedly more sophisticated, accurate, and realistic. Our aim is to assess whether the straight-line is an adequate proxy for travel distances by car and multimodal public transportation in Germany.
We compare the straight-line and the travel distances between the former school and the current university. We also distinguish between the shortest and the best route. The linear relation between the straight-line and travel distance was analyzed using ordinary least-squares regression. To examine outliers, the difference between the actual travel distance and the predicted travel distance, which is the straight-line distance multiplied by the regression slope, was used. The straight-line distance is a good proxy when the absolute difference between the actual travel distance and the predicted travel distance is less than 5 km, or the relative difference is less than 10 %. The results are based on a representative sample of 2,903 different routes taken by German students.
In 96 % of the cases, the straight-line distance is an adequate proxy for the shortest travel distance by car. However, the straight-line is a good approximation of the best car route 80 % of the time. For the shortest and best public transportation routes, the straight-line is a reliable proxy 66 % and 60 % of the time, respectively. The largest discrepancies occur in areas with physical obstacles such as lakes, rivers, mountains or wilderness and nature conservation areas. These findings suggest that future studies should use travel distances for more realistic results, as they provide significantly greater accuracy than straight-line distances.

Title

Analysis of long-term and feedback effects in individual physical activity patterns

Author(s)
Marek Brabec [1]

[1] Institute of Computer Science

Abstract
This methodological study has been conducted as part of the „Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583"co-financed by the European Union and is based on analyzing large-scale ecological momentary assessment data of daily step counts from the „Healthy Aging in Industrial Environment HAIE CZ.02.1.01/0.0/0.0/16_019/0000798 "project. It starts with traditional mixed effects linear hierarchical Markovian (AR-type) model allowing for the separation of between- and within-subject relationships to important covariates (such as sex, age, exercise identity, education, SES level) and their interactions. We show that the traditional additive modeling is not sufficient and there are important interactions with place of residence. Then, we carefully explore seasonality of the physical activity at two scales (weekly and annual) proving their significance and quantifying their relative contributions. We demonstrate that corrections for seasonality are important in that if unadjusted, they distort systematically individual dynamics and give false impression of external controllability of the activity (e.g. by intervention) where important parts are pre-determined by largely unchangeable seasonal patterns. Further insight is based on modeling autoregressive component in a nonlinear way (accounting for local saturation effects connected with over- or under-exercising that cannot be captured by linear models) utilizing the flexible generalized additive modeling approach. Since the long-term effects (related e.g. to fatigue after cumulative over-exercising spanning several days) in the AR dynamical structure are notoriously difficult to model, we approach the problem from the complexity-penalizing viewpoint and regularize the higher lag coefficient behavior based on flexible (soft) constraints. This allows us to demonstrate long-term dynamic details that operate above the seasonal behavior and hence are controllable in principle. Since our model is hierarchical with individual-specific random effects, we are able to separate and quantify several sources of inter-individual variability and show that specific parts of inter-individual structural variability in the physical activity dynamics is an important feature to consider when designing new interventions programs.

Title

Labour market participation and Unpaid Care Work in the UK: An Intersectional Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA)

Author(s)
Christie Butcher [1]

[1] The University of Sheffield

Abstract

**Context: The number of unpaid carers has risen in recent years, with a growing proportion providing high-intensity care. Balancing caregiving with employment can create significant challenges, such as financial strain, poor mental and physical well-being and reduced labour market participation. This study explores how caregiving intensity and socio-demographic characteristics combine to shape labour market outcomes, taking an intersectional approach. The ways in which the caring-employment relationship varies by socio-demographic characteristics is yet to be explored.

**Aims: We address the following research questions: RQ 1 What is the relationship between caring intensity and labour market participation? RQ 2 How does the relationship between care intensity and labour market participation vary across intersectional strata? RQ 3 To what extent is the variation in the care-employment relationship (identified in RQ2) explained by two-way, care-by-demographic-variable interactions, as opposed to higher-order interactions? Methods: Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) Regression models with random slopes are used to analyse wave 14 (2022-2023) of the Understanding Society dataset.

**Key findings: Caring is linked to reduced levels of labour market participation, even after accounting for socio-demographic variables. This association was most pronounced for people providing higher intensity care (20 hours or more). There is variation in the relationship between care intensity and labour market participation, however most of this variation is explained by the two-way, care-by-demographic-variable interactions: There is a larger gap in job hours between male non-carers and male high intensity carers compared to females, although females have lower job hours at the baseline. The gap in job hours is the largest between high intensity carers from generation X and Y and non-carers. High intensity carers with secondary school or other education work similar hours to lower intensity carers or non-carers who have no qualifications, and many fewer hours than non-carers with the same qualifications. There are fewer clear differences between ethnic groups.

**Limitations: To maintain large enough sample sizes, detail is lost in the re-categorising of care intensity into fewer categories, and the measurement of care intensity does not include detail about the individual experience of providing care.

**Conclusion: More support for carers should be put in place, such as respite care and improved access to Carers Allowance.

Title

Improving Scale Properties With a Bifactor Structure: Measuring Work as a Calling

Author(s)

Gladys Rolo-González [1] , Gabriel Muiños [2] , Ernesto Suárez [1] , Isabel Duarte-Lores [1]

[1] Universidad de La Laguna; [2] Rijksuniversiteit Groningen

Abstract

Calling has been central to psychological research in understanding what drives individuals to feel drawn to their profession. Experiencing a sense of calling has been associated with increased motivation, a stronger sense of purpose, and the perception of work as a means of contributing to others. Moreover, identifying a calling during professional training can facilitate specialization and enhance adaptability to workplace challenges. Research on calling has grown exponentially in both quantity and quality over the past decade. However, a major limitation in calling research is the lack of consensus regarding its definition and measurement. Divergent empirical approaches have hindered the development of a unified framework, limiting the construct's applicability. Reliable and valid instruments are crucial to advancing research, allowing for a deeper understanding of how calling evolves over time and its impact on career and well-being. One of the most widely used instruments in this field is the Calling and Vocation Questionnaire (CVQ), developed by Dik et al. (2012). The CVQ distinguishes three content-related factors—Transcendent Summons, Purposeful Work, and Prosocial Orientation—that define the essence of calling. Additionally, it includes two temporal dimensions—Presence and Search—that describe the stage of the process in which an individual is situated. The original structure of the CVQ proposes six first-order factors derived from combining the three content dimensions with the two temporal dimensions. However, this factorial structure has not been consistently replicated in subsequent studies, including with participants from non-Western countries, such as India, or in adapting the scale to other languages, such as Korean or Spanish—where low fit indices and lack of factor consistency have been reported. As an alternative, we propose a bifactor model in which each item is explained by combining a content factor and a temporal factor instead of being categorized into one of the six possible types. Therefore, the three content-related factors define the construct, while the two temporal dimensions represent different ways of experiencing calling.

This bifactor structure was tested through a confirmatory factor analysis in a sample of art majors (N = 427), yielding a good fit (CFI = .90, RMSEA = .059, χ^2/df = 2.302) and outperforming the original model.

Although the bifactor solution proposed in this study modifies the relation between the factors from the original model, it remains consistent with the theoretical framework of the construct. Moreover, it captures the interdependence of calling dimensions, facilitating construct interpretation and practical application. This structure allows for the independent analysis of subscales, providing greater flexibility depending on research needs.

Title

Q-Matrix Validation with Factor Retention Methods in Cognitive Diagnosis Modeling

Author(s)
Tugay Kaçak [1]

[1] Trakya University, Department of Educational Sciences

Abstract
In Cognitive Diagnosis Modeling (CDM), validating the Q-matrix is crucial to classify attribute profiles accurately. Several empirical and statistical methods have been developed to validate Q-matrix. However, most of these methods need a number of attributes to perform and begin the Q-matrix validation process. Yet, the studies which Q-matrix validation techniques to use estimate number of attributes are limited. With the aim of filling this gap, the current study evaluates factor retention methods to determine the number of attributes in CDMs by several Monte Carlo simulations. Data will be generated by varying sample size, number of attributes, test length, generating model and item quality. The evaluated factor retention methods will be three variants of parallel analysis (PCA, PAF, MRFA) with 95th eigenvalue criteria, Empirical Kaiser Criterion, Comparison Data Forest, Factor Forest, Next Eigenvalue Suffiency Test and exploratory graph analysis. Results will be assessed by accuracy and absolute bias, also doublet combinations of factor retention methods will be assesed by agreement hit rate to reveal “winner” combination instead of if there is not a “winner” single method.

Title

Longitudinal Trajectories of Presenteeism and Absenteeism: The Role of Trait Competitiveness in Early Career Researchers

Author(s)
Anja Isabel Morstatt , Daniel Spurk [1] , Simone Kauffeld [2] , Hannes Schilling [2] , Stephanie Hirschberger [2]

[1] University of Bern; [2] Technische Universität Braunschweig

Abstract

Presenteeism (working while ill) and absenteeism (not working due to illness) are well-documented as contributors to detrimental health outcomes, such as exhaustion (Demerouti et al., 2009; Komp et al., 2021), and are linked to productivity losses (Harrison & Martocchio, 1998; Johns, 2011). While prior research has extensively examined situational predictors of these behaviors, little is known about how they evolve over time (Ruhle et al., 2020). This study investigates differences in the trajectories of presenteeism and absenteeism among pre- and postdoctoral researchers in Germany and examines the role of trait competitiveness in shaping these trajectories.

We analyzed data from three waves of a longitudinal survey conducted between 2014 and 2018, with yearly intervals, involving German pre- and postdoctoral researchers (N = 334; nPostDocs = 211, nPreDocs = 123). To address challenges common in applied research, including missing data due to dropout, small sample sizes, and non-normal distributions, we employed a combination of advanced statistical techniques: Latent Growth Curve Modeling, Multiple Imputation, and Scale Transformation and Categorization.

Our results reveal that presenteeism increases significantly among postdoctoral researchers over the three-year period (ηSlope = 0.066, SE = 0.032, p = 0.038). Additionally, trait competitiveness is associated with lower absenteeism during the first wave, but only among predoctoral researchers (b = -0.044, SE = 0.022, p = 0.047). These findings highlight that the trajectories of presenteeism and absenteeism vary based on career stage (pre- vs. postdoctoral researchers) and that trait competitiveness can mitigate absenteeism in early career researchers. However, long-term health consequences for highly competitive early career researchers need to be investigated in future research.

The study findings underscore the importance of early interventions in practice to address presenteeism and absenteeism effectively.

Methodologically, we demonstrate how the combination of several analyses features can be used to leverage survey data. At the conference, we will discuss the limitations of the study, including potential biases related to, e.g., the measurement of presenteeism and absenteeism, and offer implications for future research and practice.

1.2 Session 2 : ”Innovative methods in Measurement and Evaluation”

Title

Investigating the potential of large language models to streamline psychometric test development

Author(s)
Meltem Ozcan [1] , Hok Chio (Mark) Lai [1]

[1] USC

Abstract
Despite extensive research, accessible resources, sophisticated tools, and clear guidelines for the development and use of psychological scales, researchers often bypass critical steps in this process—such as measurement invariance (MI) testing—due to the complexity and time demands of these procedures. Questionable measurement practices, such as failing to test for MI, modifying scales without proper justification, or constructing scales without conducting necessary psychometric evaluations, are commonplace in research and compromise the validity of inferences (Flake, Pek, & Hehman, 2017; Maassen et al., 2023), highlighting the need to refine and streamline existing methods to increase adherence to best practices. Large language models (LLMs), with their high capacity for pattern recognition and human-like text generation capabilities, offer many new possibilities to address these challenges. While initial studies have primarily focused on using LLMs for item generation, their potential to streamline other aspects of test development (e.g., in identifying potentially biased items or by harnessing linguistic cues to supplement statistical evidence when data are limited) remains largely unexplored. In this talk, I discuss the challenges of conventional test development, review emerging applications of LLMs in psychometrics, and present findings from a systematic investigation of whether, when, how, and to what extent LLMs may be leveraged during the highly resource-intensive and iterative test development process. This work explores and highlights LLMs' potential to enhance and complement current practices to reduce researcher burden, improve the validity and fairness of psychometric measures, foster greater accessibility for researchers in fields with limited resources, and enable a more widespread adoption of rigorous methodological practices.

Title

Towards the adaptation of the learning patterns model in Primary Education: an explanatory study

Author(s)

Anna Ciraso-Calí , J. REINALDO MARTÍNEZ-FERNÁNDEZ [1] , Carla Quesada-Pallarès [1]

[1] Universitat Autònoma de Barcelona

Abstract

This study explores the transferability of Vermunt's (1998) learning patterns model (that focuses on learning conceptions, motivational orientation, regulation strategies, and cognitive processing strategies) in Primary Education. Traditionally used in Higher Education, this model is less explored in other educational contexts, creating a gap in the literature that this research aims to address.

The research involves a mixed-methods approach (S-QUAN–>qual) and includes 218 students from three public primary schools in the Murcia region (Spain). Participants are students in 4th, 5th, and 6th grades, with a balanced gender distribution. Quantitative data were collected using the adapted version (Martínez-Fernández et al., 2015) of the Inventory of Learning Patterns of Students (ILS). Qualitative data were gathered through three focus groups with a subsample of 25 participants; where students shared their views on learning, motivations, and strategies, through activities involving collage and drawing.

Through robust exploratory factor analysis, we detected that younger students' learning patterns differ significantly from those of adults, with distinct factor structures emerging. For instance, the regulation strategies component showed clear differentiation between self-regulation, external regulation, and lack of regulation. However, other components revealed significant differences. In the learning conceptions component, there was no solid belief about learning as knowledge construction; instead, items from this subscale integrated into the conceptions of learning as use and knowledge increase. Motivational orientations also varied, with a new "challenge-orientation" emerging. This orientation reflects a motivation to tackle difficult tasks, demonstrating one's abilities. This finding aligns with previous research by Severiens and Ten Dam (1997), who identified a similar orientation in Secondary Education for adults with a history of academic failure.

In terms of processing strategies, the factor structure simplified into two types: concrete/deep processing, that involves elaboration and structuring of content, as well as transferring learning to other contexts and problem-solving (which aligns with Hattie and Donoghue's (2016) deep learning and transfer consolidation phase); and superficial processing, that includes memorization, rehearsal, and some elements of critical processing.

Random forest cluster analysis on ILS scores revealed that the largest group in the sample could not be characterized yet by a specific pattern. However, a distinct cluster of students with an undirected (UD) pattern was identified, clearly differentiated from meaning/application-oriented (MD/AD) and reproduction-oriented (RD) groups.

Qualitative analyses suggested the need to reformulate some ILS items and expand the teacher stimulation subscale to include families. Emotional aspects also emerged as significant, with students referring to regulation and processing strategies as ways to control stress and manage emotions. This highlights the importance of considering emotional factors in learning patterns, as suggested by Ahmedi and Martínez-Fernández (2023).

Overall, this research provides valuable insights into the applicability of Vermunt's learning patterns model in primary education. It highlights the need for further adaptation of the model and the instruments to better suit younger students' unique learning processes and developmental stages. The findings underscore the importance of integrating emotional and contextual

factors into the model to provide a more comprehensive understanding of primary students' learning patterns.

Title

Using pictographic single-item measures to overcome psychometric challenges: A pre-registered development and validation of the Meditation Pictographic Scale (MPS)

Author(s)
Rosa Baños [1] , Maja Wrzesien [2] , Catherine Andreu [3] , Ausiàs Cebolla [2] , Rocío Herrera [4] , Ylenia D’Elia [5] , Desirèe Colombo [6] , Oscar Lecuona [7]

[1] Faculty of Psychology, Universitat de València, Spain; CIBERObn Ciber Fisiopatologia de la obesidad y la nutrición, Madrid, Spain; [2] Faculty of Psychology, Universitat de València, Spain; [3] Faculty of Psychology, Universidad Santiago de Compostela; [4] Instituto Polibienestar, Universitat de València, Spain; [5] Brain and Creativity Institute, University of Southern California, United States; [6] Faculty of Psychology, Universitat Jaume I, Spain; [7] Complutense University of Madrid

Abstract

Concepual framework: Meditation and mindfulness research has encountered some psychometric challenges, among which are cross-cultural differences or linguistic biases. This has led to standard psychometric developments (based on verbal self-reports) to show unclear psychometric properties. In addition, these fields stress the importance of first-person phenomena as the main mechanisms of change. Thus, a psychometric challenge is drawn, from which we propose the use of pictographic single-item measurements as a Meditation Pictographic Scale (MPS). These way verbal and cultural phenomena could be mitigated and allow for a more precise measurement.
Methods: A three-stage pre-registered development was implemented. First, we reviewed literature and selected the most established meditation-related experiences that were translatable to items. Second, 10 expert meditators were interviewed on the items with semi-structured phenomenological and cognitive interviews to explore understandability, relevance and representativity of the items, along with response processes. Independent raters coded the interviews to extract insights, to produce a refined MPS. Third, 376 participants from the general population were randomized into four groups (control, and three different meditations) and responded to the refined MPS and other state measures pre and post. Linear models were implemented to explore sensitivity of the scale to meditation. To study internal structure, network analysis was implemented due to having single-item measures without assumptions of latent variables or reflective/formative models.
Results: Experts reported general understandability and provided critical improvements of pictographs. The MPS showed to be sensitive to meditation and concurrent with state measures, especially regarding body, self, affect, compassion, and time. Networks showed a relatively connected network with some exceptions. Exploratory Graph Analysis showed two stable communities of items (positive and negative experiences) and a more heterogeneous community (distortions). One item was revealed as reverse due to their negative associations with in-group items. Meditation also seemed to impact associations between items but not community structures.
Implications: The MPS shows stable psychometric properties regarding contents, response processes, internal structure, and relation with relevant variables. Thus, it seems to overcome psychometric challenges present in literature. Future research should explore the MPS on clinical and trained samples trained, and the impact of manualized meditation programs.

Title

Play-Along Interviews: A Child-Centered Method for Research with Children

Author(s)
Elisavet Pasidi [1] , Gijs van Campenhout [1] , Kirsten Visser [1] , Gideon Bolt [1]

[1] Utrecht University

Abstract

In an attempt to amplify children's under-represented voices in research, we explored the "play-along interview"as an innovative method for collecting context-rich data directly from children while informing methodological innovation based on children's lived experiences. In this presentation, we will share how we applied this method in practice and discuss its opportunities and challenges.

Building upon the "go-along"method traditionally used with adults –wherein researchers accompany participants through their environments to gain real-time insights into their experiences–we embed interviews within play, a domain where children are experts. The play-along interview empowers children's active participation in age-appropriate yet meaningful ways, addressing the limitations of existing methods. For example, traditional question-and-answer interview formats are rather hierarchical and often fail to engage young children effectively, leading to disinterest and missed opportunities to capture their authentic perspectives. Tapping into children's natural enthusiasm, creativity, and mastery of play, the play-along interviews create an engaging context for children to express their thoughts and feelings in ways that are meaningful to them. Other methods of incorporating children's perspectives in research, such as asking parents or teachers to answer on their behalf or relying on observations, are inherently limited by adults'subjective interpretations rather than children's direct accounts. In contrast, play-along interviews enable researchers to gather firsthand insights directly from children and ask follow-up questions in real time, offering a more nuanced understanding of their views. Even when children are directly involved in research through questionnaires, significant limitations exist. Younger children usually lack the literacy skills to complete questionnaires independently, while older children can still struggle with complex questions, requiring adult mediation that may constrain their responses. On the other hand, play-along interviews offer an interactive approach that adapts to children's developmental levels and communication styles, with the researcher entering the children's world, without supervising adults present.

We applied play-along interviews in the context of "risky play", which refers to exciting forms of play where children take age-appropriate risks (e.g., climbing trees). Risky play is vital for children's development, yet much research in this area has focused on adult perspectives, who tend to prioritize safety over excitement. In this study, we accompanied children as they explored their environments without supervising adults, allowing them to lead the way and engage in (risky) play. Our aim was to reveal the complex, situated ways in which residential environments influence children's play behavior. Conducting interviews in the environments where children actually play allowed for a deeper understanding of their behaviors, while enhancing their ability to recall and reflect on their experiences. This contextual understanding provided insights into how neighborhood factors (e.g., green, traffic) influence children's risky play. Additionally, we plan to use children's insights to refine existing factors from the literature to develop targeted questionnaires that are contextually relevant to children's lived experiences.

The play-along interview is a powerful method for studying children's lives, environments, and behaviors across diverse fields. Beyond this specific application, it demonstrates how participatory, child-centered methods can inform methodological innovation, advancing qualitative research and complementing mixed-methods studies.

Title

Assessment of Problematic Social Networking Sites Use Among Youth and Adolescents

Author(s)
Álvaro Postigo [1] , Covadonga González-Nuevo , Jaime García-Fernández [1]

[1] University of Oviedo

Abstract

Introduction: The negative consequences of the misuse of Social Networking Sites (SNS), particularly problematic use, are a growing concern for parents of children and adolescents. The first questionnaire designed to assess not only addictive but also comparative SNS use, known as the Problematic SNS Use Questionnaire (UPS), has been validated exclusively in adult populations.
Objective: The objective of this study was to adapt and validate the UPS questionnaire in a sample of children and adolescents.
Method: The sample consisted of 443 participants (53.4% female) with a mean age of 15.04 years (SD = 1.34) and an age range of 12 to 18 years. Participants completed an online survey that included sociodemographic questions, SNS usage frequency, the Problematic SNS Use Questionnaire (UPS), the Social and Emotional Competence Questionnaire (SEC-Q), and the Big Five Inventory-2 (BFI-2). The psychometric properties of the UPS were assessed, including internal consistency, reliability, and validity evidence concerning internal structure and relationships with other variables.
Results: Confirmatory factor analysis supported the two-dimensional structure of the original version, showing a good model fit with adequate reliability for both subscales and appropriate validity evidence concerning its relationship with other variables.
Conclusions: The UPS version for children and adolescents has demonstrated adequate psychometric properties, making it a suitable tool for the general Spanish child and adolescent population. The findings are discussed in the context of preventing problematic SNS use, emphasizing the importance of early detection and intervention strategies.

Title

Setting Stopping Rules for Progressive Tests: A Practical and Transparent Toolkit

Author(s)
Jianan Chen [2] , Ellen Irén Brinchmann [1] , Johan Braeken [2]

[1] Department of Special Needs Education, University of Oslo, Oslo, Norway; [2] Centre for Educational Measurement, University of Oslo

Abstract

In psychological assessment, a popular test design is administering items in order of progressively increasing difficulty, referred to as a progressive test. Prime examples include subtests of the well-known Wechsler intelligence test batteries. Progressive tests are based on the concept of a Guttman scale, which facilitates intuitive score interpretation, and the application of stopping rules in test administration for higher efficiency and less test burden. However, in practice, stopping rules often lack a well-documented empirical basis and justification due to the absence of clear standards and guidelines, raising general concerns about fairness and validity. To facilitate evidence-based decision-making for setting appropriate stopping rules, we propose a transparent approach that charts the impact of varied alternative stopping rules on Accuracy (e.g., test outcomes at the individual and group level, and norm tables) and Efficiency (e.g., person-specific test length). These A-E charts are based on retro-actively applying stopping rules to normative data administered without a stopping rule or with a default stopping rule. An empirical working example is used to illustrate the proposed toolkit. We show that a universal stopping rule likely does not exist and that the optimal rule varies as a function of the desired efficiency-accuracy trade-off suitable for the intended test use and target population. The proposed approach provides a pragmatic solution for practitioners, researchers, test developers, and test publishers to rethink the existing stopping rules, systematically evaluate the alternatives, and set appropriate rules.

1.3 Session 6 : ”Structural models and complex data analysis”

Title

Standard Error Estimation in the Local Structural-After-Measurement (LSAM) Approach

Author(s)
Seda Can [1] , Yves Rosseel

[1] Izmir University of Economics

Abstract
Accurate estimation of standard errors (SEs) is essential in structural equation modeling (SEM) as it quantifies the uncertainty of parameter estimates, plays a critical role in computing test statistics and p-values, and ensures robust inferences about population parameters. While standard SEM typically relies on a joint or "system-wide"estimation approach, the Local Structural-After-Measurement (LSAM) framework separates the estimation of the measurement and structural parts. Despite the increasing attention to LSAM's potential in SEM, prior studies have primarily focused on point estimates, leaving the behavior and accuracy of SEs within this framework underexplored. This study addresses this gap by evaluating SE estimation in the LSAM framework under challenging research conditions, including nonnormal data, smaller sample sizes, and model misspecification.

Two simulation studies were conducted to assess the performance of various SE estimation methods within the LSAM framework. The methods included analytic approaches, such as two-step estimation, and resampling-based methods, including parametric and nonparametric bootstrapping. Study 1 investigated a two-factor SEM, focusing on misspecification in the measurement model, while Study 2 expanded the model by incorporating observed exogenous and endogenous variables and introducing misspecification in the structural model. The simulations varied conditions such as normality versus nonnormality, correctly specified versus misspecified models, and sample sizes ranging from small to large.

The results showed that the nonparametric bootstrap excelled under nonnormal data, providing near-unbiased SE estimates regardless of model specification. The parametric bootstrap produced minimal bias under normal conditions across all sample sizes, even when models were misspecified. Analytic two-step method was effective under normal conditions but exhibited higher variability under nonnormal data, particularly in smaller samples and misspecified models.

By incorporating both parametric and nonparametric bootstrapping, this study offers new insights into the potential benefits and limitations of resampling-based SE methods within LSAM. Our findings emphasize the robustness of LSAM methods for SE estimation, particularly in research contexts characterized by data nonnormality, small sample, and potential model misspecification. Furthermore, this study expands LSAM's utility beyond point estimates to SE estimation, offering valuable implications for researchers working in less-than-ideal conditions.

Title

The Structural-after-Measurement (SAM) approach: updates and extensions.

Author(s)
Yves Rosseel [1]

[1] Ghent University

Abstract

In the Structural-after-Measurement (SAM) framework for Structural Equation Modeling (SEM), parameters of the measurement model are estimated first, followed by the estimation of the structural model parameters. This presentation focuses on the ‘local’ SAM (LSAM) approach, where summary statistics (mean, covariance matrix) of latent variables are derived in the first step. In this presentation, I will present recent LSAM developments, including: 1) incorporating binary or ordinal indicators in the measurement model, 2) integrating multiple interaction and quadratic terms into the structural model, and 3) applying the infinitesimal jackknife technique to obtain local two-step standard errors. All these extensions have been incorporated in the sam() function, which is part of the R package lavaan.

Title

A simulation study comparing structural-after-measurement versus traditional approaches to estimate nonlinear effects in structural equation modeling

Author(s)
Felipe Vieira [1] , Yves Rosseel [1]

[1] Ghent University

Abstract

Several methods have been proposed to include quadratic or interaction terms involving latent variables in structural equation models. Some examples are the latent moderated structural equations (LMS) approach and several variants of the product indicator (PI) approach. All of these methods use a system-wide estimation approach and estimate the free parameters of the model simulta-neously. They tend to perform effectively in models featuring only a limited number of nonlinear effects. However, as the complexity of the model increases with a higher number of nonlinear terms, the feasibility of joint or one-step methods progressively diminishes. Recently, several so-called structural-after-measurement (SAM) approaches to handle latent quadratic and interaction terms have been proposed in the literature. In a SAM approach, estimation

proceeds in two stages. In a first stage, we estimate the parameters related to the measurement part of the model, while in a second stage, we estimate the parameters related to the structural part of the model. In this presentation, we will present initial simulation results where different one-step and SAM approaches are compared, in a large variety of conditions, including conditions that involve a large number of latent interaction and quadratic terms.

Title

Extending Log-Logistic IRT: A Multidimensional Model and Its Implementation in R

Author(s)
David Navarro-González [1] , Fàbia Morales-Vives [1] , Pere Joan Ferrando [1]

[1] Universitat Rovira i Virgili

Abstract

Unipolar log-logistic response models (LLRM) are a family of Item Response Theory (IRT) models intended for measuring traits that take only positive values and are asymmetrically (rightly skewed) distributed in the target population. Their two most distinctive features are that: (a) the item response functions are not ogives but power functions, and (b) the amount of information decreases with trait levels. LLRMs have be found to be appropriate in the measurement of certain clinical and forensic traits, symptoms checklists, addictive behaviours, and irrational beliefs among other applications.
LLRMs were originally intended for unidimensional instruments based on binary items (Lucke, 2015) and next extended to the graded-response case (Reise et al. 2021). Our research group has further extended the seminal proposals in two directions (Ferrando, et al. 2024, 2025). First, we have also considered double-bounded continuous-response items. Second, for each item format, we have started to develop multidimensional extensions. Specifically, we shall first present a novel LLRM we propose for multidimensional instruments based on continuous-response items: the Md-LL-CRM. Our presentation of this model has two parts. First, we aim to discuss its main features and functioning, particularly: basic equations, item response surfaces, multidimensional location and discrimination parameters and information functions.
The second part is more practical and includes proposed estimation and scoring procedures for LLRMs in general and Md-LL-CRM in particular, as well as their general implementation, and we shall introduce SkewIRT, a new R package designed to facilitate the usage of LLRMs. SkewIRT provides functions for model estimation, item and test information analysis, and individual scoring. Additionally, it includes visualization tools to explore item response surfaces and multidimensional trait information. By integrating these capabilities into a single framework, SkewIRT aims to make LLRMs more accessible to applied researchers in behavioral assessment.
SkewIRT will be released in the coming months, offering two versions: a code-based version for experienced R users and a Graphical User Interface (GUI) version developed with shiny (Chang et al., 2025), designed for researchers with limited programming experience.
In this presentation, we will demonstrate the main functionalities of SkewIRT through illustrative examples.

Title

Exploring the Consensus Emergence Model as a Tool for Analyzing Intra-Individual Variability in Psychological Development

Author(s)
Alejandro Díaz-Guerra [1] , Paul Bliese [2] , Mirko Antino [1]

[1] Complutense University of Madrid; [2] University of South Carolina

Abstract

Introduction:
The Consensus Emergence Model (CEM) is an advanced analytical model widely utilized in Organizational Psychology to study the development of shared perceptions, feelings, or climates within groups over time. This approach extends traditional multilevel methodologies by incorporating the examination of residual variances within growth models, capturing the dynamic processes underlying consensus emergence. The CEM formally tests for consensus formation in longitudinal designs where timepoints are nested within participants and participants are nested within groups.
Building on this framework, we propose that the CEM is also a suitable tool for analyzing intra-individual changes in the variability of psychological constructs. By extending its application to longitudinal designs where timepoints are nested within items and items are nested within participants, the CEM provides a unique lens for studying dynamic psychological processes at the individual level.

Objective:
This study aims to explore the potential of the CEM as a valid and innovative tool for analyzing changes in intra-individual variability of psychological constructs over time.

Method:
We re-analyzed data from a previously published study on the development of the Moral Self-Concept (MSC) in children. The study involved 82 children assessed at two time points –when they were 4 and 5 years old –using a puppet-interview methodology. The original authors computed MSC coherency scores manually (based on the variability of item responses) and used paired t-tests to examine developmental changes. In contrast, we applied the CEM to investigate the possible emergence (i.e., increased coherency) of MSC and its subdimensions: helping, sharing, and comforting self-concepts. All analyses were conducted using the R programming environment.

Results:
Our findings partly contradict those of the original study. While the initial results indicated a significant increase in MSC coherency from ages 4 to 5, our analysis yielded inconclusive results (δ_1= -0.08, p = .061). However, we did find a statistically significant emergence pattern in the sharing self-concept subdimension (δ_1= -0.19, p = .030).

Conclusions:
Our preliminary findings suggest that the CEM is a valuable and complementary analytical method for studying changes in intra-individual variability. This approach may offer meaningful insights, even in scenarios where traditional longitudinal invariance tests fail to provide conclusive evidence. By expanding the use of the CEM beyond its conventional applications, we contribute to advancing the methodological toolkit for longitudinal research in Psychology.

Title

Alignment optimization and the sequential testing approach for data harmonization

Author(s)
Meltem Ozcan [1] , Hok Chio (Mark) Lai [1]

[1] USC

Abstract

Technological advances and open science initiatives have made large-scale datasets more readily available than ever before. By combining data from different studies, researchers can attain larger and more diverse samples, increase statistical power, explore novel and broader research questions, study rare outcomes, and identify variations across populations and settings. However, cross-study discrepancies, such as non-standard or partial item administration, missing items, or other operational, contextual, or methodological differences, pose significant challenges to the comparability of datasets, and may undermine the validity of pooled analyses. Traditional approaches such as the sequential testing approach, which involves iteratively constraining and freeing measurement parameters to find a model that produces unbiased estimates and places factor scores on a common metric, enable the simultaneous and flexible management of measurement error and cross-group differences by statistically linking observed test items to a latent construct. While the sequential testing approach is widely used, it is often time-consuming, error-prone, and not scalable to many groups, limiting its practicality in harmonization efforts. Alignment optimization (AO; Asparouhov & Muthén, 2014), a recently developed method which reframes the model specification process as an optimization problem, is an efficient alternative to the traditional method. In this talk, we introduce the sequential testing approach and AO methods and demonstrate their application, relative strengths and limitations, and effectiveness in producing comparable factor scores through an empirical illustration harmonizing large-scale national datasets. Our findings highlight key trade-offs between precision, scalability, and efficiency, positioning AO as a practical and robust alternative to the traditional sequential testing approach.

1.4 Symposium : ”Advancing many groups comparisons: Mixture multigroup approach for latent variable analysis”

Title

Mixture Multigroup Structural Equation Modeling: An empirical application revealing cross-national patterns in how human values predict climate policy support

Author(s)
Jeroen Vermunt [1] , Kim De Roover [2] , Meijun Yao [2]

[1] Tilburg University; [2] KU Leuven

Abstract

The increasing accessibility of large-scale international surveys has provided new opportunities for social scientists to conduct comparative research. Such studies frequently examine relations between latent constructs (e.g., how perceived economic threat affects political ideology) and compare them across groups (e.g., countries) to reveal cultural variations in value priorities, attitudes and behavioral patterns. Structural Equation Modeling (SEM) is the state-of-art method for analyzing and comparing these complex relations. While these relations often differ across groups, similarities may emerge among certain groups, leading to the formation of clusters, especially in the case of many groups.

Latent constructs are measured indirectly by multiple questionnaire items. To enable valid comparisons of their relations across groups, the measurement of latent constructs should be invariant. However, when dealing with multiple groups, there are often some differences (or non-invariances) in the measurement models (MM). It is important to capture them in the SEM model to avoid biased estimations of the structural relations. Mixture Multigroup Structural Equation Modeling (MMG-SEM) has recently been proposed as a novel method for identifying clusters of groups with equivalent structural relations, while accounting for measurement non-invariances through group-specific measurement parameters. To demonstrate its application in empirical research, we apply MMG-SEM to European Social Survey Round 8 (ESS8) data, uncovering cross-national differences and similarities in the relations between human values, climate change belief, and support for climate change policies across 23 countries.

Title

Extending Mixture Multigroup Structural Equation Modeling to deal with ordinal variables

Author(s)
Andres Felipe Perez Alonso [1] , Jeroen Vermunt [1] , Kim De Roover [2] , Yves Rosseel [3]

[1] Tilburg University; [2] KU Leuven; [3] Ghent University

Abstract

The recently proposed Mixture Multigroup Structural Equation Modeling (MMG-SEM) efficiently compares groups by clustering them based on their structural relations while accounting for the reality of measurement (non-)invariance. Currently, MMG-SEM relies on maximum likelihood (ML), which assumes continuous and normally distributed observed indicators. However, this can introduce bias when applied to ordinal data, which is often used in social sciences. In this paper, we extend MMG-SEM to accommodate ordinal data relying on the Structural-After-Measuremt stepwise estimation approach. In the first step, we implement a multi-group categorical confirmatory factor analysis (MG-CCFA) with diagonally weighted least squares (DWLS) to estimate the measurement model (MM). The second step uses ML to estimate structural relations and perform clustering. A simulation study evaluates the performance of this approach compared to traditional ML-based MMG-SEM under various conditions. The results show a better recovery of MM parameters with DWLS, particularly with fewer response categories, whereas both approaches perform similarly in structural model recovery.

Title

Evaluating the Efficacy of Mixture Multigroup Factor Analysis in Handling Non-Normal and Ordinal Data: A Simulation Study

Author(s)
Eva Ceulemans [1] , Francis Tuerlinckx , Kim De Roover [1] , Lucas Nowicki [1]

[1] KU Leuven

Abstract

In the social sciences, a common research objective is the comparison of latent variables among different groups, such as in cross-cultural studies. For making valid comparisons measurement invariance (MI) is required, which implies that constructs are measured consistently across populations. When dealing with many groups, MI often does not hold, requiring pairwise comparisons between the groups to identify the sources of non-invariance. However, such comparisons can become impractical when dealing with many groups. Mixture multigroup factor analysis (MMG-FA) offers a solution by clustering groups based on their measurement parameters. This method captures between-group differences and similarities in measurement parameters without requiring extensive pairwise comparisons. It combines cluster-specific and group-specific parameters to cluster groups on specific subsets of measurement parameters, for instance, on factor loadings to achieve metric invariance within each cluster of groups. An EM algorithm was developed for MMG-FA to drastically lower the computation time, but this is specific to maximum likelihood estimation. However, the use of maximum likelihood estimation in MMG-FA assumes continuous items and underlying multivariate normality—assumptions that are not always tenable in real-life settings. Using simulations, we investigate MMG-FA' s performance with ordinal data and underlying non-normal distributions when clustering based on factor loadings, to examine the robustness of MMG-FA to violations of the previously mentioned assumptions.

Title

MixMG-SEM with double mixture modeling to capture similarities in measurement model and in structural relations across many groups

Author(s)
Hongwei Zhao [1] , Jeroen Vermunt [2] , Kim De Roover [1]

[1] KU Leuven; [2] Tilburg University

Abstract

Comparing relations between latent constructs across groups is essential for understanding social phenomena in different contexts. A key assumption for valid comparisons of such relations is that the constructs are measured equivalently across the groups, referred to as “measurement invariance”. Specifically, partial metric invariance is sufficient –meaning that at least some factor loadings are invariant across groups –provided that non-invariant measurement parameters are appropriately accounted for in the model. To address this, we propose double-mixture multigroup structural equation modeling (2MixMG-SEM), which applies a mixture clustering of the groups to capture differences in the measurement model (measurement clusters) and another mixture clustering to capture differences in the structural relations (structural relations clusters). 2MixMG-SEM thus captures measurement non-invariance with cluster-specific measurement parameters, as opposed to mixture multigroup SEM (MixMG-SEM), which captures them with group-specific parameters. We therefore expect 2MixMG-SEM to perform better than MixMG-SEM when some groups are too small for group-specific parameters to be accurately estimated. Through a simulation study, we evaluate 2MixMG-SEM’s performance, addressing key challenges such as classification uncertainty and selecting the cluster numbers for both layers of clustering.

Title

Evaluating (Mixture) Multigroup Structural Equation Modelling with Exploratory Measurement Models

Author(s)
Jennifer Dang Guay [1] , Kim De Roover [1] , Yves Rosseel [2]

[1] KU Leuven; [2] Ghent University

Abstract

Structural equation modelling (SEM) is the state-of-the-art method for analysing relations between latent variables (e.g., attitudes or behaviours), also called ‘factors’. SEM consists of a measurement model (MM), which specifies how questionnaire items measure the factors, and a structural model (SM), which captures the relations of interests. Traditionally, SEM estimates the MM and the SM simultaneously, whereas the structural-after-measurement (SAM) approach estimates the MM first, and then the SM. When comparing relations across multiple groups (e.g., countries), measurement invariance (MI) is a prerequisite. When MI fails, it is crucial to model the measurement non-invariances to avoid biasing the comparisons. Multigroup exploratory factor analysis (MG-EFA) estimates all factor loadings, and thus allows to identify all kinds of loading non-invariances. The choice of rotation in MG-EFA can, however, affect the MM. Also, rotation per group affects the detection of loading non-invariance, as it disregards the loading agreement between groups. But this is accounted for by multigroup alignment (MG-A) and multigroup factor rotation (MG-FR). In this talk, I will present the results of a simulation study that evaluates how well MG-A and MG-FR perform to recover the measurement parameters and loading (non-)invariances (when using different rotations) in the first step of multigroup exploratory SAM (MG-ESAM), and how this, in turn, affects the recovery of the relations in the second step. We examine how MG-ESAM can be used to tackle the challenge of finding the most optimal rotation before comparing the relations across groups. Finally, I will present some specific challenges to consider when extending MG-ESAM into Mixture Multigroup ESAM to find clusters of groups based on their structural relations.

Title

Mixture Three-Step Latent Vector Autoregression to Find Individuals With Similar Dynamic Processes

Author(s)
Jeroen Vermunt [1] , Kim De Roover [2] , Leonie V.D.E. Vogelsmeier , Manuel Tobias Rein [1]

[1] Tilburg University; [2] KU Leuven

Abstract

Researchers often use vector autoregressive models to study dynamic processes of latent variables in daily life, such as the extent to which positive and negative affect carry over and interact with each other from one moment to the next. Mixture modeling allows finding clusters of individuals that are similar to each other in their dynamic processes. However, applying MMG-SEM to vector autoregressive models is not straight-forward. For example, not only metric, but also (partial) scalar invariance has to hold. To validly cluster individuals based on their dynamic processes while accounting for partial measurement non-invariance, we present an extension of the recently proposed Three-Step Latent Vector Autoregression (3S-LVAR). We discuss challenges that arise when applying the idea of MMG-SEM to intensive longitudinal data and how to tackle them.

1.5 Symposium : ”Mixed methodologies”

Title

Analysis of the relationship between self-esteem level and interest, importance and learning achievements in highly able students through mixed method design

Author(s)
Ainize Sarrionandia , Karmele Salaberria , Leire Aperribai Unamuno [1]

[1] University of the Basque Country UPV/EHU

Abstract

The scientific literature supports the relationship between self-esteem and interest or motivation, academic achievements, learning processes and academic performance of students in general, and also in highly able students. Thus, the objective of this study is to analyse how the different levels of self-esteem are related to interest, importance and learning achievements in highly able students in Primary and Secondary Education levels. The sample consisted of 25 students (8 girls and 17 boys), aged 10-13 years, from the province of Gipuzkoa. The Rosenberg Self-Esteem Scale (EAR) was applied and open questions were asked about the importance and usefulness of the subject content, and about the achievements and effort made. The results show that a majority (60%) has high self-esteem, while the rest has average (28%) and low (12%) self-esteem. It has been found that, as the level of self-esteem increases, the number and type of forms or words related to interest/importance and achievement in learning increases.
Finally, the descending hierarchical analysis reveals the importance of those forms related to the need to study more and in a more active way, which have been significantly related to a low self-esteem. In conclusion, self-esteem is important when it comes to promoting academic performance through the interest and achievement in learning of highly able students.

Title

Influence of Intelligence and Gender on Mathematics Anxiety: Verbalized Strategies to Overcome Difficulties

Author(s)
Ǎfrica Borges del Rosal [1] , Jesús del Pino Relwani Moreno , Juan Francisco Flores Bravo [2] , Adalberto González Martín [1]

[1] Universidad de La Laguna; [2] University Center for Health Sciences, University of Guadalajara

Abstract

Introduction:
Mathematics anxiety negatively impacts performance, achievement, and career choices. This study investigates how intelligence and gender influence this anxiety and explores the coping strategies used by those who dislike math. Existing research shows that lower intelligence and being female often correlate with higher math anxiety, but how these factors interact with coping mechanisms is less understood. This research employs a mixed-methods approach to analyze these relationships, aiming to answer how intelligence and gender affect anxiety levels, what strategies are used to manage this anxiety, and if these strategies vary based on intelligence and gender. By examining these questions, this study will contribute to a deeper understanding of mathematics anxiety and inform the development of tailored interventions to support learners, ultimately fostering a more positive and inclusive mathematics learning environment.
Objective:
To investigate the influence of intelligence and gender on mathematics-related anxiety and to analyze the verbalized strategies employed by individuals who dislike mathematics to overcome their difficulties.
Method:
A mixed-methods research methodology (MMR) was used, with the quantitative part employing a cross-sectional survey design and the qualitative part using content analysis. The sample consisted of 1552 students, of whom 1012 were women.
Results:
Significant differences were found based on both intelligence and gender. Women and individuals with lower cognitive abilities reported higher levels of anxiety related to mathematics. The verbalized strategies were grouped into four clusters: cluster 1 (26.6%), relying on external support, such as peers, teachers, and online resources, to seek help and guidance; cluster 2 (23.7%), using extra learning aids like video tutorials and procedural explanations, with a focus on practical exercises; cluster 3 (24.4%), simplifying and understanding basic concepts through individual effort and perseverance; cluster 4 (25.3%), developing study habits, including problem-solving, note-taking, and concentration-focused strategies.
Conclusions:
The findings highlight how gender and cognitive abilities influence mathematics-related anxiety and coping mechanisms. Women and higher math anxiety tend to rely on external support and tend to seek alternative learning strategies, while men and lower math anxiety prefer individual effort and perseverance . These insights underline the importance of tailoring support strategies to the specific needs of learners who dislike mathematics.

Title

MMR Approach in the Study of Physical Activity, Intelligence, and Gender in Adolescents

Author(s)
África Borges del Rosal [1]

[1] Universidad de La Laguna

Abstract
Introduction: One of the myths surrounding high intellectual abilities is the belief that individuals with higher intelligence are not interested in physical activity, implying a relationship between intelligence and exercise.
Objective:To analyze the relationships between intelligence and physical activity, as well as gender differences in interest in physical activity, while also studying perceptions about physical activity.
Methodology: The sample consisted of 297 secondary school students aged 13 to 16. The instruments used included an intelligence test (Herranz's G factor), a physical activity questionnaire (PAQ-A), and two open-ended questions: "Do you like engaging in physical activity?" and "Why?"Quantitative analysis included Pearson's correlation to examine the relationship between intelligence and physical activity and Student's t-test to study gender differences in interest in physical activity. Qualitative data were analyzed using the IRAMUTEQ software.
Results:No relationship was found between intelligence and physical activity, and boys showed greater interest in exercise. Qualitative analysis revealed three main themes: engaging in sports, enjoyment, and mental benefits. The significance of the independent variables used was also analyzed.
Discussion: Further research is needed to explore the relationship between intelligence and physical activity to confirm the independence of these variables. It is also important to examine the reasons behind gender differences, which show a greater interest in exercise among boys. The qualitative analysis offers three perspectives on understanding physical activity: the act of exercising itself, its recreational aspects, and the mental health benefits it provides.

Title

Motives for Lying in Mexican Adolescents

Author(s)
Beatriz Viera Delgado [1] , Jesús del Pino Relwani Moreno

[1] Universidad de La Laguna

Abstract
Introduction: Several studies have indicated that the tendency to lie is more prevalent in adolescents compared to children and adults (Buta et al., 2020; DePaulo et al., 1996; Levine et al., 2013). Studying the motivations behind this behaviour can be essential to gaining a deeper understanding of this phenomenon.
Objective: Study the different reasons for lying among the adolescent population.
Method: The methodology used was Mixed Methods Research (MMR). The sample consisted of a total of 433 adolescents (M=12.77; SD=.97) from the general population of the State of Jalisco, Mexico (42.60% women). For data collection, the CEMA-A questionnaire (Armas-Vargas, 2023) and an open-ended question about the main reasons for lying were used. For quantitative data analysis, the SPSS program, v.26, was used, and for qualitative analysis, the lexical analysis software IRAMUTEQ 0.8a7 was employed.
Results: The MANOVA was significant for the interaction of gender and age variables. The analysis of qualitative responses allowed the extraction of two classes: ”avoiding harm or punishment”(34%) and ”hiding information”(66%). Significant relationships were observed between this two classes and the study's different quantitative variables.
Conclusion: This research contributes to the existing literature by providing novel data on the motives that drive the lying behaviour in adolescents. We highlighted the importance of using MMR and suggest continuing this line of study with a sample of older participants.

Title

Multipotentiality in university students and its relationship with gender, high abilities and entrance score

Author(s)
Juan Francisco Flores Bravo [1] , Maria Dolores Valadez Sierra , África Borges del Rosal [2] , Elena Rodríguez Naveiras [2]

[1] University Center for Health Sciences, University of Guadalajara; [2] Universidad de La Laguna

Abstract

Introduction Multipotentiality, defined as the ability to excel in diverse areas of interest (Cordero, 2019), has been explored through individual differences and educational factors that favor its development. Previous studies highlight that factors such as gender and high abilities play crucial roles in how individuals explore and manage multiple talents (Kerr Huffman, 2018). For example, Jung (2019) notes that students with high abilities often have a wide variety of interests and abilities, which can lead them to face the challenge of having numerous options, but having difficulty choosing just one. Objective To analyze how multipotentiality in college students is associated with variables such as gender, high ability, and college entrance score. Method MMR was used with a transforming concurrent design. The sample consisted of 1,446 (997 women) university students from various degrees in the area of Health Sciences, who were given a questionnaire that included, among others, an open- ended question on the areas or disciplines in which they perceive themselves to have a high level of competence. Quantitative analysis was performed by means of Student39;s t test using SPSS v.27, while the open-ended responses were classified using the ALCESTE program. Results The qualitative analyses of the verbalizations refer to two classes “Personalization of competencies”(47.57%), which reveals that multipotentiality is significantly related to being male, having high abilities, and obtaining a high score in the university entrance exam, especially in careers related to Sports and Nutrition. In contrast, the class “General competencies”(52.43%) reflects that women without high abilities and with average entrance scores tend to identify less with multipotentiality, especially in the Psychology career. At the quantitative level, no significant differences were observed with respect to gender, intelligence and university entrance score. Conclusions The findings suggest that high ability and entry score are significant predictors of multipotentiality in college students. These results may guide future educational policies and support programs to foster the development of multiple competencies in diverse student populations.

Title

Subjective perception of Fear of Public Speaking. A mixed-methods research

Author(s)
África Borges del Rosal [1] , Ernesto Pereda de Pablo , Ricardo Quintero Rodríguez

[1] Universidad de La Laguna

Abstract

Introduction. Fear of Public Speaking (FoPS) or Public Speaking Anxiety (PSA), considered a specific subtype of Social Anxiety Disorder, profoundly affects the personal, academic, and professional spheres. This phenomenon is characterized by cognitive, emotional, and physical manifestations that limit the performance and social interactions of those who experience it.
Objectives. The relationship between Fear of Public Speaking and its associated manifestations is explored through a mixed research approach. Methodology. A sample of 436 university students (26% male, 74% female; mean age = 21.1 ± 3.46 years) was obtained through convenience sampling. Participants were classified into different anxiety levels using the Social Anxiety Questionnaire for Adults (SAQ-A30). Quantitative analyses were performed with SPSS v.29 software, while qualitative data, derived from open-ended responses on thoughts, emotions, and physical symptomatology, were processed with ALCESTE software.
Results. Preliminary findings revealed significant differences by gender and anxiety level, both quantitatively and qualitatively. Textual analysis identified three distinct thematic classes, revealing differentiated patterns in participants39; perceptions and the strategies they employed.
Discussion. The adoption of mixed methods research is emphasized as essential for gaining a deeper understanding of complex phenomena within the field of Behavioral Sciences.

1.6 Session 11: ”Education, Accesibility and Methodological Critique”

Title

Towards a universal design in Psychometrics: Designing an accessible course for students with visual impairments in an online environment

Author(s)
Ariadna Angulo-Brunet , Isabel Duarte-Lores [1] , José Israel Reyes [2]

[1] Universidad de La Laguna; [2] Universitat Oberta de Catalunya

Abstract

The Psychometric course is part of the Psychology degree program at the UOC. This is an online University that operates asynchronously. Students have access to various learning materials (e.g., texts, videos, forums) to study the course content and engage in activities designed to achieve the expected learning outcomes. This course introduces students to the study of a test' s psychometric properties using simulated data that represent a plausible research scenario. The course has two key objectives: first, to provide students with access to a real test manual from a commercial publisher, allowing for an in-depth analysis; and second, to engage them in data analysis using JASP software to draw meaningful conclusions.

From the outset, the course was designed with accessibility in mind, ensuring the use of open-source software that does not require high-performance computers. Additionally, it was important that the software remained actively maintained and included a support channel for troubleshooting and inquiries. Despite efforts to apply universal design principles, the course presents challenges, especially for students with visual impairments. This presentation analyzes the teaching implications of adapting the psychometric course for visually impaired students. First, it examines the specific needs that arise within the course framework, particularly identifying the competencies and learning outcomes that are impacted by the current design. Second, it explores alternative solutions tailored to the online learning environment, considering that certain adaptations used in synchronous, in-person settings may not be feasible. Finally, the presentation introduces and evaluates the implementation of an ad-hoc adaptation in the course, emphasizing the importance of actively involving students with visual impairments in an iterative process to ensure a successful adaptation.
Beyond sharing insights from this teaching experience, this presentation aims to spark discussion on the importance of universal design in course development and the value of incorporating student perspectives, especially when working with diverse student populations.

Title

Methodological Critique in Science: A Geometric and Algebraic Approach to Evaluating Research Quality

Author(s)
Sławomir Pasikowski [1]

[1] University of Lodz

Abstract

The aim of this presentation is to introduce a model of methodological critique based on the analysis of deviations from a locally or generally accepted standard of methodological correctness. A key element of this model is the concept of visibility, which defines the scope of responses on the part of both the sender and the receiver of a scientific statement. The model takes into account the awareness of errors and the methodological rigor within a given discipline, which influences the environmental legitimization of deviations.

The presented model enables the prediction of the behaviour of the creator, recipient, and disciplinary community in which a given scientific statement operates. Furthermore, it allows for an analysis of the persuasiveness of scientific communication through empirical reports. In the literature, methodological critique most often appears in a prescriptive form, whereas the proposed model has an analytical character and allows for its systematic study.

This model has two main versions: geometric and algebraic. The geometric version allows for tracking relationships between key conditions of methodological critique, such as awareness of errors, their identification, and the degree of rigor within the scientific environment. Based on these factors, it becomes possible to predict the communicative situation in which critique occurs. In contrast, the algebraic version enables the objectification of evaluations of scientific statements by analysing the ordering relations within sets of assessments or preferences. This approach allows for the construction of rankings and the estimation of their validity, supporting a more objective assessment of the quality of scientific statements.

The presentation will discuss three main groups of tools used within this model: ordering metrics, a statistical ranking model, and a ranking model based on an ideal vector. Their application will be illustrated through a comparison of the effectiveness of these methods and an analysis of their practical use in scientific evaluation.

In conclusion, the proposed model of methodological critique serves as an innovative tool for analysing scientific communication. Its structure enables a precise determination of the conditions under which methodological critique occurs and facilitates the quantification and objectification of the quality of scientific statements. The results of the conducted analyses may contribute to a better understanding of the mechanisms of acceptance or rejection of specific methodological deviations in different scientific environments and serve as a foundation for further refinement of tools used to assess research quality.

Title

A Comparison of Crossed-Random Investigations of Educational Leadership in Psychology and Economics

Author(s)
Rebecka Persson [1] , Iman Dadgar [1]

[1] Stockholm School of Economics

Abstract

One methodological challenge in social sciences is the understanding of practices in different areas that study the same topics. We address the interdisciplinary understanding between psychology and economy in research about educational leadership. A recent call has been made to extend multilevel modeling (MLM) into econometric work (Oshchepkov & Shirokanova, 2022), which we do with the case of crossed-random school data about grades and teacher ratings of principal leadership- nested both in principals and in schools. The goal of our work is to compare the treatment of crossed random data, and the causal inference claims that can be made, in the typical psychological approach versus the economical approach. Both economists and psychologists study educational leadership and often see for a goal to make causal inference (Höfler et al., 2024, Martin et al., 2021). The different expressions of analysis and conclusions and the terminology in respective tradition do however hinder conversation and knowledge exchange. We aim to promote mutual understanding between the two approaches.

The crossed-random data structure occurs when two different units partly overlap at one and the same level of the data hierarchy. In our example, the overlapping units are schools and principals. Since principals quite frequently change their jobs in our longitudinal data of 3445 principals, they do not entirely overlap, which enables analysis of the magnitude of principal' s influence on their schools. The crossed random context for data in educational leadership research serves as an example of an analysis that would be managed differently with research methods typical to psychology versus economic research. With a psychological approach, as well as in organizational research (Eckardt, 2021), an MLM approach would be employed (e.g, Snijders & Bosker). In the MLM approach, the variance of data at different levels, like students and schools, is explicitly modelled. In the econometric approach, the variance that stems from a hierarchical data structure instead is ascribed to the error variance compound of the model, by means of fixed or random effects (e.g., Böhlmark et al., 2015). We compare the analytical frameworks and the conclusions that can be made to understand whether they differ.

We focus on two aspects for the comparison of analytical procedures. They are estimation methods and the levels of analysis at which interpretations are made. The estimation method in MLM is often maximum likelihood (e.g., Rockwood, 2020) whilst econometric models employ least squares estimation techniques (Greene, 2019). The main procedural difference caused by the different estimation techniques is perhaps the simultaneous estimation enabled my maximum likelihood. The simultaneous estimation considers the interdependencies of parameters which lend accuracy to standard deviations and confidence intervals. Meanwhile economists do rigorous work on each estimate e.g., with weights, that are not easily integrated in the MLM framework. Our comparison addresses the potential consequences of the different estimation techniques through comparison of models in our example data and follow-up simulations. For the levels of analysis at which interpretations are applied in the different analytical frameworks, we discuss feasibility relative to research aims and to policy making.

Title

Instruments for measuring Mathematical Anxiety in the last 10 years: what we know and what we can do.

Author(s)
Laura Barrera Romero [1] , Carlos Fresneda Portillo [1] , Salvador Reyes de Cózar [1]

[1] Universidad Loyola Andalucía

Abstract

Math Anxiety affects students' academic performance and emotional well-being. In recent years, there has been an increase in interest in understanding and addressing this issue, which has led to the development of a wide variety of measurement scales. However, this diversity of instruments has generated a series of challenges for education and psychology professionals.

The objective of this systematic review is to analyze the diversity of scales used in the last decade to measure Mathematical Anxiety to understand the heterogeneity in terms of the conceptualization of the construct, as well as its applicability in different educational contexts, being necessary to study the aspects, elements and dimensions taken into account to define and evaluate Mathematical Anxiety in students.

The review (according to PRISMA) is carried out as follows: the inclusion and exclusion criteria are defined according to the PICO format, focusing on the target population (students with normative development of school age), as well as on the instruments for measuring Mathematical Anxiety that exist for this population. A screening process is then carried out to select the articles that best fit the information being sought.

Once the final articles have been selected, all the information that may be necessary at some point in the review is extracted from each of them, which will attend to the following classification: information on the publication of the article, information on the conceptualization of Mathematical Anxiety and information on its measurement, including in this category both the instrument used and the educational stage of the students and the age range of the sample, among others.

The results of the review, although they confirm the evolution of the instruments for measuring Mathematical Anxiety over the years in the form of concreteness, in each study, according to the population investigated, also pose challenges for professionals who work with students who experience difficulties with mathematics, such as the difficult choice of an appropriate scale for each particular case since its application requires prior knowledge of both of the case of the student and the instrument, or the lack of standardization in the measurement of Mathematical Anxiety, which makes it difficult to compare results between different studies and identify general patterns. This review revealed significant heterogeneity in the conceptualization and measurement of Mathematical Anxiety, underscoring the need for the development and validation of robust, standardized measures that are sensitive to the diverse experiences of students across different educational contexts, posing challenges for researchers and practitioners in achieving consistent and comparable assessments.

In conclusion, the proliferation of scales to measure Mathematical Anxiety, although they reflect a growing interest in its study, also poses important challenges for research and practice, making it necessary to achieve greater standardization in the measurement of the construct, as well as the development of guidelines and criteria for the selection of the appropriate scale in each case, or the creation of a generic scale (previous model) applicable to the target students without the need to attend to specific individual issues as existing scales do.

Title

Methods of music research and its role in the contemporary context

Author(s)
Ligia Farcasel [1]

[1] ”George Enescu” National University of Arts from Iași, Romania

Abstract

In contrast to research in scientific fields, where results are mostly quantifiable, art - music in particular - brings to the fore a science with a high degree of subjectivism, which makes traditional research methods only partially suitable. Here, rather than measuring the object under study, the aim is to understand it with a view to assimilation, because music is concerned with aspects that are less tangible and more closely related to states. The effects of music on the individual and, by extension, on society are far-reaching. However, only an honest perspective of this reality can be the starting point for recognizing the important role of music and, therefore, of music research.

Since the history of music research is relatively young, its methods and techniques are still undergoing a fervent process of development and refinement within this humanistic science which is called 'music science'or 'musicology'. Throughout the centuries, musicologists have tended to practice musicology largely through historical exposition and syntactical and morphological musical analysis. Later, scholars were no longer satisfied with this way of working and began to relate historical musicological analysis more deeply to different contexts and fields, thus bringing the science closer to the understanding of non-musicologists. Thus, the offer of specialized literature is nowadays really rich and constantly growing and is increasingly taking on a general-comprehensive aspect.

So what are the most appropriate methods for musical research? How far can we go in freedom of choice? To what extent can techniques from other fields be applied here? These are some of the questions that shape the intrigue of the current approach. Music research nowadays is more than ever about linking music analysis itself to various wider contexts, be they social, cultural, political, etc. Contextualization is the optimal way to define and achieve meaning in art and ultimately in life. For neither artistic research nor any other kind of research finds its meaning unless it is mirrored in reality. Consequently, the observation of musical phenomena through methods borrowed from diverse fields has become an indispensable practice, which ensures its increased relevance in the vast field of global research. Musicology is no longer a closed discipline, addressed only to connoisseurs, but it is a science with integrative abilities, it has the means to make itself understood from any point of view, and consequently it is proving to be a science of real utility, both in the academic and in the social context.

Title

Effects of Technology-Enhanced Mathematics Learning: A Raw IPD Multilevel Meta-Analysis of Single-Case Experiments

Author(s)
Wim Van Den Noortgate [1] , Nadira Dayo [1]

[1] Faculty of Psychology and Educational Sciences and ITEC, an imec Research Group, KU Leuven, Leuven, Belgium

Abstract

Although single-case experiments (SCEs) are increasingly used in many domains, they are often excluded from meta-analyses, which means that a wealth of information is not utilized. The main purpose of this study was therefore to meta-analyze raw individual participant data (IPD) of SCEs about the effects of technology-enhanced mathematics learning (TEML). A four-level hierarchical meta-analysis was performed, with measurement occasions nested within time series, which are nested within participants, which in turn are nested within studies. The analysis included 166 studies (143 journal articles, 21 theses, 1 book chapter, and 1 conference paper), containing 15,246 measurement occasions, 899 time series, and 587 participants (63% male). In addition to estimating the overall effect, we studied the moderating effects of publication year, author country, document type, age, gender, grade, SCE type, digital tools (tools for outsourcing mathematics, dynamic mathematical tools, data analysis software, program and language tools, extended reality, mathematics content learning platform and artificial intelligence tools), mathematics standards (numbers and operations, algebra, geometry, measurement, and statistics and probability), and disabilities (EBD, IDD, OHI, SLD, and brain injury). The results revealed not only a significant positive immediate effect of TEML but also a significant interaction effect between time and intervention, indicating that after the intervention started, the positive effect increased over time. During the presentation, the moderator effects at both study and participant levels will also be discussed. The meta-analysis of SCEs offers granular and robust effects of TEML at the individual level, accommodating the hierarchal structure of raw IPD. The purpose of the presentation is also to emphasize the value of (the meta-analysis of) SCEs, and more generally to promote methodological diversity and innovation.

Keywords: Single-case experiments, technology-enhanced learning, mathematics, raw IPD, multilevel meta-analysis.

1.7 Session 12 : ”Validity and Reliability in Psychological measurement”

Title

Person- or situation-specific? Factors explaining convergent validity and discrepancy between self-report and digital trace of smartphone use

Author(s)
Martin Tancoš [1] , Michal Tkaczyk [1] , David Smahel [1] , Steriani Elavsky [2]

[1] International Research Team on Internet and Society (IRTIS), Masaryk University; [2] University of Ostrava

Abstract

There is a consensus in the existing literature that self-reported measures provide an inaccurate picture of actual digital behavior. At the same time, there is a high level of heterogeneity in discrepancies between logged and self-reported media use. The prior research concerned with factors explaining this heterogeneity is limited to self-reported measures used in cross-sectional designs, leaving factors related to intraindividual variability in media use and to methodological aspects of repeated measures designs largely understudied. To address this gap, the current study examines the effects of several methodological (duration of the study), contextual (weekend versus weekday), and participant factors (smartphone use, phone controlling efficacy) on convergent validity and accuracy of self-reported measures of smartphone use in 14-day EMA study conducted on the sample 114 adolescents (13 to 17 years old, 57% boys, 825 observations). The accuracy of self-reported smartphone use was lower for adolescents who spend more time using smartphones and on days when participants'smartphone use was less fragmented. Convergent validity of self-report decreased with each day spent in the study. Obtained results support prior findings suggesting that the inaccuracy of self-reports is not solely due to random error but is related to key variables under investigation, such as time spent using smartphones and its characteristics, and this study extends them to repeated measures design. They also show that researchers should address the so-called fatigue effect when designing repeated measures studies to reduce systematic error.

Title

Evaluating the concurrent validity of traditional and alternative measures used in university admission processes in Spain

Author(s)

Susan Niessen [1] , Juan F. Luesia [2] , Milagrosa Sánchez Martín [2] , Stephen Sireci [3]

[1] University of Groninguen; [2] Universidad Loyola Andalucía; [3] University of Massachusetts Amherst

Abstract

Admission processes to access higher education should include the necessary tools to comprehensively measure the competencies and skills related to academic success.
Previous studies have highlighted that relying solely on cognitive tests presents various disadvantages and alternative solutions have been explored, including the application of other types of instruments aimed at capturing not only cognitive competencies. The objective of this study is to analyze the concurrent validity of a classic criterion used in admission tests, such as high school GPA (HSGPA) and cognitive measures, as well as a more innovative instrument called curriculum-sampling tests (CST) that are hypothesized to reflect a combination of cognitive ability, motivation, study time, and tacit knowledge. CSTs have been shown to predict first-year GPA as effectively as HSGPA, but little is known about their relationship to other admissions criteria.
Nine tests were administered to 1,184 applicants during the 2024-25 admission processes for specific degrees at a private university in Spain. Specifically: a) four cognitive tests, which were computed as a single score in cognitive competence, b) four non-cognitive tests, and c) a CST specifically designed.
Of the 1,184 applicants, 369 students enrolled at the university. Correlation and regression analyses were conducted to identify which measures predict HSGPA and CST. The results revealed that: a) predictors of HSGPA were total cognitive score (β=.65, $p < .001$), organizational skills (β=.13, $p = .022$), and self-efficacy (β=.13, $p = .024$); and the model accounted for 47% of the variance; b) predictors of CTS in Psychology were total cognitive score (β=.37, $p < .001$), and self-efficacy (β=.17, $p = .012$); and the model accounted for 17% of the variance. The only predictor of CTS in Pharmacy was the total cognitive score (β=.25, $p = .009$); and the model accounted for 5% of the variance. Predictors of CTS in Medicine were the total cognitive score (β=.29, $p < .001$) and critical thinking disposition (β=.12, $p = .002$); and the model accounted for 9,7% of the variance. No significant predictors of CTS emerged in Nursing, but the total cognitive score showed a trend (β=.12, $p = .077$).
The results of this study show that HSGPA and CST both exhibit strong relationships with cognitive and non-cognitive aspects relevant for success in university. However, for non-cognitive aspects, the relationships are weaker and less consistent. The results suggest HSGPA is useful in admission processes, as it captures cognitive and non-cognitive components of candidates. Similarly, CST appears to be a promising tool to evaluate a combination of cognitive and non-cognitive competencies, although it shows differences across degrees that require further analysis.
This study provides evidence of the importance of including measures in admission tests that capture non-cognitive components of participants, with CST showing great potential. Emphasizing broader admission criteria allows institutions to gather diverse validity evidence to support the intended use of the obtained scores, as well as to gain insights into prospective students that enable universities to implement concrete actions to foster specific competencies or skills.

Title

Reliability and validity of experimental measures of value-driven attention: a meta-analysis on individual differences studies

Author(s)
Pablo López-Mártinez [1] , Francisco Garre Frutos [2] , Juan Lupiáñez [2] , Pablo Solana [2] , Miguel A. Vadillo [3]

[1] Universidad de Málaga; [2] Universidad de Granada; [3] Universidad Autónoma de Madrid

Abstract

Value-driven attention refers to the automatic allocation of attentional resources to reward-associated stimuli, even when such allocation conflicts with task goals. This phenomenon has been proposed as a manifestation of individual differences in “attentional sign-tracking” and has been linked to psychopathological traits. However, the experimental measures commonly employed in this field are difference scores, which often fail to meet psychometric standards. The low reliability in these measures casts doubt on the validity of many findings. To address these concerns, we conducted two meta-analyses. The first one examined the reliability of measures typically employed in studies of value-driven attention. Using 18 publicly available raw datasets, we found that eye-tracking measures consistently demonstrated superior reliability compared to response time measures. This advantage is explained by lower correlations among component measures in eye-tracking data, which reduces measurement error and enhances reliability in differences scores often employed in experimental psychology. Additionally, our analyses indicate that response time measures can attenuate observed effects by up to 31% relative to eye-tracking, even assuming perfect reliability in the other measure. The second meta-analysis assessed the presence of publication bias and the expected replicability rate (statistical power) of the existing literature using z-curve analysis. Analyzing 25 studies (27 contrasts), we estimated a quite low replication rate of only 19%. In conclusion, the present results challenge the validity and reliability of the existing literature on value-driven attention and its individual differences. We advocate for the use of highly reliable measures (such as eye-tracking), alongside pre-registration of hypotheses, adequately powered study designs, and transparent reporting practices.

Title

Multiple Imputation of missing values for randomized controlled trials: A step-by-step tutorial using mice

Author(s)
Victor Ciudad [1] , Oscar Lecuona [2] , Ariadna Angulo-Brunet , Ricardo Olmos [3]

[1] Universitat de València; [2] Complutense University of Madrid; [3] Universidad Autónoma de Madrid

Abstract

Concepual framework: Randomized Controlled Trials (RCTs) are a widely used research protocol in applied research. Among others, a major challenge of RCTs is the presence of missing data due to participant dropout. This leads to loss of power and estimation bias. Multiple imputation (MI) is becoming increasingly popular to deal with missing data in Randomized Controlled Trials. However, MI can produce biased results if not carried out properly. In addition, the required assumptions and steps to develop proper MI for RCTs can be challenging. There is a scarcity of practical guidelines to implement MI for such protocols.
Objectives: In this article we provide a step-by-step tutorial on (1) how to assess missing data in a RCT and how to avoid common misconceptions and pitfalls, (2) how to implement MI in and RCT using the mice package, (3) how to analyze RCT data in the MI framework such as implementing linear models, comparison of effect sizes, and plotting results, and (4) how to develop sensitivity analysis to assess robustness and impact of MI.
Sample: We illustrate this tutorial with a case RCT for wellbeing in social workers (N = 82) comparing two interventions (mindfulness-based intervention, and wellbeing-based intervention) across four measurements (pre intervention, post intervention, 2-month and 4-month follow-ups). Participants were mostly female (92.7%), single (46.3%) and with undergraduate studies (62.2%). Self-report measurements of depression, anxiety, and mindfulness are used as outcomes.
Implications: MI showed stability of the findings and tests used for this dataset, while also the strength of increasing power of conclusions. This tutorial can aid applied researchers to use MI with rigor in their RCT designs. Limitations and extensions of the field are also addressed

Title

Prior sensitivity analysis in Bayesian SEM and its application in R

Author(s)
Mauricio Garnier-Villarreal [1]

[1] Vrije Universiteit Amsterdam

Abstract

There are many research scenarios in which informative (or user-specified) priors have an impact on posterior inference. In addition, diffuse priors have also been found to influence final model estimates in important ways. Given that prior specification has the potential to alter obtained estimates (sometimes in an adverse way), it is always important to assess and report prior impact alongside the final model results being reported for a study. It is important to never blindly rely on default prior settings in software without having a clear understanding of their impact. A sensitivity analysis of priors allows the researcher to methodically examine the impact of prior settings on final results. The researcher will often specify original priors based on desired previous knowledge. After posteriors are estimated and inferences are described, the researcher can then examine the robustness of results to deviations in the priors specified in the original model.

Many Bayesian researchers recommend that a sensitivity analysis accompany original model results. This practice helps the researcher gain a firmer understanding of the robustness of the findings, the impact of theory, and the implications of results obtained. In turn, reporting the sensitivity analysis will also ensure that transparency is promoted within the applied Bayesian literature. Note that there is no right or wrong finding within a prior sensitivity analysis. If results are highly variable to different prior settings, then that is perfectly fine–and it is nothing to worry about. The point here is to be transparent about the role of the priors, and much of that comes from understanding their impact through a sensitivity analysis.

Here we will present a proposed process and specific steps for prior sensitivity analysis in Bayesian Structural Equation Modeling (BSEM) and how to apply it in R. The proposed steps include comparing the model's predictive accuracy with the Leave-One-Out Information Criteria (LOO-IC) and the Widely Applicable Information Criteria (WAIC). Then we will how to compare the model's overall fit, like CFI, SRMR, and gamma-hat, as in BSEM we estimate the posterior distribution of fit indices, we recommend the comparison of the whole posterior between multiple priors. Lastly, we will show how to compare parameters of interest between models with different priors, we will compare the full posterior distribution, point estimates, and variability.

To conclude we will provide recommendations on how to interpret the prior sensitivity comparisons.

Title

Protocol for Developing Validation Studies (PROVAL): A Comprehensive Framework and Illustrative Application

Author(s)
Isabel Benítez Baena [1] , Andrés González [2] , Jose-Luis Padilla Garcia [3]

[1] University of Granada. Mind, Brain and Behavior Research Center (CIMCYC); [2] University of Granada; [3] University of Granada Faculty of Psychology: Universidad de Granada Facultad de Psicologia

Abstract

Validity is one of the most extensively addressed aspects in the literature, as gathering validity evidence is essential to examine interpretations that support the intended uses of the measures. The specific needs of the target instrument, purposes, and assessment context should guide decisions about which sources of validity evidence to prioritize. Most validation studies currently include a pilot phase to evaluate items and test or questionnaire functioning, alongside validity evidence based on internal structure and relationships with other variables. Some also incorporate expert judgment to provide validity evidence based on test content. However, other sources of validity evidence such as those based on response processes and testing consequences, are less frequently explored. Researchers often follow routine procedures without fully reflecting on the most appropriate validity evidence that could provide better support for the intended purpose of measures. This study introduces PROVAL, a nine-step protocol designed to guide researchers in developing validation studies. PROVAL encourages a reflective, tailored approach to instrument validation by helping researchers identify and prioritize the most relevant validity evidence for their specific needs. An illustration of how to apply PROVAL in designing validation studies will be presented, and its contributions to optimizing resources and supporting instrument purposes will be discussed.

1.8 Session 8 : ”Psychometric evaluation in forced-choice tests”

Title

Empirical Evaluation of Psychometric Properties in Forced-Choice Tests: Comparing Binary and Graded Preference Formats Across Test Designs

Author(s)
Rodrigo S. Kreitchmann [1] , Diego F. Graña [2] , Francisco J. Abad [2] , Miguel A. Sorrel [2]

[1] Universidad Nacional de Educación a Distancia; [2] Universidad Autónoma de Madrid

Abstract
Graded preference (polytomous forced-choice) tests have been proposed as a solution to the lower reliability observed in binary forced-choice tests compared to traditional Likert-format assessments. This format retains key advantages, such as larger robustness to social desirability, while increasing the available information by expanding the number of options. However, despite favorable evidence from simulation studies, empirical validation with real data remains scarce. This is particularly relevant because certain response patterns (e.g., if respondents tend to make binary decisions despite the graded forced-choice format), may undermine the expected benefits of graded preferences. Moreover, it remains unclear whether the format change affects test blocks differently depending on whether they are homopolar (equally-keyed) or heteropolar (unequally-keyed). Accordingly, the present study compares graded and binary forced-choice formats against the Likert format under two forced-choice test designs: one using only homopolar blocks and another incorporating 30% heteropolar blocks. Empirical data are analyzed using a Thurstonian Item Response Theory model. Forced-choice tests were developed from a Likert-based personality item bank rated for social desirability, with optimal pairing to maximize expected reliability. We present findings on reliability (marginal and conditional), structural validity (factor loading patterns and model fit), convergent validity, discriminant validity, and criterion-related validity. Results indicate that the graded preference format outperforms the binary format in reliability, particularly in tests composed solely of homopolar blocks. In terms of validity, higher correlations with established Big Five inventories and criterion variables emerged for the graded preference format compared to binary forced choice. In conclusion, the graded preference format enhances the reliability and validity of forced-choice questionnaires, especially when using only homopolar blocks. Given that homopolar blocks naturally facilitate better matching for social desirability, adopting the graded preference format becomes particularly advantageous. Finally, we provide general recommendations for constructing forced choice and graded preference tests.

Title

Statistical foundations of person parameter estimation in the Thurstonian IRT model for forced-choice and pairwise comparison designs

Author(s)
Safir Yousfi [1]

[1] German Federal Employment Agency

Abstract

The statistical foundations of person parameter estimation for the multivariate Thurstonian item response theory (TIRT) model of pairwise comparison and forced-choice (FC) ranking data are elaborated, and several misconceptions in IRT and TIRT are addressed. It is shown that directional information (i.e. multivariate information as defined by Reckase & Kinley, 1991; Applied Psychological Measurement, 15, 361) is not suited to quantify the precision of the estimates unless the Fisher information matrix is diagonal. The asymptotic covariance can be quantified by the inverse Fisher information matrix if the genuine likelihood is used and by the inverse Godambe information for independence likelihood estimation that results from ignoring within-block dependencies of pairwise comparisons. Analytical expressions are provided for the genuine likelihood and the Fisher information matrix for a generalized TIRT model that comprises binary pairwise comparison and ranking designs, which enables maximum likelihood estimation (MLE) and Bayesian estimation (maximum a posteriori probability with normal and Jeffreys prior) of person parameters. The bias of the MLE is quantified, and methods of bias prevention and bias correction are introduced. The correct marginal likelihood of graded pairwise comparisons is provided that might be used for person parameter estimation based on the independence likelihood.

Title

Ipsativity indices for forced-choice assessments

Author(s)
Rodrigo Schames Kreitchmann [1] , Diego Graña Rollón [2] , Francisco J. Abad [2] ,
Miguel A. Sorrel [2]

[1] National University of Distance Education; [2] Universidad Autónoma de Madrid

Abstract

Ipsativity is an important concern in psychological assessment, particularly with forced-choice response formats. It refers to the lack of comparability of test scores between persons, providing information only about the predominance of traits within a person and little or no information about the person's absolute standing in each trait. In other words, while it is possible to correctly estimate a person's difference across traits being measured (i.e., their relative predominance), the person's average or sum across the traits is unidentified. This problem often originates from multicollinearity between trait scores. For instance, in forced-choice formats, endorsing statements associated with a given trait implies not endorsing statements from remaining dimensions, resulting in negative interdependence between scores. As a result, it can largely affect score validity, as the sum of covariances with external variables and the sum of the trait variance-covariance matrix are biased toward zero. In practice, when scores are fully ipsative, the composite/average scores across traits tend to be constant, so the absolute position of examinees is unidentified. However, the amount of ipsativity in forced-choice scores greatly depends on the psychometric characteristics of the blocks of stimuli.

This study proposes two methods to quantify ipsativity/normativity as the proportion of variance of the respondent's true absolute trait standing (composite score) within observed scores. In essence, for a given respondent, the average (or sum) of the posterior covariance matrix of the scores is equivalent to the variance of the person's average (or sum) score across the different traits. Two indices are derived, the first being a theoretical approximation based on the inverse of the test posterior information, and the second being an empirical index, based on posterior trait covariances conditional on the observed response patterns. In essence, low values for these indices suggest ipsativity, whereas high values denote normativity. Furthermore, through this formulation, one can quantify the measurement error due to ipsativity conditional to the trait score. That is, depending on a test's psychometric properties, it may be easier to properly identify the absolute standing of people in different regions of the trait continuum.

A simulation study was conducted to illustrate the two proposed indices. Ipsative data were generated for 5-factor datasets through a binary forced-choice design following the Thurstonian IRT model. Questionnaire lengths ranging from 10 to 60 forced-choice pairs were considered. The accuracy of the indices is quantified as the ability to recover the true ipsativity/normativity (squared correlation between true and estimated person-wise trait averages), as well as its impact on simulated continuous external variables (i.e., on external validity). Two types of external variables were simulated: one with correlations of 0.3 with all the measured traits (i.e., higher correlation with the common variance across traits), and the other with a correlation of 0.3 solely with one of the traits (i.e., higher correlation with the relative score of the corresponding trait). The simulation results support good recovery of both theoretical and empirical ipsativity/normativity indices, as well as a positive relationship between the indices and the estimated correlations with external variables.

Title

A fit index for latent class analysis of dichotomous scale

Author(s)
Pier-Olivier Caron [1]

[1] Université TÉLUQ

Abstract
Latent class analysis (LCA) is a powerful statistical method for identifying unobserved subgroups within a population based on categorical data. However, selecting the optimal number of latent classes remains a challenge and there is no consensus on which fit index to use. Based on the properties of dichotomous variables, this paper introduces a new fit index that capitalizes on the recovery of the model implied covariance matrix from the response probabilities to measure its discrepancy with the sample covariance matrix S. Based on the pattern matrix
Based on the pattern X where each row represents one of the 2^I binary I-tuples, such as
X=[■(x_1,1&⋯&x_(1,I)@⁝&⊠&⁝@x_(2^I,I)&⋯&x_(2^I,I))],
where x_(i,j)∈{0,1} ∀i∈{1,2,⋯,2^I },j∈{1,2,⋯,I}, I is the number of item, the pattern probabilities are
P_i=∑*(k=1)^K*⊠*∏*(j=1)^I⊠ 〖p(x_(i,j)^((k))) c_k 〗 ,
where K is the number of classes and c the class probability, we derived the implied covariance matrix
S(θ)=(XP)^' X-MM^',
where M_j=∑*(i=1)^(2^I)*⊠ 〖*X*(i,j) P_i 〗 .
Using the square difference of the Fisher transform of both covariance matrices, we derived a pseudo χ^2statistic. A Monte Carlo simulation was carried out to compare the accuracy and bias of three versions of this fit index with nine usual fit indices (AIC, BIC, saBIC, χ^2, CAIC, AIC3, Lo-Mendell-Rubin, Vuong-Lo-Mendell-Rubin, and the bootstrap LRT). The simulation shows new among the three versions tested, two had very good properties: less bias and more accurate than other indices. The other one had very good accuracy but tended to narrowly miss the correct number of classes leading excessive over-extraction when it failed. Future developments are discussed, i.e., investigating the asymptotic properties of the underlying pseudo-χ^2 distribution, improving the current criteria and extending the index for ordinal scales.

Title

What ipsative measures can tell us about the General Factor of Personality

Author(s)
Anna Brown

Abstract

Background. The General Factor of Personality (GFP) is a higher-order factor consistently found in personality inventories, explaining correlations between all personality traits in the socially desirable direction. The long-standing controversy about the GFP is whether it is a real thing (individual differences can really be reduced to a single continuum from "bad personality" to "good personality"), or it is an artefact of response biases, most notably socially desirable responding.

I argue that this controversy cannot be resolved while using single-stimulus response formats easily lending themselves to socially desirable responding. Instead, researchers could measure personality with forced-choice response formats, which prevent socially desirable responding and thus are better suited to study the GFP. The use of Thurstonian Item Response Model (TIRT, Brown & Maydeu-Olivares, 2011) further ensures that the scale scores extracted from such questionnaires are normative and are free from ipsative constraints.

Method. This research investigated the construct validity of GFP in two studies, using the same personality inventory - the Customer Contact Styles Questionnaire (CCSQ published by SHL) consisting of 128 items arranged in 32 forced-choice blocks of 4 items - but different external measures for construct validation.

Study 1 recruited N=246 undergraduate psychology students to complete the CCSQ under two conditions –research (low stakes) and selection (high-stakes) condition. The GFP was extracted using the Thurstonian IRT and validated using three measures: the Balanced Inventory of Desirable Responding (BIDR), the Situational Test of Emotional Management (STEM), and the Geneva Emotion Recognition Test (GERT).

Study 2 recruited N=219 call centre employees to complete a validation study using the CCSQ in both the forced-choice and single-stimulus formats. The GFP extracted from both response types was validated using the figures of incentive bonus that the employees received according to their performance.

Results. In Study 1, the GFP accounted for 20.3% and 21.7% of trait variance in low and high stakes forced-choice data, respectively. The GFP extracted from high-stakes forced-choice data correlated with the tendency to manage impression measured by BIDR (r=.24) and with emotional intelligence measured by GERT and STEM (r=.27 and r=0.21). The GFP extracted from low stakes forced-choice data correlated only with STEM (r=0.22).

In Study 2, the GFP accounted for 27.85% and 35.63% trait variance in forced-choice and single-stimulus data, respectively. The GFP extracted from the forced-choice and single-stimulus formats correlated with the incentive bonus (r=.252 and r=.269, respectively). However, these format-specific GFPs correlated with each other only moderately at 0.329, suggesting distinct constructs.

Discussion. Past research has related the GFP to social effectiveness by finding large overlaps with the measures of trait emotional intelligence (van der Linden, Dunkel & Petrides, 2016) and assessment centre ratings (van der Linden, Bakker & Serlie, 2011). Using a personality inventory that combines both the single-stimulus and forced-choice formats, in both high and low stakes, this research found that the GFP has different meanings depending on response formats. The construct validity of the GFP and its relationships with demographics and external constructs are discussed.

Title

Optimal Design in Linear Paired Comparisons for Thurstonian IRT models

Author(s)
Heinz Holling

Abstract

In this talk, we present optimal designs for Thurstonian IRT models based on linear paired comparisons. For this paired comparison type, optimal designs of item pairs are characterized by combinations of those values of factor loadings which optimize predetermined criteria, as correlation between estimated and true trait scores. In order to apply these models in the selection of personnel, only positive factor loadings are admitted. This condition requires the development of novel types of optimal designs. Beyond properties of optimal designs developed in the literature so far, two more requirements have to be particularly taken into account: (a) the restriction of the design region, and (b) the constraint that alternatives have to load on mutually distinct factors, respectively. In this talk, we present solutions for the optimal design problem which substantially outperform current methods in the literature in terms of precision and sample size required. These results will carry over to Thurstonian IRT models with binary or ordinal response.

1.9 Symposium : ”Intervention programs evaluation: effect size, moderator variables and methodological quality”

Title

Convergent-discriminant validity evidence of the Methodological Quality Scale for Observational Methodology: A multitrait-multimethod analysis

Author(s)
Daniel López-Arenas , José Mena Raposo [1] , Salvador Chacón Moscoso [2] , Susana Sanduvete-Chaves [1]

[1] Universidad de Sevilla; [2] Universidad de Sevilla, Spain; Universidad Autónoma de Chile, Chile

Abstract
Introduction: designs based on observational methodology allow the systematic recording and subsequent quantification of the spontaneous behavior displayed by participants in natural contexts. These research methods are frequently used in psychology, as well as in the social, educational and health fields due to their multiple advantages, such as a low level of intervention, independence with respect to standardized measurement instruments or their flexibility when applied in non-standardized intervention contexts. A Methodological Quality Scale for Studies Based on Observational Methodology (MQSOM), a tool to measure the methodological quality of these studies, has recently been validated with adequate psychometric properties (RMSEA = 0.000, NNFI = 1, GFI = .98, AGFI = .97). The MQSOM comprises a second-order factor of Methodological quality (ω = .87; D = .55) containing two first-order factors: Quality of design (6 items; ω = .90; D = .46; ICC = .933 - .967) and Quality of measurement and analysis (5 items; ω = .68; D = .67; ICC = .797 - .988).
Objective: the aim of this study is to present the evidence of convergent and discriminant validity of MQSOM.
Methods: a multitrait-multimethod analysis (MTMM) with Spearman correlations was carried out to examine the relationship between the dimensions of MQSOM and those of the methodological quality instruments Rigorous Mixed-Methods (RMM), Guidelines for Publishing Evaluations Based on Observational Methodology (GREOM) and Mixed Methods Appraisal Tool (MMAT), circumscribed to the field of Mixed-Methods studies. Ninety-six articles based on observational methodology were coded with MQSOM and each of the contrast instruments.
Results: adequate levels of inter- and intra-coder reliability were obtained (ICC between .73 and 1). MQSOM dimension of Design showed empirical evidence of convergence with MRMM (ρ between .22 and .47), GREOM (ρ between .22 and .34) and MMAT (ρ = .21). It also showed empirical evidence of discriminant validity with the contrast instruments (ρ between -.05 and .03 regarding MRMM; ρ between -.03 and .03 regarding GREOM; ρ = -.04 regarding MMAT). MQSOM dimension of Measurement and Analysis showed empirical evidence of convergence with MRMM (ρ between .21 and .61), GREOM (ρ between .22 and .61), and MMAT (ρ between .21 and .64). Conclusions: these results support the use of MQSOM, a brief instrument that addresses methodological quality in observational methodology in a diagnostic way, measuring the quality of design, measurement and analysis of results in studies based on observational methodology, but also in a prescriptive way, serving as a reference for applied researchers, editorial boards and other decision-making committees.
Funding: This work was supported by the research project PID2020-115486GB-I00 funded by the Ministerio de Ciencia, Innovación y Universidades, MICIU/ AEI/10.13039/501100011033, Government of Spain; and the Chilean government project FONDECYT Regular 1250316 funded by the National Fund for Scientific and Technological Development, ANID.

Title

The goodness of fit indexes RMSEA and SRMR using ULS and RULS in Structural Equation Modeling: a review of its cut-off point

Author(s)
Julia Sánchez García , Francisco Pablo Holgado Tello , José Mena Raposo [1] ,
Juan Carlos Suárez Falcón

[1] Universidad de Sevilla

Abstract

The use of Likert scales in the field of social research is becoming more and more common every day, it is necessary to investigate which is the most appropriate methodology to carry out the analysis of the data obtained. If they are ordinal, they should be treated as such, however, they are frequently analyzed considering them as continuous variables. One of the most widely used techniques to obtain construct validity evidence through internal structure of the nomological models, is Confirmatory Factor Analysis. Using simulation studies in which four factors have been manipulated (number of factors, number of items response categories, skewness and sample size) our objective is twofold: firstly, when ordinal variables are used, analyze the type I error and power of the most common fit indices, such as RMSEA and SRMR obtained using ULS and RULS estimation methods; and secondly, using Receiver Operating Characteristic Curve (ROC) review the cut-off points of RMSEA and SRMR. It is found that, depending on the estimation method chosen, the type I error and power differ, as well as the values reported by RMSEA and SRMR. RULS seems to obtain better results regardless of experimental factors manipulated. Finally, it is found that it would be convenient to review the cut-off points for these global fit indices recommended by the literature.

Title

Training program outcomes for mental health professionals: The role of methodological quality, study type, and timing. A meta-analysis

Author(s)
José Lozano Lozano , José Mena Raposo [1] , Salvador Chacón Moscoso [2] ,
Susana Sanduvete-Chaves [1]

[1] Universidad de Sevilla; [2] Universidad de Sevilla, Spain; Universidad Autónoma de Chile, Chile

Abstract

This study aims to analyze the effectiveness of training programs designed for mental health professionals. The analysis focuses on randomized controlled trials (RCTs) and cluster-randomized studies, examining the impact of these interventions across three levels of outcomes (based on Kirkpatrick & Kirkpatricks'model): knowledge acquisition, attitude changes, and behavioral modifications. The study includes 18 eligible studies, each meeting rigorous inclusion criteria, and evaluates the moderating effects of methodological quality, study type, and intervention duration. Methodological quality was assessed using the 10-item Methodological Quality Scale, providing a standardized measure to gauge the robustness of the included studies. The analysis further investigates the differential effects of research design studies (RCTs versus clusters) and intervention
and measurement times. Three distinct meta-analyses were conducted to integrate the outcomes across the selected levels. Preliminary findings suggest a positive overall effect size, with decreasing magnitude observed as the analysis progresses from knowledge to attitudes and, ultimately, to behaviors. These results align with the hypothesis of diminishing returns through the hierarchical pyramid of training impact. This work underscores the critical importance of methodological rigor and contextual factors in determining the efficacy of training programs in mental health services. Insights from this analysis provide actionable evidence to enhance future program design, implementation, and evaluation.

Funding: This work was supported by the research project PID2020-115486GB-I00 funded by the Ministerio de Ciencia, Innovación y Universidades, MICIU/AEI/10.13039/501100011033, Government of Spain; and the Chilean government project FONDECYT Regular 1250316 funded by the National Fund for Scientific and Technological Development, ANID.

Title

Risk of bias in clinical psychology meta-analyses (2000-2020): An overview

Author(s)
Julio Sánchez-Meca [1] , Alejandro Sandoval-Lentisco [2] , Jose Antonio Lopez Lopez

[1] University of Murcia (Spain); [2] Universidad Autónoma de Madrid

Abstract
One of the biggest limitations of meta-analyses is that the information they provide can be affected by the biases of the included primary studies. To address this, evaluations of primary study risk of bias (RoB) can be performed and incorporated into the meta-analysis. However, research on this topic in clinical psychology is scarce. In this study, we examined this issue using a sample of clinical psychology meta-analyses that included RoB assessments. First, we evaluated meta-analysts'assessment practices. Second, we summarized the RoB ratings of the primary studies included in the meta-analyses. Lastly, we examined the relationship between RoB ratings and effect sizes. We found some suboptimal practices in the assessment procedures, such as only half of the studies reporting that the assessment was conducted in duplicate. Regarding RoB ratings, the domains with the highest ratings were random sequence generation, blinding of outcome assessment, and incomplete outcome data, with about half of the primary studies rated as low RoB. The lowest ratings were found for allocation concealment and, especially, blinding of participants and personnel. Importantly, we found a positive association between the publication year of the primary studies and a lower RoB in most domains. Lastly, performing our own re-analysis, we found an association between RoB and effect sizes, which contrasts with the results of the analyses reported in the meta-analyses that combined those studies. We recommend caution when interpreting a lack of modulation of effect sizes in meta-analyses, as they may not have sufficient statistical power for moderator analyses.
Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NBI00

Title

Effectiveness of psychoeducation on myositis: Quality of life and well-being

Author(s)
María Palacín Lois , Inma Armadans Tremolosa , Angela Castrechini Trotta , Albert Selva O'Callaghan , Salvador Chacón Moscoso [1] , Susana Sanduvete-Chaves [2]

[1] Universidad de Sevilla, Spain; Universidad Autónoma de Chile, Chile; [2] Universidad de Sevilla

Abstract

Background: This study investigated the effectiveness of a psychoeducational intervention on the quality of life and well-being of patients with myositis, a rare condition that significantly impacts daily life. Methods: All myositis patients in a specific healthcare region were invited to participate. Thirty-four eligible patients were randomly assigned to either an intervention group or a control group. The intervention group received five 100-minute sessions focused on understanding how myositis impacts daily life. Both groups were assessed before and after the intervention using validated tools to measure quality of life, well-being, and self-efficacy in managing the disease. Results: Patients in the intervention group showed improvements in quality of life, well-being, and self-efficacy compared to their pre-intervention scores. These improvements were more pronounced in the intervention group compared to the control group for 70% of the variables studied. Notably, the intervention group experienced a greater reduction in sedentary behavior and an increase in satisfaction with social relationships. Conclusions: This randomized controlled trial, conducted on a representative sample of myositis patients, provides evidence that a psychoeducational intervention can effectively improve healthrelated quality of life, well-being, and self-efficacy in managing myositis. Funding: This study was funded by the Instituto de Salud Carlos III (grants PI22-00708), co- financed by the European Regional Development Fund; the research project PID2020-115486GB-I00 funded by the Ministerio de Ciencia, Innovación y Universidades, MICIU/AEI/10.13039/501100011033, Government of Spain; and the Chilean government project FONDECYT Regular 1250316 funded by the National Fund for Scientific and Technological Development, ANID.

Title

Validity evidence of the Hospital Anxiety and Depression Scale (HADS) in Chilean patients with chronic kidney disease

Author(s)
Jose Antonio Lozano Lozano , Erica Villoria , Francisco Pablo Holgado Tello ,
Salvador Chacón Moscoso [1] , Susana Sanduvete-Chaves [2]

[1] Universidad de Sevilla, Spain; Universidad Autónoma de Chile, Chile; [2] Universidad de Sevilla

Abstract

Background: Chronic kidney disease (CKD) is a global health issue that significantly impacts patients'quality of life due to physical and emotional symptoms. Anxiety and depression are common in these patients, negatively affecting their prognosis and treatment adherence. The Hospital Anxiety and Depression Scale (HADS) is a popular tool for assessing these disorders, but it has not been validated in Chilean renal patients. Methods: In a sample of 442 CKD patients from hospital centers in Chile, the factor structure, internal consistency, and concurrent validity of the HADS were evaluated using confirmatory factor analysis, Cronbach's alpha, McDonald's omega, and correlations with the Depression Anxiety Stress Scale (DASS-21), respectively. Results: Analyses showed a good fit for the two correlated factors model, with anxiety and depression subscales demonstrating high internal consistency. Significant correlations between HADS and DASS-21 confirmed concurrent validity. Conclusions: These findings suggest that the HADS is a valid and reliable tool for assessing anxiety and depression in Chilean CKD patients, facilitating timely psychological interventions and improving patients'quality of life. Future studies should include more diverse samples and assess the temporal stability of the scales to confirm these findings.

Funding: This work was supported by the research project PID2020-115486GB-I00 funded by the Ministerio de Ciencia, Innovación y Universidades, MICIU/AEI/10.13039/501100011033, Government of Spain; the Chilean
government project FONDECYT Regular 1250316 funded by the National Fund for Scientific and Technological Development, ANID; and Fondecyt Iniciación 11200099, Government of Chile.

1.10 Symposium : ”Recent Developments in Meta-Analysis”

Title

Can We Include Dichotomous Variables in Meta-Analytic Structural Equation Modeling? Mind the Prevalence

Author(s)
Suzanne Jak [1] , Belén Fernández-Castilla , Hannelies de Jonge [2] , Kees-Jan Kan

[1] University of Amsterdam; [2] Methodology & Statistics, Psychology, Leiden University

Abstract

Meta-analytic structural equation modeling (MASEM) is a method to systematically synthesize results from primary studies, allowing the researchers to simultaneously examine multiple relations among variables by fitting a structural equation model to the pooled correlations. Incorporating dichotomous variables (e.g., having
a specific disease or not) into MASEM poses challenges. While primary studies that investigate the relation between a dichotomous and continuous variable typically report standardized mean differences (e.g., Cohen's d), in the specialized MASEM software it is not possible to directly include standardized mean differences. Instead, MASEM typically uses correlation matrices as input. A proposed solution is to convert the standardized mean differences to point-biserial correlations. Here lies a complication because, in contrast to a standardized mean difference, the point-biserial correlation depends on the distribution of group membership. Through three Monte Carlo simulation studies, we investigated which conversion formula is suitable when one wants to include a dichotomous variable in MASEM. We varied the prevalence, sampling plan, and within-study sample sizes, and the distribution of participants over two groups. Our results show that which conversion is suitable, and which is not depends on the aim of the meta-analyst. We have extended our freely available web application to fill the existing gap and to assist the meta-analyst with their conversions.

Title

Fitting two-level structural equation models to meta-analytic data

Author(s)
Suzanne Jak [1] , Mike W. L Cheung

[1] University of Amsterdam

Abstract
In a recent paper we presented a way of incorporating mean structures in meta-analytic structural equation modeling (MASEM). MASEM with means is applicable when the studies included in the meta-analysis used the same indicators, measured on the same scales. The meta-analytic data consist of the studies'covariance matrices and mean vectors. The MASEM then restricts the vector of meta-analyzed means and covariances to the structure of the hypothesized SEM, and quantifies the heterogeneity of the model-implied covariances and means across studies. In this presentation we explain how the heterogeneity matrix of the model implied means can be interpreted as what is often referred to as ΣBETWEEN in two-level SEM, while the model-implied pooled covariance matrix can be interpreted as ΣWITHIN. We illustrate how to fit SEM models to the heterogeneity matrix of the model implied means in the R-package OpenMx, and compare the results with those obtained from fitting two-level models directly on raw data in lavaan. These new modeling options have implications for meta-analytic research (e.g., extending the range of models that can be evaluated) as well as for two-level SEM (e.g., fitting models on summary statistics, flexibility in adding random effects)

Title

Reporting Biases Analyses in Psychological Meta-analyses: Current Practices and Robustness of Conclusions

Author(s)
Rubén López Nicolás [2] , Alejandro Sandoval-Lentisco [1] , Miguel A. Vadillo [1]

[1] Universidad Autónoma de Madrid; [2] Universidad de Castilla la Mancha

Abstract

Reporting biases are well-known phenomena that can undermine the credibility of published scientific findings and potentially distort meta-analytic effect estimates. These biases arise when the decision to publish or report results is influenced by their nature or direction. Traditionally, methods for assessing small-study effects and evaluating the robustness of results against publication bias have been widely used to address this issue. However, in recent years, novel approaches to detecting and correcting for reporting biases have emerged and gained attention. The proliferation of methods for assessing reporting biases presents challenges, as their sensitivity, specificity, and accuracy can vary under different conditions, with no single method consistently outperforming others under all conditions. Consequently, the wide availability of alternative methods could introduce researcher bias into these analyses, creating a paradox where reporting bias may itself be present in reporting bias assessments.

This project has two main aims. First, we investigated current practices in reporting bias analysis among recent meta-analyses. Second, we examined the potential impact of the variety of available approaches for reporting bias analyses on the robustness of conclusions.

We included meta-analyses published in Psychological Bulletin from January 2020 to May 10, 2024. The articles selected met the following criteria: (a) they included at least one meta-analysis, (b) they were not re-analyses of previously published meta-analyses, and (c) their unit of analysis in the synthesis was primary studies. Additionally, for analyses related to the second aim, meta-analyses had to meet the following criteria: (a) the original data of the meta-analysis had to be openly available in a machine-readable format, (b) the meta-analysis had to be based on traditional standardized effect sizes for group differences or bivariate associations, and (c) reporting bias had to be assessed in the original meta-analysis. We collected data on the prevalence of reporting bias assessments, the methods used to assess reporting bias, the number of reporting bias methods applied, whether reporting bias assessments were pre-registered, any deviations from pre-registered protocols, and the conclusions reached regarding the presence of bias. Second, for the subset of meta-analyses meeting the second aim's criteria, we reanalyzed their primary data using a set of pre-registered methods. Then, the number of methods indicating the presence of bias, based on pre-registered criteria, were counted and these results were grouped according to the original conclusions reached.

Title

Comparing Type I error and power rates in meta-regression with multiple effect sizes: A study of analytical approaches

Author(s)
Belén Fernández Castilla [1] , Jose Antonio Lopez Lopez , María Rubio Aparicio [2]

[1] Universidad Nacional de Educación a Distancia; [2] University of Murcia

Abstract

Moderator analyses play a crucial role in meta-analysis, as they help to identify relationships between study characteristics and the effect size magnitude. When multiple effect sizes are reported within studies, various methods can be used to perform moderator analysis or meta-regression. These include three-level models (which may or may not account for variability in moderator effects across studies), Robust Variance Estimation (RVE) methods (with or without the wild bootstrapping technique), and multilevel models combined with RVE. In this study, we conducted a simulation to compare the performance of these methods in terms of Type I error rates and statistical power when performing meta-regressions, focusing specifically on qualitative moderator variables (such as study design or sample type). This focus arises from the common occurrence of unbalanced effect size distributions across moderator categories (i.e., most effect sizes belong to one category, while few belong to others), and it remains unclear which method performs best under these conditions. Additionally, we provide an empirical example of how these differences among methods affect real meta-analyses.

To simulate typical meta-analyses, we generated standardized mean differences under varying conditions, such as the number of studies, effect size differences across moderator categories, and average outcome numbers, among others. We analyzed qualitative variables with two or three categories to represent study or effect size characteristics, and the effect sizes were distributed in balanced, unbalanced, or highly unbalanced ways across moderator categories. When simulating three categories, we also used Tukey's multiple comparison correction to assess differences across categories.

Results showed that when the qualitative variable referred to effect size characteristics, the three-level model that did not account for moderator effect variability (the one commonly implemented in practice) had highly inflated Type I error rates, while other methods maintained acceptable rates. Power levels were generally lower when the moderator referred to effect size characteristics, and these were minimally affected by unbalanced effect size distributions across categories. When the moderator referred to study characteristics, all methods exhibited acceptable Type I error rates, but power was inadequate, particularly when effect sizes were highly unbalanced. Across all conditions, three-level models combined with RVE provided the best Type I error-power balance, although power remained very low.

In conclusion, this study suggests that, in the presence of multiple effect sizes within studies, multilevel models should always be applied with RVE correction when conducting meta-regressions. Additionally, further advancements are needed to generally improve power for detecting moderator effects.

Title

Correcting for Publication Bias in Moderator Effects: A Simulation Study

Author(s)

Robbie van Aert , Franziska Rüffer [1] , Jelte M. Wicherts , Marcel van Assen

[1] Tilburg University

Abstract

Moderator analysis in meta-analysis is commonly used to study whether certain study characteristics can explain the heterogeneity in effect sizes. Understanding why effect sizes vary between contexts is important for selecting the right intervention for the right context and for guiding further research. In order to rely on the results from moderator analyses, the moderator effect estimates need to be unbiased. When publication bias is present, this cannot be guaranteed. Previous research has demonstrated that moderator effects in (mixed-effects) meta-regression may be both, under- or overestimated, depending on the characteristics of the meta-analysis. In practice, one would not only like to understand the influence of publication bias on moderator effects but also how to correct for it. For this purpose, we have conducted an extensive simulation study to assess how well publication bias models can account for publication bias in moderator effect estimates. In total, 1,728 simulation scenarios were generated by varying the true effect sizes, the amount of heterogeneity, the number of studies in the meta-analysis, the primary study sample sizes, and the amount and type of publication bias. We focused on generating estimates from a meta-regression model with either a single binary or a single continuous moderator using the conventional mixed-effects meta-regression model that does not correct for publication bias and different publication bias models. The included publication bias models were step function selection models (Hedges, 1992; Hedges & Vevea, 1996), PET and PEESE and PEESE MRA which allows for different amounts of publication bias at each level of the moderator (Stanley, 2008; Stanley & Doucouliagos, 2014). In this talk, we will present the main results from this simulation study and give recommendations on which of these models can correctly account for publication bias in meta-regression analysis and in which contexts they are applicable.

Title

Correcting for publication bias in multivariate and multilevel meta-analysis: A multivariate step function selection model approach

Author(s)

Robbie van Aert

Abstract

Univariate meta-analysis models assume that all effect sizes included in the meta-analysis are independent. This assumption is violated if, for example, two outcomes are reported in a study that are of interest to the meta-analyst or a study reports multiple experiments administered by the same researchers in the same lab. The multivariate and multilevel meta-analysis model allow to model dependent effect sizes and these models have recently gained in popularity among meta-analysts in psychology.

One of the largest threats to multivariate and multilevel meta-analysis is publication bias, but there are currently no methods available that correct for publication bias in these models. Selection model approaches are nowadays frequently used to correct for publication bias in a meta-analysis. In this presentation, we extend the univariate step function selection model approach to multivariate and multilevel meta-analysis. We propose a strict and more relaxed selection model that assigns a different publication probability to studies that have only statistically significant outcomes or at least one significant outcome.

We illustrate how the multivariate step function selection model approach can be used in a sensitivity analysis by applying it to the data of a published multivariate and multilevel meta-analysis. Two simulation studies tailored to these two applications show that the multivariate step function selection model approach outperforms the multivariate and multilevel meta-analysis model that do not correct for publication bias. We conclude this presentation with offering guidance for applying the proposed method in practice and discussing limitations of the method as well as opportunities for future development.

1.11 Poster Session 1

Title

Conditions of education and methodological preferences of young researchers - results from preliminary studies

Author(s)
Martyna Jarota [1]

[1] Department of Educational Research, Faculty of Educational Sciences, University of Lodz, Poland

Abstract
Academic education holds a special place in acquiring knowledge and skills in research methodology. According to studies (Papanastasiou & Papanastasiou, 2004), its quality has a significant direct impact on the formation of attitudes towards science. Furthermore, it is assumed that methodological preferences may be a consequence of university education (Szmatka et al., 1996). Therefore, from the perspective of observing the development trends of individual scientific disciplines, it is important to address the issue of methodological preferences and the conditions of education (including academic education) that may shape them. The main research question of the study was: How do the relationships between the conditions of education and the methodological preferences of young researchers develop? To answer this research question, partially structured interviews were conducted with Polish PhD students representing social science disciplines. A decision was made to use team-based random sampling to recruit respondents for the study. The analysis of the collected qualitative data was conducted with the support of computer software.

Title

Visual Strategies in Educational Assessment: An Analysis of Eye Movements in Multiple-Choice Item Resolution

Author(s)
Susana Sanz [1] , Miguel A. Sorrel [2] , Maria Pilar Aivar [2] , José David Moreno Pérez [2]

[1] San Pablo CEU University; [2] Universidad Autónoma de Madrid (UAM)

Abstract

Background and objectives:
Multiple-choice items are widely used in assessment contexts as they enable efficient sampling of a broad range of content. However, not all students employ the same strategies when answering tests. Besides their level of knowledge on the subject being assessed, their response behavior can influence performance, making it essential to categorize these behaviors to better understand the response processes. In this regard, the study of eye movements has proven useful in related fields such as reading comprehension. This project aims to analyze information processing in multiple-choice items using eye-tracking techniques while students solve these tasks. Three objectives are proposed: 1) To examine eye movement patterns in assessment contexts using multiple-choice items, establishing categories based on information processing strategies; 2) To explore the relationship between eye movement measures and psychometric properties of items; and 3) To analyze the effects of violations of item-writing guidelines, which may impact response processing, and determine whether general processing strategies change depending on these violations.
Methods:
Eye movements during the reading of multiple-choice items will be recorded using an EyeLink 1000 Plus eye-tracker. Considering the characteristics of eye-tracking research, in which each participant is measured multiple times in a very precise way, the sample size calculated for the required statistical modeling and a minimum power of 0.80 was 43 participants. Data collected will be analyzed using multilevel models, in which we will consider two random factors (item and participant), and different fixed factors according to the objectives of each study. Separate models will be developed for each dependent variable analyzed, considering accuracy rates and different eye movement measures. Random intercepts and slopes for participants and items will also be included as random effects.
Proposed studies and expected contributions:
Study 1. Establishing a baseline framework for processing under standard assessment conditions: This study will develop a general classification, grouping participants' processing styles through eye-movements when facing multiple-choice items.
Study 2. Relationship between psychometric properties and eye movement patterns: This study will examine the relationships between eye movement measures and item characteristics, providing insights into how items are processed based on their difficulty, the selection rates for distractors, and the discrimination of options, contributing validity evidence for response processes when answering multiple-choice items.
Study 3. Effect of including item-writing guidelines'violations: This study will assess the impact of some guidelines'violations on the psychometric properties of items.
Study 4. Relationship between processing level and item-writing guidelines'violations: This study will further investigate the relationship between participants' processing strategies and some item-writing guidelines'violations, offering conclusions on differences in item difficulty and discrimination. This will inform improved multiple-choice item design, supported by validity evidence against non-recommended item-writing practices.
Conclusions:
This is an emerging field of study for which no empirical evidence currently exists in scientific literature. The goal of this project is to establish the foundations of the relationships between eye movements and the resolution of multiple-choice items, enabling the design of more reli-

able and valid knowledge assessment tools.

Title

Linking Evidence-Based Architectural Research Methodologies with Medical/Health Sciences: A Qualitative and Mixed-Methods Framework for Infection Prevention in Healthcare Facilities

Author(s)
Tomislav Meštrović [1] , Marko Jelovac [2]

[1] Health Metrics Centre, University Centre Varaždin, University North, Croatia / Institute for Health Metrics and Evaluation, US / University of Washington School of Medicine, Seattle, US;
[2] Faculty of Architecture of the University of Belgrade, Serbia

Abstract

Infection prevention in the built environment is a growing concern, most notably in healthcare facilities where disease transmission risks can be high. Consequently, addressing this challenge through informed architectural design requires a systematic, interdisciplinary and evidence-based research methodology that integrates insights from both architecture and medical/health sciences. Here we explore how qualitative and mixed-methods research methodologies enhance objectivity in architectural decision-making for infection prevention, ensuring that interventions are not only scientifically sound, but also contextually relevant and user-centered.

Qualitative research methodologies provide a deep, contextual understanding of how spatial configurations influence not only infection risk, but also user behavior and adherence to hygiene protocols. Through ethnographic studies, detailed case studies, interviews and spatial observations, architects can assess how users interact with different architectural features such as ventilation systems, sanitation infrastructure and circulation patterns. Moreover, expert input from medical and nursing professionals, as well as public health experts, leads to more structured and more streamlined in architectural decision-making for infection prevention, ensuring that interventions are up-to-date and in accordance with their work needs. These methods also help in identifying social and psychological barriers to effective infection control, ensuring that design solutions are not solely driven by technical considerations, but also by actual human needs and behaviors. Socio-spatial mapping techniques can help identify unseen risks within building (such as high-touch areas or overlooked transmission pathways) that conventional quantitative models may fail to capture.

A mixed-methods approach further strengthens objectivity by integrating qualitative insights with structured assessment tools, such as post-occupancy evaluations and pattern analysis of high-risk infection zones. By triangulating multiple data sources –including epidemiological evidence, behavioral mapping and material performance assessments –this approach ensures that architectural strategies are grounded in empirical evidence. Additionally, mixed-methods research allows for iterative design improvements based on continuous feedback from healthcare professionals, facility managers and end-users, leading to a responsive and adaptive approach to infection prevention.

By emphasizing qualitative rigor and methodological integration, the interdisciplinary potential of architectural research in developing infection-resistant environments becomes very evident. Rather than relying solely on subjective decisions or prescriptive guidelines, this approach enhances objectivity by capturing the complexities of real-world architectural use. In our view, this is a way to reframe architecture as an epidemiological tool, where spatial configurations and material choices are viewed not just as passive design elements, but as active agents in disease transmission dynamics. Informed by qualitative and mixed-methods research, such approach can lead to resilient, adaptive, sustainable and health-centered space, ensuring that infection prevention measures and workforce needs are embedded in both architectural planning and everyday practice.

Title

Speech Analysis Module: An Open-Source Audio Processing Library for Onset Detection and Stimuli Preparation in Psycholinguistics

Author(s)
Emma Rico Martín [1] , Alberto Domínguez Martínez [1] , Agustina Birba [2] , Beatriz Bermúdez Margaretto [3]

[1] Instituto Universitario de Neurociencia, Universidad de La Laguna, Tenerife, Spain; Departamento de Psicología Cognitiva, Social y Organizacional. Universidad de La Laguna, Tenerife, Spain.; [2] Cognitive Neuroscience Center. University of San Andrés, Vito Dumas 284, B1644BID, Victoria, Buenos Aires, Argentina; [3] Departamento de Psicología Básica, Psicobiología y Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain; Instituto de Integración en la Comunidad - INICO, Universidad de Salamanca, Salamanca, Spain.

Abstract

Introduction:
Language research has gained increasing importance in recent years, especially with the rise of AI. Precise measurement of naming reaction times is critical in psycholinguistic research, particularly in experimental psychology paradigms involving spoken responses. Although some approaches accurately detect voice onset, most fail or provide less accurate predictions in the presence of noise or poor-quality signals, and some are not free. Another common requirement in psycholinguistic research is the preparation of auditory stimuli, along with other essential tasks such as subject response transcription, preprocessing, and standardization. Here we present a novel Python-based audio processing library that integrates several open-source libraries—such as Librosa, SciPy, NumPy, Matplotlib and Whisper—to facilitate robust onset detection and stimuli preparation. To evaluate its performance, we conducted a naming experiment, demonstrating the library's effectiveness in accurately estimating reaction times from spoken responses.
Methods:
The library encompasses a comprehensive suite of functions tailored to experimental settings. Key functionalities include:
• Stimuli Synchronization: Audio files are systematically retrieved and prepared using functions that trim silence (via Librosa's trimming algorithms), adjusting the duration with precise zero-padding. This is useful for stimuli preparation, particularly in neuroscience research where precise synchronization with brain data is essential.
• Signal Filtering: A Butterworth bandpass filter can be applied to refine the audio signal by reducing noise and isolating specific frequency bands typically associated with the human voice.
• Noise Gating: A dynamic noise gate further enhances signal quality by attenuating background noise below a predetermined threshold.
• Onset Detection: After applying signal filtering and noise gating, the library leverages Librosa's onset strength envelope to identify transient events. By computing the energy envelope of the audio signal and applying backtracking and sensitivity adjustments, this approach ensures accurate detection of speech onsets, crucial for measuring reaction times.
• Visualization and Spectral Analysis: The library offers visualization of waveforms and spectrograms, allowing researchers to verify signal quality and processing outcomes.
To evaluate the effectiveness of this approach, we compared the accuracy of the library's automatic onset detection in a naming experiment with manual selection using Praat visualization. Additionally, we provide an example of audio preprocessing in a natural speech inhibition task.
Results:
Analysis revealed that the automatic naming reaction time estimator performed robustly, with an absolute difference of 38.29 ms and an R^2 of 0.87 when compared against manual measurements across 322 audio files. These metrics indicate high accuracy and reliability in capturing

the temporal dynamics of spoken responses.
Conclusions:
This open-source audio processing library offers a versatile, transparent, and effective tool for psycholinguistic research. By integrating well-established Python libraries, it simplifies the complex process of audio signal processing—from onset detection to stimuli preparation—thus reducing manual effort and increasing measurement precision in language research. The encouraging results obtained underscore its potential as a valuable resource for experimental psychology, enabling researchers to derive meaningful insights from auditory data with enhanced accuracy.

Title

A Comparative Study of Dynamic Structural Equation Modeling and Structural Equation Modeling in Longitudinal Actor-Partner Interdependence (APIM) Mediation Analysis

Author(s)
Shuncheng He , Wooyeol Lee

Abstract

The Actor-Partner Interdependence Model (APIM) is a widely used framework for analyzing dyadic data, which enables the capture of both actor and partner effects. Recently, the APIM has been extended to assess mediation effects in dyads. The APIM is traditionally implemented within the framework of Structural Equation Modeling (SEM), APIM assumes static relationships, which may overlook the temporal dynamics inherent in dyadic interactions. Furthermore, when applying SEM to longitudinal APIM data, researchers often simplify models to fit the data, rather than tailoring the model to the research question (Planalp et al., 2017; Savord et al., 2023).

Dynamic structural equation modeling (DSEM; Asparouhov & Muthén; 2019) has emerged as a promising alternative by integrating time series analysis, SEM, and multilevel modeling to capture the dynamic and multilevel structure of the data. Fang et al. (2024) proposed a method to investigate examining mediation effects in intensive longitudinal dynamic data within the DSEM framework and showed that residual DSEM (RDSEM) outperforms traditional DSEM in de-trending mediation analysis. However, few studies have applied the DSEM framework to APIM, and Savord et al. (2023) only demonstrated the feasibility and characteristics of APIM within DSEM through the utilization of illustrative data in Mplus, without conducting a comprehensive comparison between the SEM and DSEM frameworks in the context of APIM data.

The purpose of this study is to compare SEM, DSEM, and RDSEM within the APIM mediation analysis framework using Monte Carlo simulations, with a focusing on the accuracy of parameter estimates, Type I error rates, and power performance under different modeling conditions. Furthermore, this research investigates the power and sample size requirements of APIM mediation models within SEM and DSEM frameworks. The findings will provide methodological guidelines for empirical researchers to make informed decisions about selecting the most suitable model framework and optimizing study design for APIM mediation analysis.

Title

Comparative Analysis of EEG Acquisition Systems: Neuroscan with EasyCap vs. OpenBCI with Florida Research Cap

Author(s)

Emma Rico Martín [5] , Damian Enrique Jan Cordón [1] , Manuel De Vega Rodríguez [1] , Ksenia Travina [1] , Iván Padrón González [2] , Melany del Carmen León Méndez [3] , Michele Robelli [4] , Agustina Birba [6] , Yennifer Ravelo González [5] , Hipólito Marrero Hernández [5]

[1] Instituto Universitario de Neurociencia, Universidad de La Laguna, Tenerife, Spain; [2] Instituto Universitario de Neurociencia, Universidad de La Laguna, Tenerife, Spain; Departamento de Psicología Evolutiva y de la Educación. Universidad de La Laguna. Tenerife, Spain; [3] Instituto Universitario de Neurociencia, Universidad de La Laguna, Tenerife, Spain; Departamento de Psicología Psicología Clínica, Psicobiología y Metodología. Universidad de La Laguna, Tenerife, Spain; [4] Università Degli Studi di Trieste; [5] Instituto Universitario de Neurociencia, Universidad de La Laguna, Tenerife, Spain; Departamento de Psicología Cognitiva, Social y Organizacional. Universidad de La Laguna, Tenerife, Spain; [6] Cognitive Neuroscience Center. University of San Andrés, Vito Dumas 284, B1644BID, Victoria, Buenos Aires, Argentina

Abstract

Introduction:
Electroencephalography (EEG) is vital for cognitive neuroscience, traditionally using wet electrode systems like Neuroscan with EasyCap for high signal fidelity and broad scalp coverage. Recently, dry electrode systems such as OpenBCI with the Florida Research Cap have emerged, offering rapid setup and improved participant comfort. This study directly compared these two systems in both resting-state and Go/NoGo task paradigms.
Methods:
Twenty participants (Mage = 20.1, SD = 2.07) completed two EEG sessions on the same day. Participants first underwent a recording with the OpenBCI dry-electrode system (16 channels), followed by a session with the Neuroscan wet-electrode system (62 channels). To ensure comparability, Neuroscan data were downsampled from 500 Hz to 125 Hz, and analysis was restricted to the matching 16 channels. Resting-state recordings included 3 minutes with eyes open and 7 minutes with eyes closed to capture baseline neural oscillations. The Go/NoGo task consisted of 200 trials (80% Go, 20% NoGo) to evaluate reaction times and error rates as measures of inhibitory control. Data were bandpass-filtered (0.5–45 Hz) with a 50 Hz notch filter to remove line noise, and Independent Component Analysis was used to eliminate eye movement and muscle artifacts before segmenting data for ERP analysis (e.g., N2, P3 components) and power spectral density estimation.
Results:
The Neuroscan wet-electrode system demonstrated higher signal-to-noise ratios, lower impedance levels, and more robust ERP components (notably a stronger N2 amplitude) during the Go/NoGo task, with fewer discarded epochs due to artifacts. In contrast, the OpenBCI system offered faster setup times (typically under 10 minutes) and was rated as more comfortable by most participants. Although the dry electrodes were slightly more susceptible to motion and other artifacts, the topographic distributions and temporal characteristics of the EEG signals were comparable across systems. Additionally, time-frequency analysis of the resting-state data yielded comparable results for both systems.
Discussion and Conclusion:
These findings highlight a trade-off between signal quality and ease of use. Neuroscan's wet system is preferable for high-precision applications requiring extensive scalp coverage and minimal noise, such as source analysis and connectivity studies, albeit with longer preparation and the need for specialized gel application. The OpenBCI dry system, while exhibiting minor reductions in SNR and a higher risk of artifacts, provides a quick, user-friendly alternative ideal for portable setups, mobile brain-computer interfaces, and large-scale field studies. Fu-

ture research should focus on enhancing dry electrode materials, refining artifact suppression algorithms, and increasing channel counts to further bridge the performance gap between dry and wet EEG systems in both clinical and research environments.

Title

Investigating Teacher Candidates’ Attitudes Toward Scientific Research: A Methodological Approach

Author(s)
Elisabeth Desiana Mayasari [1]

[1] University of Łódź, Poland

Abstract

This study explores teacher candidates’ attitudes toward scientific research and examines how these attitudes influence their engagement with and perception of research in educational contexts. Understanding future educators’ perspectives is crucial for enhancing their ability to effectively participate in and apply scientific research in their teaching practices.
The study aims to identify teacher candidates’ attitudes toward scientific research and assess how these attitudes impact their preparedness for engaging with educational research in their professional careers. This research employs closed-ended and open-ended survey items to gather insights into teacher candidates’ views on research. A total of 469 teacher candidates from various Indonesian universities participated by completing a 37-item survey measuring five key variables related to their attitudes toward scientific research: (1) reinforcement, (2) practice-based engagement, (3) feelings toward scientific research, (4) self-efficacy, and (5) critical thinking.
The findings indicate that Indonesian teacher candidates generally have positive attitudes toward research. Additionally, the study reveals that teacher candidates perceive research as a journey of personal growth and contribution, encompassing five main themes: (1) intrinsic motivation, driven by a passion for research topics and personal interest; (2) challenges and overcoming difficulties, where participants face obstacles but grow through persistence; (3) time management and organization, fostering discipline and efficiency; (4) broader learning and critical thinking, enhancing awareness and analytical skills; and (5) positive impact and contribution, where research is viewed as valuable for knowledge and society. These findings have implications for curriculum development and teacher training programs, highlighting the need to foster a positive and proactive attitude toward scientific inquiry among teacher candidates.

Title

A quantitative approach to the evaluation of response processes.

Author(s)
Luis Manuel Lozano Fernández , Celia Serrano-Montilla [1] , Andrés González [2]

[1] UNED; [2] University of Granada

Abstract

Data are the basis of every psychometrical inference that psychologists make. The quality of the procedure followed by the researchers is not relevant if data are deficient; every kind of inference based on them will be erroneous.

Of the five sources of validity evidence proposed by the Standards, those related to the response process (along with the consequences) have received the least attention. The Standards provide few sources of validity evidence that allow us to assess whether the cognitive process involved in responding to an item (or set of items) is appropriate for the intended use and the definition of the evaluated construct. Among the different aspects that can affect how a respondent answers a question is the item format. The number of response alternatives or the verbal labels linked to the numerical alternatives, among others, can produce the cognitive process implied in the answer that does not match the expected process. Several answer biases, such as central tendency and acquiescence (among others), can produce erroneous ability estimation of the construct score.

One of the current strategies to approach the study of response styles from the theoretical framework of IRT are the IRTree models. These models are becoming increasingly popular due to their flexibility, ease of interpretation, and “easy” implementation. As the name suggests, these models adopt a tree structure in which multiple pseudo-items are generated representing different cognitive processes. Although there are some multidimensional models nowadays, traditionally, it is assumed that these processes are unidimensional.

This approach is particularly beneficial because it allows the study of various response behaviors involved in each process. In this work, we analyze the response styles of a 5-point Likert-type scale that evaluates perfectionism in adolescents.

Since the number of alternatives is odd, the model called the midpoint primary process (MPP) will be used, in which the variance is decomposed into three nodes: a) central response tendency; b) agreement with the item; c) extreme response. Analyzing the results, it can be seen how this new approach, through IRTree models, allows us to have a ‘purer’ estimate of the ability level of the evaluated individuals after controlling for both the central and extreme response tendency effects. This approximation, as argued by different studies, allows for better psychometric inference, instilling confidence in the validity of the assessments.

IRTree models are shown as a promising approach to evaluating the underlying response processes when answering questionnaires. Obtaining ability levels from which the possible effects of response tendencies, such as central and extreme response biases, have been subtracted will provide a better approximation to other sources of validity evidence. Of course, this approximation to the evaluation of the cognitive processes implied in the answer to questionnaires can be complemented by the classical approaches such as, for example, cognitive interviews or thinking aloud.

Title

Deep Learning -Based Approaches for Continuous-Time Dynamical Systems

Author(s)
Charles Driver [1] , Manaswi Mondol [1]

[1] University of Zurich

Abstract

This project explores deep learning approaches for modelling continuous-time dynamical systems, with a focus on handling missing data and quantifying uncertainty. The models developed in this project use recurrent neural networks to learn temporal dependencies in time-series data, providing a flexible alternative to parametric state-space models such as those implemented in the ctsem R package. By integrating methods for uncertainty estimation and interpolation, the models allow for better comparison with structured approaches in psychological and biological research.
A key advantage of these models is their ability to adaptively infer system dynamics without strong parametric assumptions. Compared to traditional latent-variable models, deep learning approaches can accommodate complex, nonlinear dependencies while still enabling meaningful interpretation through tools such as impulse response functions and conditional predictions. Interpolation techniques further enhance prediction smoothness and enable retrospective analysis of missing time points, making these models applicable to real-world scenarios where measurement irregularities are common.
Beyond predictive accuracy, these models contribute to the interpretability of learned representations by providing uncertainty estimates, confidence intervals, and dynamic response characteristics. Their ability to approximate probability distributions over trajectories allows for better risk assessment and decision-making in scientific forecasting applications. By bridging deep learning with established continuous-time modelling frameworks, this work highlights the potential for hybrid approaches that integrate the strengths of both paradigms.

Title

Mindfulness in Psychotherapist Education: Feasibility and Effects on Psychological Well-being

Author(s)
Andres E. Zerpa [3] , Emiliano Díez [1] , María Teresa Miró [2] , Maria A. Alonso [3]

[1] Instituto Universitario de Integración en la Comunicación (INICO); [2] Facultad de Psicología y Logopedia (ULL); [3] Instituto Universitario de Neurociencia (IUNE)

Abstract

Mindfulness training has proven to be an effective tool for enhancing attentional regulation and reducing adverse psychological states in mental health professionals. However, its integration into university psychotherapist training programs remains limited. This pilot study aimed to assess the feasibility of incorporating mindfulness training into the academic curriculum of future psychotherapists and to analyze its impact on relevant psychological variables. To this end, 51 students were divided into two groups: an experimental group that received mindfulness training over nine weeks and a control group that continued with their regular academic activities. To examine the effects of the training, pre-intervention and post-intervention assessments included the Depression Anxiety Stress Scale-21 (DASS-21) for adverse emotional states and the Five Facet Mindfulness Questionnaire (FFMQ) along with the State Mindfulness Scale (SMS) for mindfulness levels. The results showed that the experimental group, which received mindfulness training, exhibited a significant reduction in stress, anxiety, and depression levels compared to the control group. Additionally, improvements were observed in mindfulness factors related to present-moment awareness and attentional regulation. These findings highlight the feasibility of incorporating mindfulness into psychotherapist training and its potential to enhance both personal well-being and the professional competencies of future psychotherapists.

Title

Meta-Analysis on the Effectiveness of Psychological Interventions: A Study on Replicability and Reproducibility

Author(s)
María Rubio-Aparicio [1] , Laura Badenes-Ribera [2] , Rosa M. Núñez-Núñez [3] , Julio Sánchez-Meca [1]

[1] University of Murcia (Spain); [2] University of Valencia (Spain); [3] University Miguel Hernández (Spain)

Abstract

In recent years, the importance of meta-analyses has been highlighted within the context of evidence-based practice. The reproducibility and replicability of meta-analyses have become critical areas of research due to the inherent complexity in data extraction and analysis. Conducting a meta-analysis involves a series of methodological decisions and steps that can significantly influence the final results.
This study aims to assess the reproducibility and replicability of meta-analyses related to the effectiveness of psychological interventions.
The reproducibility and replicability of the results of 8 randomly selected meta-analyses were empirically evaluated from a pool of 100 meta-analyses published on the effectiveness of psychological interventions.
Statistical data were extracted from individual studies, effect sizes for each individual study were calculated, meta-analytic computations were performed, and the results were compared with those of the original meta-analyses.
The potential implications of errors and inconsistencies in this process are analyzed.

Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

Title

Reporting of standard deviations and pre-post correlations: implications for effect size estimation in meta-analysis.

Author(s)

María Rubio-Aparicio [1] , Julio Sánchez-Meca [1] , Juan J. López-García [1] ,
Manuel J. Albaladejo-Sánchez [1] , José A. López-López [1] , Fulgencio Marín-Martínez [1]

[1] University of Murcia (Spain)

Abstract

One key advancement in the development of meta-analysis methodology is the standardization of the effect size (ES), which allows for the integration of studies under a common metric, but raises the question of which standard deviation should be used. In repeated measures designs, where the ES is defined as the standardized mean change, some authors recommend using the standard deviation of the difference scores (SD). However, some studies do not report the SD, leading to the exclusion of such studies. To prevent the loss of valuable data, the SD can be estimated from the pre-test and post-test variances, along with the pre-post correlation coefficient. This correlation is also necessary to estimate the variance of the ES, which is then used in the model weights. A common challenge is that some studies do not report this correlation, requiring researchers to impute it in some way. A widely used approach is conducting sensitivity analyses by imputing different correlation values and comparing the resulting estimates, although some authors recommend using the average correlation reported in previous studies. We analysed a dataset of primary studies from multiple meta-analyses in clinical psychology to determine the percentage of studies that report the SD and the pre-post correlation, as well as the range of reported correlation values. Finally, we conducted a sensitivity analysis by recalculating the combined ES for each meta-analysis using the mean correlation reported and the most commonly observed values. The implications of these findings are discussed.
Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

Title

Citation Bias in Ego Depletion Research: A Follow-Up to Hardwicke et al. (2021)

Author(s)
Miguel A. Vadillo [1] , Alejandro Sandoval-Lentisco [1]

[1] Universidad Autónoma de Madrid

Abstract

Introduced in the late 1990s, the notion that self-control relies on limited resources and becomes depleted with repeated exertion (termed ego-depletion) quickly gained popularity in social psychology. However, over the past decade, this effect has faced heavy criticism, with failed multi-lab replication attempts and null meta-analytic findings rendering it highly controversial. Nonetheless, proponents of the theory identified possible weaknesses in such studies. Hardwicke et al. (2021) observed that between 2017 and 2019 (immediately following the publication of a failed replication) the number of citations of classic ego-depletion articles barely declined. Moreover, these citations were predominantly favorable to the effect's existence while omitting contrary evidence. In this follow-up, we re-evaluate citation trends of classic ego-depletion articles, extending the analysis over a longer time span, during which two additional important multi-lab replication studies have been published. Building on Hardwicke et al. (2021) framework, we examine (1) the number of citations of the original ego-depletion articles up to 2025, (2) whether these citations are favorable or unfavorable, (3) how often evidence challenging ego-depletion is omitted, and (4) whether citations explicitly defend the effect's credibility. We discuss our findings and their implications for self-correction mechanisms in scientific literature.

Title

Multilevel Models in Single-Case Design: A Systematic Review of Existing Research and Gaps

Author(s)
Cristina Rodríguez Prada [1] , Ricardo Olmos [1] , Antònia Busquets Cantallops [1] , Irene Montáns [1]

[1] Universidad Autónoma de Madrid

Abstract

Multilevel models (MLMs) have been increasingly used in single-case design (SCD) research, providing a statistical framework to account for the hierarchical structure of repeated measurements within individuals. However, the extent to which MLMs have been applied and the specific methodological aspects that have been prioritised remain unclear. This systematic review maps the existing literature on MLMs in SCD, identifying key research trends and areas that require further attention. Our findings indicate that studies have primarily focused on model estimation procedures, fixed and random effect specifications, and handling of autocorrelation. However, critical aspects such as statistical power, Type I error rates, model selection strategies, the use of more complex models, and the treatment of non-normal data have received less attention. By outlining what has been explored and what remains underdeveloped, this review provides a foundation for future research to refine the application of MLMs in SCD and improve methodological rigour in psychological research.

Title

Analysis of the Prevalence of Child Sexual Abuse by Geographical Area.

Author(s)

M Victoria Cerezo , Marta Ferragut , Lucía Palacios Rodríguez , Jesús Ruiz Ramos

Abstract

Introduction: Child sexual abuse (CSA) is a phenomenon that occurs worldwide and has significant negative consequences for those who experience it. Various studies have observed that the prevalence of this phenomenon varies considerably depending on the geographical area in which the study was conducted.
Objective: To analyse the mean prevalence for each geographical area and determine whether it affects the estimation of the mean prevalence of CSA.
Method: The search for articles was carried out in the following databases: Web of Science, SCOPUS, ERIC Ebsco, PsycINFO, and Dialnet. The data were analysed separately according to the sex of the participants (both, female, and male). A meta-regression was performed to assess the impact of the geographical area variable on the estimation of the mean prevalence of mixed CSA, and a mean prevalence was estimated for each of the coded areas.
Results: A total of 31 studies published between 2009 and 2023 were included. The mean prevalence values of child sexual abuse range from 0.06-0.09 in Europe, 0.09-0.11 in Asia, 0.11-0.15 in North America, and 0.15-0.18 in South America. Not enough articles were found to analyse prevalence in Africa.
Conclusions: In general, higher mean prevalence values for CSA were found in the American Continent, particularly in South American countries, whereas the lowest mean prevalence values for sexual abuse were found in Europe. Additionally, in all geographical areas, the mean prevalence of CSA in the female sample was higher than in the mixed and male samples.

Title

Effect of Physical Activity on Total Gestational Weight Gain, Adherence to IOM 2009 Recommendations, and Incidence of Gestational Diabetes Mellitus: A Systematic Review and Meta-Analysis

Author(s)
Marta Fornos Rodríguez [1] , Javier Ibias Martín [2]

[1] Facultad de Psicología, UNED. Área Sanitaria III Avilés (Asturias); [2] Facultad de Psicología UNED

Abstract

Purpose: Physical activity can reduce the incidence of gestational diabetes mellitus (GDM) and gestational weight gain (GWG). This meta-analysis aims to (a) examine the effect of physical activity on GWG and GDM incidence and (b) identify potential moderators.

Methods: A systematic review was conducted using the PubMed, Embase, and Cochrane Clinical Trial databases. For total GWG, the standardized mean difference was used to estimate the effect size, calculating Hedge's g and its variance. The log OR was calculated for the proportion of women exceeding IOM recommendations and GDM incidence. A random-effects method (DerSimonian-Laird) was used to estimate the overall effect size and inter-study variance. Moderator analysis was also conducted.

Results: The effect size of physical activity on total GWG was statistically significant (0.24, $p < 0.001$), with high heterogeneity (72.65%, $p < 0.001$). The effect on the log OR of exceeding IOM recommendations was also significant (-0.48, $p < 0.001$), with significant heterogeneity (73.66%, $p < 0.0001$). The effect size on GDM incidence was moderate and significant (-0.48, $p < 0.05$), with significant but lower levels of heterogeneity (51.96%, $p < 0.05$). Pre-pregnancy overweight status significantly affected total GWG and the log OR of exceeding IOM recommendations. Supervised activity and timing of intervention were significant for GDM incidence ($p < 0.05$, $p < 0.001$). The type of physical activity also significantly affected the log OR of exceeding IOM recommendations ($p < 0.001$) and GDM incidence ($p = 0.027$).

Conclusions: Physical activity reduces total GWG, the proportion of women exceeding IOM recommendations for total GWG, and GDM incidence. Overweight status, supervised activities, and aerobic/mixed exercises enhance these effects. On the other hand, factors such as diet, activity duration, and frequency appear to be less relevant.

Title

Interviews in the Digital Age: Comparing Data from Online and Face-to-Face Methods

Author(s)
Georgina Guilera [1] , Emilio Rojo [2] , Karina Campoverde [1] , Juana Gómez-Benito [1] ,
Maite Barrios [1]

[1] University of Barcelona; [2] International University of Catalonia

Abstract
Introduction
The increasing use of online interviews raises questions about their equivalence to face-to-face interviews in terms of quantity and quality of data collected. This systematic review, conducted following PRISMA guidelines, synthesizes empirical studies comparing both methods to assess their advantages, limitations and impact on research outcomes.
Methods
A search was performed in Web of Science, Medline, Scopus, PsyArticles, PsycInfo, and ERIC using a strategy incorporating terms related to face-to-face and online interviewing, as well as their comparison. Empirical articles with quantitative data published in English-language peer-reviewed journals were selected, while theoretical reviews, reports, and grey literature were excluded. Title and abstract screening was performed using the AI-driven ASReview tool, yielding 114 articles for full-text assessment.
Results
The studies analyzed show differences in the quantity and quality of data collected. Online interviews have advantages such as reduced costs and greater accessibility, but pose challenges in establishing rapport, capturing non-verbal communication, and achieving depth in responses. In some cases, data collected in online interviews may be less rich in emotional or expressive nuances.
Conclusions
Both online and face-to-face interviews have distinct strengths and limitations that may influence data quality. The decision on which modality to use should be based on the objectives of the study and the characteristics of the sample. Future studies could explore hybrid approaches that optimize the benefits of both methodologies.

Title

The JD-R Model in Volunteering: A Longitudinal Approach to Estimation and Missing Data Treatment

Author(s)

Luis Manuel Blanco-Donoso [1] , José Luis Cifri-Gavela [1] , Vanessa Elizabeth Da Silva Larez [1] , Nuria Real-Brioso [1]

[1] Universidad Autónoma de Madrid

Abstract

The Job Demands-Resources (JD-R) model is widely used in organizational psychology but remains largely unexplored in other contexts such as volunteering. Despite theorizing causal relationships that unfold over time, previous studies have primarily relied on cross-sectional data, failing to test these relationships with a longitudinal design, which poses significant methodological challenges. Additionally, the treatment of missing data can substantially influence parameter estimates, leading to divergent conclusions and limiting the model's replicability.

To address these issues, we present an empirical application of the JD-R model that explores the salutogenic effects of job characteristics on well-being using a longitudinal dataset of Spanish volunteers. This study evaluates key methodological decisions related to model estimation and missing data handling, comparing their impact on the interpretation of results. Specifically, we propose two plausible structural equation models (SEM) based on the theoretical framework and compare three missing data treatments: multiple imputation, maximum likelihood estimation, and Bayesian approaches.

Based on our results, we propose clear recommendations for applying the JD-R model in longitudinal research and discuss the limitations of different missing data treatments. These insights aim to enhance methodological rigor, improve reproducibility, and offer applied researchers a more reliable framework for studying organizational psychology processes over time.

Title

Longitudinal Stability of Repeated Covariate Equating in High-Stakes Assessments

Author(s)
Michaela Vařejková [2] , Patrícia Martinková [1] , Eva Potužníková [3]

[1] Institute of Computer Science of the Czech Academy of Sciences and Faculty of Education, Charles University; [2] Institute of Computer Science of the Czech Academy of Sciences and Faculty of Mathematics and Physics, Charles University; [3] Faculty of Education, Charles University

Abstract

In high-stakes educational assessments, ensuring score comparability across multiple test forms is essential for fairness and validity, particularly when test-taker groups differ in ability. Traditional test equating methods rely on anchor items, which allow direct score adjustments between test forms. However, when anchor items are unavailable, covariate equating provides an alternative by using external variables correlated with the test scores (such as grades, prior test scores, or school type) to adjust for ability differences between test-taker groups. One challenge in this approach arises when the covariates themselves are measured using multiple test forms, which may introduce bias into the equating process. A proposed solution is repeated covariate equating, where the covariates are equated before being incorporated into the primary equating process. By analyzing trends across multiple exam years, we aim to determine how repeated covariate equating evolves over time and how changes in the tested population influence test score comparability.
This study explores the longitudinal stability of repeated covariate equating by analyzing test score comparability over multiple years (2016 –2024) in the Czech Republic's national Matura exam, with a focus on the English Language test. Given that the test is administered in both spring and autumn terms, differences in student characteristics across sessions must be accounted for to ensure comparability. Additionally, as test-taking patterns shift over time —due to changes in student cohorts, evolving test-taking behaviors, and shifts in the composition of non-mandatory subject selections —the effectiveness of equating methods may vary across years. Our study examines whether score comparability is maintained consistently when applying repeated covariate equating over an extended period.
Using kernel equating for non-equivalent groups, we adjust for school type, gender, and Czech Language scores as covariates. We equate the autumn test administration to the corresponding spring administration within each year and examine year-to-year equating outcomes to assess whether and how the equating results vary over time. Specifically, we analyze differences in equated scores across years, evaluate the stability of the regression coefficients for covariates, and assess changes in the standard errors of equating. Furthermore, we investigate whether pre-equating the Czech Language test scores (which are themselves measured using multiple test forms) enhances the stability of equated English scores over time.

Title

A Cautionary Note on Simulating Multilevel Data

Author(s)
Diego Iglesias [1] , Miguel A. Sorrel [1] , Ricardo Olmos [1]

[1] Universidad Autónoma de Madrid

Abstract
Simulation studies are widely used in methodological research to assess the performance of statistical methods in scenarios relevant to the behavioral and social sciences. In some cases, the real-world scenario being simulated can be statistically complex. Specifically, in the case of multilevel data, where observations are nested within clusters, simulating level-2 variables and effects may not be straightforward. For example, when simulating random intercept variation, slopes, and level-2 predictors, a common approach is to sample observations from a distribution with a fixed mean and variance and then repeat the sampled values as many times as there are observations per cluster. However, this repetition alters the variance of the resulting variables, which no longer matches the population values previously specified. In this study, we illustrate the impact of this issue in two contexts: parameter estimation and power analysis in multilevel models. Through illustrative simulations, we show that when multilevel data are generated using this approach, having few clusters leads to simulated datasets in which the variance of intercepts, slopes, and level-2 predictors is 10% lower than the intended value. This discrepancy biases the results obtained when fitting statistical models to these simulated data, leading to an underestimation of the variance of intercepts and slopes, as well as the statistical power associated with level-2 effects. We present how the data generating process can be adjusted to correct this issue and discuss practical situations in which failing to account for it could have a greater impact.

Title

Understanding Patterns of Technology Engagement among Students: A Latent Profile Analysis

Author(s)
Nicla Cucinella [1] , Cristiano Inguglia [1] , Francesco Preiti [2] , Francesca Liga [3] , Costanza Baviera [4] , Sonia Ingoglia [1]

[1] Department of Psychology, Educational Science and Human Movement, University of Palermo; [2] Department of Psychology, University of Campania "Luigi Vanvitelli"; [3] Department of Clinical and Experimental Medicine, University of Messina; [4] Department of Research and Innovation in Humanities, University of Bari ”Aldo Moro”

Abstract

Understanding how students engage with technology in their daily lives requires considering both its role as a tool and its potential to distract from school activities. This study examines technology-related attitudes, perceived utility, and patterns of distraction among 803 middle-school students (mean age = 12.2 years, SD = 0.9; 46.6% male) using Latent Profile Analysis (LPA). Four distinct profiles emerged: (1) minimal technology users, (2) underconfident and unaware but easily distracted users, (3) efficient technology users, and (4) overstimulated technology users. These profiles represent a continuum of technology engagement, balanced against varying levels of distraction. Additionally, the study assessed how students'perceived maternal and paternal support in completing schoolwork influences their profile membership. The findings highlight the dual nature of technology as both a valuable resource and a potential source of distraction from school activities, emphasizing the need for tailored strategies to foster mindful and balanced technology use. Future research should further explore how these profiles interact with broader psychosocial and developmental factors.

Title

Applying Growth Mixture Models to the longitudinal study of depressive symptomatology

Author(s)
Zaira Torres Romero [1] , Adrián García Mollá [2] , José M. Tomás [2] , Irene Fernández [3] ,
Laura Galiana [3]

[1] Universidad de Valencia; [2] Department of Methodology for the Behavioral Sciences, University of Valencia, Spain; [3] Universitat de València

Abstract

Background. Depressive symptomatology is highly prevalent among older adults and lack of treatment perpetuates its negative consequences on older adults'functional ability over time. Aside from older adults'vulnerability to depressive disorders, subclinical symptoms have been also shown to be associated with functional disability, as well as to worse prognosis of certain health conditions. The longitudinal study of depressive symptomatology and its associated factors from a person-centered perspective can inform which factors make individuals vulnerable to less favorable trajectories over time. In this work, we aim at identifying different trajectories of depression over a 10-year period and testing the effect of relevant predictors documented in the literature. Method. We employed data from waves 4, 5, 6, 7, 8 and 9 of the Survey of Health, Ageing and Retirement in Europe (SHARE), a biannual, longitudinal, panel study aimed at adults of at least 50 years of age. The sample was formed by 56600 individuals who entered the study in wave 4, most of which were female (56.0%) and married (68.3%). Mean age was 65.93 (SD = 10.01). We employed Growth Mixture Modelling (GMM) to test up to five classes of depression trajectories conditioned on the effects of age, gender, widowhood and socioeconomic status. After the best fitting model was retained, we examined the effect of covariates onto the intercept and slope of each trajectory as well as onto latent class membership. Results. The best-fitting GMM presented three trajectories of depression: AIC= 489365.68, BIC= 489886.62, ABIC= 489692.76, Entropy= .668, ALMR LR test= 9612.01 ($p<.001$), BLRT= -249439.84 ($p<.001$). Although the entropy value was not as high as the two-class GMM, the significant ALMR LR test and BLRT favored the three-class GMM. Moreover, the average latent class probabilities for most likely latent class membership of the corresponding latent class were above the .70. Class 1 (n= 9986, 26.42%) had an intercept of 4.08 ($p<.001$) and a slope of -0.10 ($p=.036$). Class 2 (n= 23492, 62.15%) had an intercept of 1.05 ($p<.001$) and a slope of 0.34 ($p<.001$). Class 3 (n= 4323, 11.43%) had an intercept of 6.71($p<.001$) and a slope of 0.14 ($p=.019$). The observed class-varying effects of age, gender and socioeconomic status indicated that latent class membership moderated the relationship between these covariates and the within-class trajectories. Discussion. In contrast to other research, we included the covariates within the GMM. Not doing so could affect model specification, retained number of classes, and estimation of class proportions and class membership. Specifically, older age, female gender and socioeconomic difficulties were associated to increased likelihood of belonging to a less favorable trajectory. Having widowed was consistently associated to higher initial depressive symptomatology in all classes. Differences among latent classes'intercepts indicated that each trajectory characterized by having a low, medium or high initial level of depressive symptomatology. Differences among latent classes'slopes suggested that depressive symptomatology of individuals classified in Class 1 decreased over time, while it increased for those individuals classified in Class 2 and Class 3.

Title

Developing a Child-Centered Instrument to Measure School Well-Being in Early Childhood: A Mixed-Methods Approach

Author(s)
Silvia Guerrero Moreno , Laura González Moreno [1] , Maria Jesús Pardo Guijarro [1] ,
Marina Oliva Lozano

[1] Universidad de Castilla La Mancha

Abstract

The measurement of school well-being in early childhood presents unique methodological challenges, given the scarcity of self-reported instruments tailored for young children. The present study addresses this gap by developing a new instrument inspired by the Maryland Safe and Supportive Schools Climate Survey (Bradshaw et al., 2014), specifically designed for children aged three to twelve. The questionnaire consists of 22 items crafted in a child-friendly format, integrating visual cues and simplified language suitable for early developmental stages.

A mixed-methods approach was adopted to enhance the robustness of the instrument. An Exploratory Factor Analysis (EFA) revealed a five-factor structure: enjoyment of school, relational climate, sense of belonging, perception of aggression, and outdoor school space. The "sense of belonging" emerged as a central factor, showing strong correlations with other dimensions, emphasizing the critical role of teacher-student relationships.

Complementing the quantitative analysis, qualitative data were collected through open-ended questions and thematic mapping of children's narratives. An inductive approach guided the qualitative analysis, allowing themes to emerge organically from the data. This facilitated a deeper understanding of children's perceptions of school well-being and provided a richer context to the quantitative findings.

Furthermore, the process of crafting items suitable for young children is discussed, focusing on the cognitive and linguistic considerations essential for ensuring validity and reliability in early childhood assessments. The study underscores the importance of methodological innovation in educational research and highlights the value of integrating qualitative insights to enhance the psychometric properties of child-focused instruments

Title

Evaluation and intervention in a case of mild cognitive impairment, amnestic subtype

Author(s)

Alicia Méndez González , Alberto Domínguez [1] , María A. Alonso [1]

[1] ULL

Abstract

Mild Cognitive Impairment (MCI) is considered one of the main predictive factors for the development of dementia and Alzheimer disease. Given its progressive nature, early and preventive intervention is crucial to slowing cognitive decline and delaying the onset of neurodegenerative disorders. Memory impairment is among the first symptoms of MCI,making it essential to conduct a comprehensive assessment of different memory subtypes
to better understand the specific deficits of each patient. This study aimed to evaluate distinct memory subtypes autobiographical memory, prospective memory, working memory, and both verbal and visual long-term memory to develop a targeted and individualized intervention plan. A single-case observational design was implemented, with pre and post-intervention assessments conducted to measure cognitive changes over time. The intervention spanned six weeks and was designed to stimulate and rehabilitate memory functions through
structured cognitive exercises and personalized strategies. The results demonstrated noticeable improvements in various memory subtypes following the intervention, suggesting that individualized cognitive training may enhance memory performance in individuals with MCI. These findings underscore the importance of tailored therapeutic
approaches in mitigating memory decline and highlight the potential benefits of early cognitive intervention in slowing the progression toward dementia. Future research should further explore the long-term effects of personalized cognitive rehabilitation programs in diverse patient populations.

1.12 KEY NOTE: Rethinking Measurement for the 21st Century

Title

KEY NOTE: Rethinking Measurement for the 21st Century

Author(s)
Javier Suárez-Álvarez [1]

[1] University of Massachusetts Amherst

Abstract

As public trust in standardized testing declines, AI-driven methods such as machine learning and natural language processing are increasingly being applied to optimize traditional measurement approaches. While these innovations offer important gains in efficiency, cost, and scalability, there is a risk that, without also addressing broader concerns of trust, equity, and relevance, psychometrics may become increasingly disconnected from evolving scientific standards, societal needs, and ethical principles.

Psychometrics has been instrumental in establishing psychology and education as scientific disciplines, sharpening clinical diagnosis, advancing prevention, promoting educational equity, and exposing systemic inequities. Yet, as we navigate the complexities of an increasingly diverse and technology driven 21st century, it is necessary to ask whether our current assessments, still largely grounded in 20th-century measurement theories and assumptions, are adequately equipped to meet the evolving needs of today's test users.

This presentation offers a critical yet constructive reflection on how fragmented assessment systems, outdated assumptions, and rigid adherence to technical standards detached from the lived realities of those being assessed can unintentionally limit our collective impact and overlook opportunities to better serve society. By revisiting classic debates, I invite us to question long-held measurement mantras and consider how the field can evolve to better serve a rapidly changing world. Through concrete examples, I advocate for assessment systems that are responsive to real-world contexts, address the diverse needs of test users, and thoughtfully balance implementation trade-offs by considering opportunity costs.

Ultimately, aligning psychometrics with the demands of the 21st century will position us to leverage AI-driven methods not only to optimize traditional measurement, but also to become more scientifically interdisciplinary, socially responsive, and ethically grounded in advancing societal progress.

1.13 Session 16 : ”Mixed methods and Behavior assessment”

Title

A Mixed-Method Network Analysis Approach for Enhancing the Development of Assessment Instruments and Obtaining Validity Evidence: A Questionnaire to Measure Police Officers Attitudes towards Intervention in Gender-Based Violence

Author(s)

Celia Serrano Montilla [1] , Jose-Luis Padilla Garcia [2] , Luis Manuel Lozano Fernández

[1] UNED; [2] University of Granada Faculty of Psychology: Universidad de Granada Facultad de Psicologia

Abstract

In this study, we propose a mixed-method network analysis approach to enhance the development of measurement instruments and gather robust validity evidence. By integrating semantic and psychometric networks, this method uncovers the relational dynamics between behaviors (semantic definitions) and items (final instrument version) for obtaining content validity evidence and internal structure validity, while it also identifies key elements critical for advancing the development process of the instrument. Specifically, the mixed method approach is illustrated by two studies carried out in the development of an instrument to assess police attitudes toward intervention in gender-based violence. The first study compiled six focus groups involving 36 specialized and non-specialized gender-based violence police officers and applied a mixed-method semantic network analysis that combined qualitative data (i.e., themes extracted from police discussions on their perspectives about intervention in gender-based violence) and quantitative data (i.e., co-occurrence of these themes within police discussions). Multidimensional scaling (MDS) was used to visualize the network along with node and network centrality indices. The goal was to obtain domain definition validity evidence and identify key behaviors shaping police attitudes toward intervention in gender-based violence. The second study drew on 272 police responses to the 32-item first version of the APIVG-S, administered online via the Unipark platform. It aimed to detect key items and obtain validity evidence based on the internal structure. For psychometric network analysis, we estimated partial correlation networks using Spearman's rank-correlation and the glasso regularization method, with a tuning parameter set to 0.5. The accuracy and stability of the edge estimates were assessed through nonparametric bootstrapping. On the one hand, the semantic network analysis revealed that attitudinal behaviors were grouped along two theoretical dimensions (reactive and proactive), providing validity evidence for domain definition. Similarly, the behavior “conventional policing in gender-based violence”emerged as the most central node in the semantic definition, highlighting which behaviors warrant a greater number of items in the table of specifications. On the other hand, the psychometric network analysis showed a medium level of connectivity (sparsity index of 0.544) and provided validity evidence based on the internal structure, as the most interconnected items corresponded to the same theoretical dimensions. The network was found to be relatively accurate and stable. The benefits of the mixed-method network analysis approach are discussed in terms of the joint interpretation of qualitative and quantitative data, incorporating the experiences of the target population (and not only experts) into the instrument development process, and identifying different indicators of the importance of behaviors in the semantic definition and the instrument's items.

Title

The value of mixed methods research over the present decade in Psychology: A critique study

Author(s)
Olatz Lopez-Fernandez [1]

[1] Universidad Nacional de Educación a Distancia

Abstract

Introduction: Mixed methods research (MMR) refers to integrating quantitative (QUAN) and qualitative (QUAL) approaches within a research study. This century has recognised MMR as a third methodological approach. During the last decade, international bodies such as the Journal of Mixed Methods Research (JMMR) and the Mixed Methods International Research Association (MMIRA) have established the milestones of the MMR. A work-in-progress review observed that psychology is one field in which there have been fewer advancements.

Objective: This communication aims to extract from a scoping review of the JMMR over the first half of the present decade: (i) the MMR features applied to the Psychology field, (ii) what MMR-specific novelties have recently appeared in other areas that could be applied to our field, (iii) the MMR global implementation and report facets holistically applied interdisciplinary, and (iv) a reflection about what measure to be taken to advance in MMR in Psychology.

Results: Despite publishing four issues annually, the advancements observed in the JMMR reflect a significant positive academic evolution based on the impact factor, as indicated by the Journal Citation Reports and the Scimago Journal and Country Rank. Although it is a multidisciplinary journal, the JMMR is situated in the first decile of the interdisciplinary social sciences category. Regarding the Psychological field, theoretical and conceptual foundations, these advancements relate to a set of paradigms (e.g., the transformative approaches). Research designs are more complex, especially the qualitative approach (e.g., grounded theory with a relational design). However, most advancements come from other fields: health, social, and educational sciences. There are many paradigmatical proposals (e.g., to avoid ambiguities, communicate better QUAL and QUAN integration, construct the study with an artisanal attitude, embrace complexity, and promote more social justice). Regarding the designs, while convergence is more prevalent, others, such as multilevel and case studies, are also present. The major expansion has been on data analysis techniques (e.g., mixed cross-sectional analysis, network analysis, cost-effectiveness analysis, hybrid inductive/deductive thematic analysis), with a rise in the application of mixed-content digital techniques using visual elements (e.g., digital photographs, vignettes, Venn diagrams, mapping strategies). The MMR global advancements include integrating QUAN and QUAL methods and pushing for a comprehensive review of terminology, quality standards, and reporting guidelines.

Conclusions: A recent examination reveals an encouraging trend in MMR, particularly in interdisciplinary areas where Psychology is underrepresented. This field could benefit from the application of technological and data analysis techniques, especially with enhanced graphic support. Additionally, there appears to be increasing complexity in the theoretical assumptions surrounding MMR, along with significant developments in analytical methods. However, research designs remain the least advanced technical-methodological aspect, representing a weak area where Psychology excels and could contribute to progress. Lastly, the quality of research outputs and the dissemination of MMR within our field need to be addressed in this second decade of the 2020s.

Title

A systematic review and an internal consistency analysis of behavioral habit measures

Author(s)
Pablo Martínez López [1] , David Luque [1] , Miguel A. Vadillo [2] , Francisco Garre Frutos [3]

[1] Universidad de Málaga; [2] Universidad Autónoma de Madrid; [3] Universidad de Granada

Abstract

Research with animal models has shown that repeating an action with enough frequency transforms it from goal-directed to habitual. In contrast to goal-directed, habitual behavior is insensitive to changes in outcome value, inflexible, and guided by the specific context where it was formed. However, the key prediction from animal studies that training leads to habit behavior has not been consistently reproduced in humans. This fact poses a crucial translational problem, requiring a valid procedure for inducing habits in humans. In this work, we reviewed the methodological and theoretical foundations of experimental paradigms that assess habit formation in the laboratory. Across eight studies and seven experimental paradigms, we found mixed results regarding the sensitivity to detect habit formation that each measure had. Only two studies included a paradigm with a measure sensitive to overtraining, while others relied on ad-hoc individual differences explanations to account for its lack of sensitivity at the group level. In this context, future research should evaluate the construct validity of each measure and, additionally, characterize potential individual differences in habit performance. However, it is first necessary to test the reliability of the measures. Here, we reanalyzed public data from each study and calculated the internal consistency of each paradigm's measures using split-half reliability via a permuted random split procedure. Internal consistency of habit measures ranged from 0.17 to 0.84. The results of this work suggest which habit measures are most promising psychometrically and contribute to the establishment of a common protocol for measuring habitual behaviors in humans.

Title

The Environmental Decision Task: A new behavioral paradigm for studying the money-environment trade-off

Author(s)
Frederik De Spiegeleer [2] , Kobe Millet [1] , Bert Weijters [2]

[1] Vrije Universiteit Amsterdam; [2] Ghent University

Abstract

Whether or not people make pro-environmental decisions often depends on the extent to which personal consequences outweigh environmental consequences. Recently, some decision tasks have been introduced to study pro-environmental decision-making when there is a trade-off between environmental and individual consequences (e.g., money-environment trade-off). These tasks make it possible to study pro-environmental decision-making in controlled lab settings. The main goal of the current set of studies was to develop a new, intuitive, easy-to-perform, and easy-to-conduct task to study the money-environment trade-off.
In the Environmental Decision Task (EDT), participants decide whether to receive money or invest it in the fight against climate change. Three studies (N = 1,294) were conducted to validate the EDT by analyzing its relationship with self-reported pro-environmental and pro-self propensities, revealing weak to moderate correlations that support the task's ability to capture relevant trade-offs. Additionally, social desirability did not substantially relate to the decision-making in the task, and open-ended responses supported that participants'choices aligned with the task's intended purpose. While we have examined EDT with many trials (up to 48) and consequential choices, we demonstrate that just nine hypothetical trials are sufficient to reliably examine decision-making under the money-environment trade-off. This makes the paradigm particularly suitable for short (online) studies.
In the presentation, we will discuss how such tasks can be used as dependent variables when presenting participants with different conditions, allowing us to systematically investigate how contextual factors influence decision-making processes. We will also use this example of the study to explain how these behavioral tasks can be used to extract different latent factors considering people's decision tendencies when confronted with different trade-offs and how they relate to many different categories of consumption behavior. Furthermore, we will delve deeper into the importance of examining qualitative data to better understand the strategies people actually use when making decisions in such tasks.

Title

The advancements in mixed methods research over the present decade as outlined in the Journal of Mixed Methods Research: A scoping review

Author(s)

Olatz Lopez-Fernandez [1]

[1] Universidad Nacional de Educación a Distancia

Abstract

Introduction: Mixed methods research (MMR) refers to integrating quantitative and qualitative approaches within a research study. This century has recognised MMR as a third methodological approach. During the last decade, the Journal of Mixed Methods Research (JMMR) has established the milestones of the MMR together with the Mixed Methods International Research Association (MMIRA).

Objective: This communication aims to conduct a scoping review of the articles that have made significant contributions to the advancement of the MMR over the first half of the present decade (2020-2025) to examine (i) the impact factor evolution, (ii) the emerging theoretical and conceptual trends, (iii) the techno-methodological elements, (iv) the present MMR implementation and report facets.

Results: Despite publishing four issues annually, the advancements observed reflect a significant positive evolution in the impact factor, as indicated by the Journal Citation Reports and the Scimago Journal and Country Rank. The JMMR is situated in the first decile of the interdisciplinary category of social sciences. Regarding the theoretical and conceptual foundations, these advancements relate to a set of paradigms (e.g., complexity theory and transformative approaches). Concerning techno-methodological characteristics, sampling strategies are now clearly stated; research designs are more complex, especially the qualitative approach (e.g., action research, grounded theory); and data analysis techniques have emerged at technical and technological facets (e.g., mixed cross-sectional analysis, social network analysis, cost-effectiveness analysis, participatory visual methods, hybrid inductive/deductive thematic analysis). At the beginning of the decade, integrating computational methods into textual analysis became prevalent (i.e., language processing utilising extensive textual data corpora). There was a rise in the application of mixed-content digital techniques and increased use of visual elements (e.g., digital photographs, vignettes, Venn diagrams, mapping strategies). The debate about the dichotomy of quantification versus qualification of the methods, designs, techniques, and data analyses, plus the integration of both, has pushed for a comprehensive review of terminology, quality standards and guidelines for reporting.

Conclusions: A recent examination reveals an encouraging trend in MMR in the interdisciplinary arena. The significant advancements are at its theoretical and methodological facets, enhancing visual support. The quality of research outputs and strategic editorial management has probably provided outstanding impact factors during this decade, and the efforts toward improving the quality of MMR and its dissemination are commendable achievements.

1.14 Session 7 : ”Clustering and classification methods in psychology”

Title

Multiphase Optimization Strategy (MOST) for Equitable Cluster Randomized Interventions: Design Considerations and Statistical Modeling

Author(s)

Tania B. Huedo-Medina [2] , Nekane Balluerka [1] , Elena Pérez Setién [3] , Natalia Alonso-Alberca [1]

[1] University of the Basque Country; [2] Fundación Ikerbasque; [3] Universidad del País Vasco UPV/EHU

Abstract

The Multiphase Optimization Strategy (MOST) is a principled framework that integrates behavioral science, engineering, implementation science, economics, and decision science to optimize interventions. MOST enables researchers to strategically balance effectiveness, affordability, scalability, and efficiency. In this presentation, we provide an overview of MOST, highlighting key experimental designs for intervention optimization and guiding principles for selecting intervention components based on empirical evidence.

A particular focus is placed on factorial experimental designs for optimizing cluster-randomized interventions, where participants are nested within higher-level units (e.g., students within schools, employees within organizations). We discuss key methodological considerations, including the use of mixed-effects modeling to analyze clustered data.

Additionally, we introduce an approach to optimizing interventions for health equity within MOST. We define intervention equitability as the extent to which health benefits are distributed evenly across populations, rather than disproportionately benefiting already advantaged groups. If equitability is a priority, it should be explicitly incorporated as a criterion alongside effectiveness, affordability, scalability, and efficiency. Using a hypothetical case study, we illustrate how MOST can be applied in cluster-randomized trials to achieve an optimal balance that integrates equitability as a core objective.

This work underscores the importance of a systematic, evidence-based approach to optimizing interventions that not only maximize impact but also promote fairness in health outcomes.

Title

Using the Stochastic Block Model for Clustering in Psychological Networks

Author(s)
Nikola Sekulovski [1]

[1] University of Amsterdam

Abstract

Psychological network analysis has emerged as a powerful tool for modeling psychological constructs as complex systems of interacting variables. However, existing statistical methods do not explicitly incorporate theoretical assumptions about clustering, despite the central role that clusters play in many psychological network theories. In this work, we propose to fill this gap by using the Stochastic Block Model (SBM) as a prior distribution on the network structure. The SBM assumes that variables belong to latent clusters, with the probability of an edge depending only on the cluster membership of the nodes. By integrating this prior into the well-established Bayesian graphical modeling framework, we enable researchers to formally incorporate theoretical expectations about clustering into the statistical model and to infer the clustering configuration using the data at hand. We demonstrate the advantages of this approach through a simulation study. We further illustrate its practical utility by reanalyzing 30 openly available empirical datasets, where we find evidence of clustering in several cases. This study contributes to bridging the gap between psychological network theory and statistical modeling and provides a new statistical method for estimating the number of clusters and cluster membership of the variables in the network.

Title

Some Simple Methods for Creating a Pooled Cluster Solution using K-Means Clustering in Multiply Imputed Data

Author(s)
Joost van Ginkel [1] , Anikó Lovik [1]

[1] Leiden University

Abstract

K-means clustering is a widely used technique to cluster cases in a dataset into a number of groups. When data are incomplete, missing data need to be treated prior to carrying out K-means clustering. Multiple imputation is a widely recommended procedure for dealing with missing data, which creates multiple plausible complete versions of the incomplete dataset. When applied to each of these imputed datasets, K-means clustering requires a method to combine the cluster solutions of the several imputed datasets into one overall cluster solution. Several combination methods have been proposed, such as majority vote, multiply imputed cluster analysis, and partitions pooling. These methods either come with practical problems, or try to resolve these problems using rather involved procedures. In the current presentation we propose two simple generalizations of the K-means clustering algorithm for complete data to multiply imputed datasets, which bypass all the problems that the other methods try to resolve. In a simulation study it is shown that the two newly proposed methods better recover the underlying cluster structure than the existing methods.

Title

Clusterwise-IVA: a new method to uncover patient heterogeneity by clustering subjects based on temporal changes in underlying processes

Author(s)
Tom Wilderjans [2] , Jeffrey Durieux [1]

[1] Leiden University; Erasmus University Rotterdam; [2] Leiden University

Abstract

In different fields of science (e.g., neuroscience and psychology) researchers'main focus consists of disclosing the processes and temporal changes therein underlying longitudinal (big) multivariate data. In neuroscience, for example, patients are regularly scanned with fMRI (e.g., yearly sessions) with the aim of disclosing temporal changes in functional connectivity (FC) patterns (i.e., correlated brain regions collaborating in psychological functioning) related to a particular disease (e.g., depression or dementia). In this regard, longitudinal changes in FC patterns that are typical for dementia patients were identified in previous studies (Dautricourt et al., 2021). To extract temporal changes in FC patterns from longitudinal multi-subject fMRI data, Independent Vector Analysis (IVA) was proposed, which performs ICA on the data of each session and restricts the associated components to be dependent across sessions. As such, the extracted (dependent) components capture longitudinal changes in FC patterns.

When studying brain diseases (e.g., dementia), however, often patient heterogeneity in the underlying processes and in the temporal changes therein is present. This heterogeneity often can be linked to the existence of subtypes of the disease (i.e., heterogeneous patterns of disease development across patient groups pointing at different types of dementia). To uncover this patient heterogeneity, we propose to cluster the patients based on similarities and differences in the temporal change profiles underlying subjects'FC patterns. By studying the differences in temporal change profiles across clusters, important insights regarding disease subtypes can be gained.

To our knowledge, however, no statistical procedure exists that is able to simultaneously extract the patient clusters and the temporal changes in FC patterns underlying each cluster. To this end, we propose Clusterwise IVA, which clusters subjects and at the same time estimates the longitudinally changes in FC patterns characterizing each subject cluster. An Alternating Least Squares (ALS) algorithm will be used to optimally estimate the Clusterwise IVA parameters. In our presentation, Clusterwise IVA will be explained and the performance of the method will be evaluated by means of an extensive simulation study and/or an illustrative application to longitudinal multi-subject fMRI data from Alzheimer patients.

References

Dautricourt, S., de Flores, R., Landeau, B., Poisnel, G., Vanhoutte, M., Delcroix, N., Eustache, F., Vivien, D., de la Sayette, V., & Chételat, G. (2021). Longitudinal changes in hippocampal network connectivity in Alzheimer's disease. Annals of Neurology, 90(3), 391-406. https://doi.org/10.1002/ana.26168

Title

Method for Sample Size Determination for Cluster Randomised Trials Using the Bayes Factor

Author(s)
Camila Natalia Barragan Ibañez [1] , Herbert Hoijtink [1] , Mirjam Moerbeek [1]

[1] Utrecht University

Abstract
Determining the sample size is a key step for a robust research design, yet most current methods for sample size determination are based on the null hypothesis significance testing (NHST), an approach with multiple pitfalls. Methods using the Bayes factor as an alternative for hypothesis testing are still scarce in multilevel models. We have designed a method for sample size determination for two-treatment parallel cluster randomised trials using the Bayes factor. In this method, the sample size—either cluster size or the number of clusters—is determined to ensure a probability of finding a Bayes factor larger than a user-specified threshold. Through the simulation of realistic scenarios, we evaluated the effect of several factors influencing the required sample size. The results of this simulation study will be presented, along with general recommendations for a priori sample size determination in cluster randomised trials, and an introduction of the R functions available to researchers for this purpose.

Title

Bayesian Sample Size Determination for Longitudinal Trials with Attrition

Author(s)

Ulrich Lösener [1] , Mirjam Moerbeek [1] , Herbert Hoijtink [1]

[1] Utrecht University

Abstract

When investigating the effects of an intervention over time, researchers often rely on longitudinal data. To analyze such data, multilevel/hierarchical models are fundamental as they account for the nested structure of the data. Inferences about treatment effects are drawn by testing hypothesis about the model parameters. Traditionally, this is done through null hypothesis significance testing (NHST) via p-values. However, in recent decades, methodologists and statisticians have increasingly advocated for Bayesian hypothesis evaluation (BHE) as an alternative. BHE offers several advantages, such as providing direct probabilistic statements about informative hypotheses and avoiding some of the drawbacks associated with NHST. Despite this, many applied researchers struggle to adopt BHE in more complex models such as multilevel models as they lack accessible tools and guidance.
For example, ethical committees and funding agencies require a motivation of the sample size by means of sample size determination (SSD) but in absence of closed-form equations for BHE, researchers must instead rely on Monte Carlo simulations. Available software that perform these simulations for BHE is often limited to simpler models such as t-tests and ANOVA, leaving researchers without a practical solution to perform SSD for their longitudinal trials.
To address these challenges, we present an open-source R software designed to carry out simulation-based Bayesian SSD in multilevel models for longitudinal trials in a user-friendly way. The software also accommodates various patterns of attrition (dropout), which are common in longitudinal studies and can significantly impact the power of these experiments. In addition, we provide a practical explanation of how to use and interpret the Approximate Adjusted Fractional Bayes Factor, an inferential tool for BHE that stands out for its simple and computationally inexpensive calculation.
By lowering the technical barriers to using BHE in multilevel models, we hope to contribute to bridging the gap between methodological advancements and practical application, enabling researchers to leverage the full potential of Bayesian methods in their work.

1.15 Session 9 : ”Psychometric Applications in Health and Wellbeing”

Title

Intersectional sleep disparities: association between multiple social intersections, perceived neighborhood deprivation and sleep disturbance in Europe

Author(s)

Enrique Alonso Perez

Abstract

Background: The prevalence of sleep disturbance, related with social status and privilege, is unevenly distributed within societies. Individual social determinants that are embedded within broader neighborhood contexts intersect and jointly shape sleep disparities. This study incorporates a quantitative intersectional framework to better understand the structural inequalities in sleep disturbance, with a focus on the social-ecological model of sleep and how individual and social context factors interact.
Methods: Our sample consisted of 17,035 individuals aged 50 and older from waves 4 and 5 of the Survey of Health, Aging and Retirement in Europe (SHARE). We created 72 unique intersectional strata by interacting individual axes of social inequality (sex/gender, family caregiving, education, occupation) with perceived neighborhood deprivation. To investigate the variations in sleep disturbance across intersectional strata, we employed intersectional Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA).
Results: Intersectional strata explained a fair magnitude of the variance in sleep disturbance (6.3%). The most disadvantaged groups, particularly women with low-education, low-skill occupations who were caregivers in perceived highly-deprived neighborhoods, exhibited the largest number of sleep disturbance. Sex/gender and perceived neighborhood deprivation were the main predictors of such differences. While some multiplicative effects were found, additive effects predominated.
Conclusions: Given the importance of sleep for health, coupled with increasing social inequalities, our findings suggest that intersectionality is a valuable framework for mapping and addressing sleep disparities. Tailored interventions should go beyond individual factors to include community-level measures, targeting socially vulnerable groups, especially women experiencing neighborhood deprivation.

Title

Decision Tree-Based Adaptive Testing in Psychodiagnostic Screening of multiple mental health conditions

Author(s)
Pasquale Anselmi [1] , Daiana Colledani [2]

[1] University of Padova; [2] Department of Psychology, Faculty of Medicine and Psychology, Sapienza University of Rome

Abstract
Traditional psychometric screening is often time-consuming and tedious, potentially compromising respondent engagement and data quality, especially among vulnerable populations. This study examines the suitability of machine learning-based computerized adaptive testing (ML-based CAT) for accurate and efficient screening of multiple mental health conditions. The study employed a cross-validation approach, based on real-data simulations, to train and test the performance of ML-based CAT in categorizing respondents as at-risk or not at-risk for pairs of disorders simultaneously assessed (i.e., depression and specific phobia; agoraphobia and social anxiety). The results indicate that ML-based CAT exhibited high diagnostic accuracy, while significantly reducing the number of items administered by more than 50%. The findings suggest that the approach based on the simultaneous classification of two disorders is more efficient than the approach based on the classification of the single disorders separately, with negligible loss of diagnostic accuracy. These results indicate that ML-based CAT has the potential to improve the efficiency and accuracy of mental health assessments by offering a versatile and effective alternative to traditional methods.

Title

What do patients and surgeons know and believe about shared decision-making when choosing treatments for colorectal cancer? A qualitative work in progress

Author(s)
Bárbara Horrillo [2] , Ángels Roca [3] , Héctor Guadalajara [4] , Jonathan McFarland [5] , Víctor Horno [6] , Maite Carreras [3] , María Luisa Sánchez de Molina Rampérez [4] , Olatz Lopez-Fernandez [1]

[1] Universidad Nacional de Educación a Distancia; [2] CES Cardenal Cisneros, Universidad Villanueva; [3] ASIA Asociación incontinencia; [4] Hospital Universitario Fundación Jiménez Díaz, Universidad Autónoma de Madrid; [5] Universidad Autónoma de Madrid, The Doctor as a Humanist; [6] CES Cardenal Cisneros

Abstract

Introduction: Shared decision-making (SDM) represents a collaborative approach in healthcare that actively involves patients in clinical decision-making processes. In recent years, SDM has gained traction in Spain, although its implementation poses certain complexities, and there are still few studies grounded in scientific evidence. Physicians attempting to incorporate SDM often encounter challenges due to limited time and training. However, they endeavour to foster more transparent and personalised communication, even to educate patients about their options. The present research pretends to explore the implications of applying SDM in the context of colorectal cancer (CRC), a disease recognised for its high prevalence at national and international levels. The aim is to understand what CRC surgeons and patients know and think about SDM in treatment options. Specifically, we aim to determine their experiences or expectations with SDM, the implications they associate with SDM, and the aspects they value or have concerns about.

Methods: a qualitative ethnographic study is currently in progress, utilising three research techniques: in-depth interviews with CRC surgeons, semi-structured interviews with patients recently diagnosed with CRC, and focus groups involving patients undergoing treatment for CRC, former patients, and clinicians. All interviews and focus group discussions have been transcribed, and a pilot exploratory and inductive content analysis has been conducted to develop a thematic analysis.

Results: The medical discourse surrounding SDM in CRC reveals opinions regarding decision-making that often present as somewhat ambiguous, exhibiting a negative bias. Healthcare professionals seem to possess limited knowledge on the subject, which leads to restrained communication; nevertheless, they appreciate offering patients a range of options. Key terms frequently referenced include patient, decision, surgeon, and time. The predominant emerging themes are uncertainty, empathy, communication, and medical expertise, encompassing the doctor-patient relationship, decision-making processes, and surgical considerations. Patient interviews highlight a consensus on the value of training related to treatment options. However, concerns have been expressed about the implications of understanding their health status and using specific technological tools. Significant terms include treatment, application, and decision, and the critical themes identified centre on uncertainty, technology, information preferences, and family involvement. In stakeholder discussions, the terms most frequently cited are patient, decision, information, and doctor. Positive views are associated with the decision-making process regarding cancer treatment, and concerns arise from feelings of uncertainty, a perceived lack of options, and articulation difficulties. The overarching themes from these discussions emphasise empathy, the doctor-patient relationship, decision-making, uncertainty, communication, emotional support, and the role of technology.

Conclusions:
The preliminary findings reveal an understanding of SDM in CRC in Spain, although the con-

cept could be ambiguous, its careful application could lead to empowered choices and benefits. Healthcare professionals recognise the importance of enhancing communication and are eager to expand their knowledge of SDM. Patients value being informed about treatment options and seek family support in their decision-making journey. Both groups generally maintain a positive perspective on SDM implementation while acknowledging the potential risks.

Title

Sex-Based Differential Item Functioning in the Broad Autism Phenotype-International Test

Author(s)

Javier Mayoral López [1] , Marta Godoy Giménez [1] , Ángel García Pérez [1] ,
María Casado Sánchez [1] , José Fornieles Alonso [1] , Pablo Sayans Jiménez ,
Andrés Soler Martínez [1]

[1] University of Almería

Abstract

Background: The Broad Autism Phenotype (BAP) refers to subclinical levels of autism-related traits, which, according to the dimensional model of autism, are continuously distributed beyond individuals with autism spectrum disorder (ASD) and their relatives, extending into the general population. Recently, the Broad Autism Phenotype—International Test (BAP-IT) was developed to assess the BAP in two countries, Spain and the United Kingdom. Initial validation studies showed strong psychometric properties (e.g., configural invariance and partial metric invariance, high reliability, and adequate validity evidence). However, the BAP-IT has not yet been examined in relation to one of the most salient issues regarding the assessment of BAP and ASD: the lack of evidence of invariance between sexes. This, together with the predominance of BAP and ASD studies on samples mainly composed of males and the differential expression of some BAP/ASD traits between sexes, could have contributed to the underdiagnosis of women. Objective: This study aims to evaluate the dimensional structure of the BAP-IT by applying the Rasch Rating Scale Model. Specifically, we seek to assess differential item functioning across sexes in the United Kingdom and Spain. Method: The study included two community samples from Spain (SP1 = 970; SP2 = 460) and one from the UK (EN = 530) that completed the BAP-IT in their respective languages. Results: The psychometric properties of both BAP-IT versions were good, and both showed no DIF between sexes (differences greater than 0.50 logits). Mean sex comparisons revealed no differences in the UK sample and small size differences in the ES sample. Conclusion: These results are not only important for supporting the international use of the BAP-IT but also a first step in studying BAP sex differences and their interplay with culture and language.

Title

Proposing an enhanced continous norming approach for non-normal data and nonlinear trends in adapting the WISC-V to the Basque language

Author(s)

Marcos Jiménez [2] , Andone Sistiaga [1] , Arantxa Gorostiaga [1] , Jone Aliri [1] , Nerea Lertxundi [1] , Nekane Balluerka [1]

[1] University of the Basque Country (UPV/EHU); [2] Vrije Universiteit Amsterdam

Abstract

The Wechsler Intelligence Scale for Children (WISC-V) is widely recognized as one of the best instruments for assessing the intellectual abilities of children aged 6 to 16 years, playing a central role in psychopedagogical assessments and school evaluations.

However, until now, no adaptation existed for Euskera—the primary language used in schools throughout the Basque Country. Consequently, children in this region are disadvantaged when taking the Spanish version of the test, as their Spanish proficiency is generally less developed than their Euskera counterpart. Using a sample of 660 children and adolescents who had Basque as their mother tongue, we adapted the WISC-V to the Basque culture by conducting reliability and validity analyses that largely replicated the factor reliabilities and structure of the Spanish version. Additionally, we developed norm tables for 33 distinct age groups. Given that each age group included only about 20 participants, we used the full sample to derive norm scores through a process known as continuous norming. The simplest form of continuous norming involves regressing the first two moments (mean and variance) of the raw scores on age using polynomial models. This process yields theoretical distributions from which quantiles can be extracted, allowing norms to be established that follow a normal distribution with a specified mean and standard deviation. However, this approach has two significant shortcomings. First, polynomial regression is susceptible to either overfitting or underfitting. Second, assuming a normal distribution is problematic when ceiling and floor effects are present or when the distribution of scores deviates from a Gaussian shape. To address these issues, we propose an enhanced continuous norming approach that significantly improves the fit of the first three moments—including skewness. This method employs a more robust regression technique that accounts for nonlinearities through the use of splines and also accommodates deviations of the normal distribution with a truncated skew-normal distribution. Based on this proposal, a comparison of results with the traditional and improved continuous norming approaches in these WISC-V data is discussed.

Title

Methodological quality of meta-analyses and systematic reviews on the psychological interventions for breast cancer: An Umbrella Review of Their Effects on Anxiety, Depression, Distress, and Quality of Life

Author(s)

Elena Pérez-Setién [1] , Eider Egaña-Marcos [1] , Marilia. I Gonzalez-Mojica [2] , Tania B. Huedo-Medina [1] , Natalia Alonso-Alberca [1] , Nekane Balluerka [1]

[1] University of the Basque Country; [2] University of Connecticut

Abstract

Introduction: In the last decade, there has been notable increase on oncology population secondary studies –systematic reviews and meta-analyses–focusing on complementary and integrative methods in oncology. Uncertainty remains regarding the efficacy of psychological interventions for breast cancer patients and survivors, largely due to heterogeneity among existing systematic reviews and meta-analyses results, which often focus on specific therapies or psychological interventions without conducting comprehensive moderator analyses. Furthermore, variations in methodological rigor impact the reliability and applicability of findings, underscoring the need for a systematic evaluation of methodological quality to ensure robust evidence synthesis.

Purpose: The aim of this umbrella review is to evaluate the methodological quality of the meta-analytic evidence of psychological interventions among cancer population. Additionally the review seeks to synthesize the evidence regarding the efficacy of the interventions on reducing anxiety, depression, and distress, and on improving quality of life among breast cancer patients and survivors, as well as to classify moderators that influence intervention efficacy.

Methods: A comprehensive literature search was conducted in WoS, Medline, FSTA, Scopus and PsycInfo from inception up to December 2024. Search terms encompassed keywords related to cancer, psychological interventions, and systematic review or meta-analysis. The eligibility criteria applied to studies were as follows: 1) used systematic or meta-analytic methods, 2) included randomized-controlled trials, 3) evaluated psychological interventions (with any type of comparator), 4) focused on breast cancer survivors or patients, 5) reported on at least one of the following outcomes: anxiety, depression, distress or quality of life. To assess compliance with quality standards a modified version of the Assessment of Multiple Systematic Reviews (AMSTARMedsd2) was used.

Results: Preliminary analysis of 16 meta-analytic studies exclusively focused on the breast cancer population were conducted. The modified AMSTARMedSD2 assessments showed that the meta-analyses completely satisfied from 30.77% to 61.54% of the AMSTARMedSD2 items (M = 50.48, SD = 8.14). Quality of life was the most frequently examined outcome (81.25%), while only 37.5% of the reviews assessed all four outcomes of interest (i.e. anxiety, depression, distress, quality of life). Twenty-five percent of the meta-analyses also reviewed additional lifestyle/wellbeing interventions (e.g., nutrition, exercise, acupuncture). Studies including interventions with at least cognitive-behavioral approaches were the most frequently reviewed (68.75%), followed by mindfulness-based interventions (25%). Across studies effect sizes indicated improvements in the evaluated outcomes for all participants. Commonly assessed moderators included type of intervention (evaluated in 50% of the studies), follow-ups (50%), and study quality (21.43%). Detailed analyses of intervention efficacy and the moderator patterns will be further discussed.

Conclusion: This umbrella review provides comprehensive evidence on the methodological quality of systematic reviews and meta-analyses evaluating psychological interventions for breast cancer patients and survivors. By identifying methodological strengths and weaknesses through AMSTAR-based assessments, this study highlights areas for improvement in future

reviews. The findings provide comprehensive evidence supporting the efficacy of psychological interventions in improving psychological and quality of life outcomes among breast cancer population. These insights are expected to inform future research design and contribute to the development of guidelines for complementary and integrative treatments.

1.16 Symposium : ”Artificial Intelligence and Large Language Models: Item Development and Validation, Educational Interventions, and Emotion Analysis of Videos”

Title

Leveraging AI Tutors to Enhance Student Learning: A Controlled Educational Intervention Study

Author(s)
Mariana Teles [1]

[1] University of Virginia

Abstract

This study investigates the effectiveness of AI-tutored learning environments in implementing evidence-based learning techniques among undergraduate students. Drawing from cognitive science principles, particularly those outlined in Willingham's (2023) work, we developed an innovative intervention utilizing AI tutors to simulate personalized learning environments focused on three key areas: effective note-taking, complex text comprehension, and exam preparation strategies.

In this controlled experiment, 40 first-year psychology students were randomly assigned to experimental (n=20) and control (n=20) conditions. The experimental group participated in eight structured sessions with AI tutors over one semester, while the control group maintained standard learning practices. The intervention's effectiveness is being assessed through a mixed-methods approach combining quantitative academic performance metrics with qualitative analysis of student responses.

Our analytical framework employs a novel combination of traditional pre-post comparisons and advanced natural language processing techniques. Specifically, open-ended student responses are being analyzed using zero-shot classification implemented through Facebook's BART model, complemented by sentiment analysis using the transformemotion package in R. This methodological approach allows for both systematic categorization of learning outcomes and nuanced understanding of students'emotional engagement with the AI-tutored environment.

While data collection is ongoing, this study contributes to the growing body of research on AI-enhanced educational interventions and provides a methodologically rigorous framework for evaluating their effectiveness. The findings will have important implications for implementing scalable, evidence-based learning support systems in higher education.

Title

Generative Psychometrics via AI-GENIE: Automatic Item Generation and Validation via Network-Integrated Evaluation

Author(s)
Lara Russell-Lasalandra [1]

[1] University of Virginia

Abstract
The rapid advancement of artificial intelligence (AI), particularly large language models (LLMs), has introduced powerful tools for various research domains, including psychological scale development. This study presents a fully automated method to efficiently generate and select high-quality, non-redundant items for psychological assessments using LLMs and network psychometrics. Our approach called, Automatic Item Generation and Validation via Network-Integrated Evaluation (AI-GENIE), reduces reliance on expert intervention by integrating generative AI with the latest network psychometric techniques. The efficacy of AI-GENIE was evaluated through Monte Carlo simulations using the Mixtral, Gemma 2, Llama 3, GPT 3.5, and GPT 4o models to generate item pools that mimic Big Five personality assessment. The results demonstrated improvement in item selection efficiency, with overall average increases of 9.78-17.80 in normalized mutual information in the final item pool across all models. After, each model in AI-GENIE generated a Big Five inventory that was administered to independent, representative samples (N = 1000 each) in the U.S. The empirical results show that the items produced across all models were diverse, theoretically consistent, and structurally stable. Taken together, these findings demonstrate that AI-GENIE is a highly effective tool to automate and streamline scale development and validation processes.

Title

Performance-Based Item Development and Validation in Silica: LLMs and Generative Psychometrics for Struc- tural Validity and Item Difficulty

Author(s)
Hudson Golino [1]

[1] University of Virginia

Abstract

Recent advances in large language models (LLMs) present opportunities for developing performance-based items in educational and psychological assessment. We introduce P-AI-GENIE (Performance-based Automatic Item Generation and Network-Integrated Evaluation), an extension of AI-GENIE that focuses on generating and validating performance items. The talk will cover how items can be developed and automatically validated in silica, including the estimation of item difficulty via Exploratory Graph Analysis without collecting data in humans. We seek to demonstrate P-AI-GENIE's potential for streamlining performance assessment development while maintaining measurement quality.

Title

Decoding Emotion Dynamics in Videos using Dynamic Exploratory Graph Analysis and Zero-Shot Image Classification.

Author(s)
Aleksandar Tomašević [1]

[1] Novi Sad University

Abstract

We propose a novel approach for modeling and understanding the dynamics of emotion facial expression recognition (FER) scores. Recent advancements in deep learning and transformer-based neural network architectures enable the time series analysis of FER scores extracted from images and videos. This type of data can be important for psychological research of affective dynamics and emotion expression dynamics. However, the properties of such data are not well understood in the current literature. We propose a new method to simulate FER scores based on a modified version of the Damped Linear Oscillator with a measurement model (DLO-MM). We use this model to conduct a large-scale simulation and use dynamic Exploratory Graph Analysis to investigate the dimensionality of the data and use network scores to recover the values of the latent dimensions—positive and negative sentiment of the expressed emotions. Our results show that the DLO-MM model can be used to simulate FER scores for different patterns of emotion dynamics and that DynEGA can be used to uncover the latent structure of emotion dynamics expressed through FER scores. All methods presented in the paper are implemented in the transforEmotion R package and the tutorial section provides a step-by-step guide on how to simulate FER scores using DLO-MM and how to estimate FER scores from YouTube videos using transformer-based machine learning models.

1.17 Symposium : ”Innovations in test development and validation”

Title

Constructing, Improving, and Shortening Tests for Skill Assessment with Competence-based Test Development

Author(s)
Jürgen Heller , Luca Stefanutti , Pasquale Anselmi [1] , Egidio Robusto

[1] University of Padova

Abstract

An assessment conducted within competence-based knowledge structure theory (CbKST) aims to uncover the skills that an individual possesses based on their observed responses to test items. This process involves first deriving the set of items that the individual is capable of solving (the knowledge state) from the set of items they actually solved (the response pattern), and then inferring the set of skills the individual has available (the competence state) from the knowledge state. A good test ensures that uncertainty about the individual's competence state is as small as possible. Competence-based test development (CbTD) is a recent method for constructing tests proposed within CbKST. It exploits concepts originally introduced in rough set theory to construct tests that are as informative as possible about individuals'competence states (i.e., adding any item does not increase the informativeness of the tests) and, if desired, also minimal (i.e., no item can be eliminated without reducing the informativeness of the tests). Given a fixed set of competence states that exist in a population of individuals and a fixed set of competencies (each of which being the set of skills required to solve an item), CbTD produces tests that differ in the competencies but are all equally informative about individuals' competence states. Both conjunctive and disjunctive tests can be developed. In conjunctive tests, all skills associated with an item are necessary for solving it, whereas in disjunctive tests, any of the skills associated with an item is sufficient for solving it. The talk presents CbTD and illustrates some real-life applications to the construction of a test from scratch, and the improvement and shortening of existing tests.

Title

Exploratory Structural Equation Modeling (ESEM) in Comparison with CFA Models

Author(s)

Palmira Faraci

Abstract

Psychometric evaluations of psychological assessment measures have shown that several instruments produce inconsistent factor structures across groups and contexts and provide questionable reliability and predictive validity. A key conceptual issue concerns how a theoretical construct is defined vs. how it is measured. Given that psychological constructs cannot be observed directly, but only inferred through rating scales, the methodology used to validate psychometric instruments may be the central issue. When cross-loadings are constrained to zero in estimation models, dynamic interactions between factors cannot be captured. Therefore, more innovative approaches to scale validation may be needed.

Exploratory Structural Equation Modeling (ESEM) has emerged as a viable option for overcoming some of these challenges, combining the finest features of exploratory and confirmatory factor analysis within the traditional SEM framework (Asparouhov & Muthén, 2009; Morin et al., 2020). Therefore, this contribution focuses on ESEM as a technique that provides a compromise between the mechanical iterative approach of finding optimal factorial solutions through rotations and the restrictive a priori theory-driven modeling approach to promote the rational use of a methodology that can support a clearer representation of the complexity of psychological constructs (Marsh et al., 2014). The purpose of this presentation is to provide a brief overview of ESEM and results from empirical studies comparing ESEM and CFA models.

Specific types of ESEM are presented as useful strategies to extend the applicability of this technique within more complex analytical frameworks. Set-ESEM enables the simultaneous estimation of multiple constructs and finds an optimal balance between CFAs and Full-ESEMs in terms of parsimony, data-model fit, rigor, flexibility, and well-defined factor estimation (Marsh et al., 2020). ESEM-within-CFA allows for the re-specification of an ESEM model within a CFA framework for more complex research questions (e.g., hierarchical structures, partial mediation, longitudinal mediation, latent change score models) (Morin & Asparouhov, 2018).

The comparison between two 4-factor solutions with 20 items and 26 cross-loadings ($|\lambda| = .103 - .417$, M = .174) reveals a reduction in correlations between factors: CFA ($.63 < r < .81$, Mr = .74), ESEM ($.49 < r < .74$, Mr = .61). The comparison between two 2-factor solutions with 7 items and 3 cross-loadings ($|\lambda| = .130 - .208$, M = .16) shows a reduction of the factor correlation as follows: CFA ($r = .63$), ESEM ($r = .56$). The comparison between two 3-factor solutions with 10 items and 12 cross-loadings ($|\lambda| = .101 - .444$, M = .234) shows a reduction of the factor correlations: CFA ($.74 < r < .79$, Mr = .77), ESEM ($.37 < r < .46$, Mr = .40).

The choice of the “best”model reflects a combination of adherence to theory and research question, goodness of fit, interpretation of parameter estimates, and parsimony. Of course, the choice is rarely so straightforward when based on real data, and researchers must balance goodness of fit, parsimony, theoretical considerations, and interpretation of parameter estimates. Golden rules about which models are best are inappropriate and even counterproductive.

Title

Addressing Ordinal Variables through Integrated IRT and CTT Methods in Cultural Capital Measurement

Author(s)
Anselmi Pasquale , Aurora Castellani [1] , Roberto Cubelli , Giulia Balboni [2]

[1] University of Perugia; [2] University of Bologna

Abstract

In quantitative measurement, Likert scales are often treated as continuous variables, potentially distorting results due to their ordinal nature. This study addresses the issue of appropriately handling ordinal variables by integrating classical test theory (CTT) and item response theory (IRT) to validate a novel Scale of Cultural Capital (SCC). SCC consists of 14 items measuring three dimensions: cultural fruition, cultural technical skills/knowledge, and involvement in groups/associations (Balboni et al., 2019). The SCC was administered online to 923 adults, 51% women, aged 20 to 66 years M(SD) = 41.70(12.44), with an educational level lower/equal (48%) or higher (52%) than a high school degree.

First, the original 5-point response scale was reduced to 4 points due to underrepresented response categories, with contiguous low-frequency categories being merged. Second, exploratory factor analyses were conducted on a random half-group of participants (n = 461), using the weighted least squares method, oblimin rotation, and a polychoric matrix (KMO = .83; Bartlett's test p < .05), as suggested for ordinal data. Based on the Parallel Analysis and MAP test, alternative solutions from 5 to 1 factors were explored. The results confirmed the three-factor solution as the most appropriate, consistent with the theoretical model. Third, confirmatory factor analysis conducted in the remaining participants using the DWLS method for ordinal data showed that the three-factor model exhibited an adequate fit (CFI = .961, SRMR = .075, RMSEA = .069) and was better than alternative one- and two-factor models. Cronbach's ordinal alpha (Zumbo et al., 2007) revealed good scale reliability (α = .84).

Invariance analyses for gender, age, and education level were conducted on the total group, comparing nested models with progressive constraints (configural, metric, scalar with threshold constraints to ensure equivalent ordinal category boundaries, and comparison of latent means) also using RMSEA_D (Savalei et al., 2023). Scalar invariance was achieved across gender (CFI = .958; RMSEA = .063; SRMR = .074), with women showing higher latent means for cultural fruition (Cohen's d = .31) and cultural technical skills/knowledge (d = .12). Partial scalar invariance across age (CFI = .957; RMSEA = .065; SRMR = .071) was achieved by freeing the thresholds of the foreign language usage item, as younger participants required a lower latent level to select higher response categories. Younger participants showed latent means that were lower for involvement in groups/associations (d = -.24) and cultural fruition (d = -.07), but higher for cultural technical skills/knowledge (d = .42). Freeing the loadings of two items allowed for achieving partial metric invariance and scalar invariance across educational levels (CFI = .923; RMSEA = .072; SRMR = .084). Participants with a higher educational level showed higher latent means on all dimensions of cultural capital. Concerning IRT, the RMSEA value of all items was below .05, indicating an overall good item fit.

Utilizing suitable methodologies for ordinal variables, the present study validated the three-factor structure of the SCC and its stability across gender, age, and level of education.

Title

On the Way to State Specific Response Errors: A Generalized Local Independence Model

Author(s)
Jürgen Heller , Alice Jenisch [1]

[1] University of Tübingen

Abstract
Knowledge structure theory is a psychometric approach for representing the knowledge of participants in a precise, non-numerical way. The most prominent probabilistic model in knowledge structure theory is the basic local independence model. One of its fundamental assumptions is the constancy of the response error probabilities (guessing and slipping) across all participants. However, it seems to be implausible that a student with no knowledge in a domain guesses the correct answer of an item with the same probability as an experienced student, who is ready to learn the item. Therefore, it would be desirable to let an item's error probabilities depend on a person's knowledge state and, in particular, on how close the item is to the knowledge state in some proper sense. Different options of capturing the discrepancy between an item and a knowledge state are discussed, and first results of simulation studies based on a generalized local independence model with state-dependent error probabilities are presented.

Title

Probabilistic Information and Network Evaluation System (PINES): A Bayesian Framework for Advancing Psychometric Testing

Author(s)
Matteo Orsoni [1] , Giulia Balboni [1] , Sara Giovagnoli , Sara Garofalo , Noemi Mazzoni , Matilde Spinoso , Mariagrazia Benassi

[1] University of Bologna

Abstract

As educational and cognitive assessments advance, there is a growing need for innovative, evidence based methodologies that offer deeper insights into students'abilities, knowledge representation, and response reliability. Contemporary assessment systems face the challenge of capturing nuanced insights into student learning while ensuring measurement validity, going beyond traditional scoring, offering valuable perspectives on students'cognitive processes, knowledge structures, and response patterns while detecting potential validity threats such as rapid guessing or cheating behavior. We propose the Probabilistic Information and Network Evaluation System (PINES), a novel framework that integrates Bayesian networks (BNs) and information theory to enhance psychometric scoring and reliability assessment. PINES incorporates item interdependencies and provides deeper insights into the cognitive processes underlying these dependencies. The framework begins by constructing a Directed Acyclic Graph (DAG) to model item relationships, from which conditional probabilities for each item are calculated based on responses to related items. This approach ensures that the scoring system accurately captures the interconnected nature of test items. PINES employs self-information to measure the "surprise"or unexpectedness of each response, given its expected probability derived from the DAG. This allows to generate a weighted score that reflects the informativeness of each response, allowing also the framework to identify potentially anomalous or unreliable responses. By computing confidence intervals, PINES enhances the interpretability and robustness of the results. Furthermore, the framework employs entropy-based metrics to evaluate the uncertainty in response distributions. For each item, PINES measures how a respondent's answers deviate from the sample average entropy, enabling the detection of specific cognitive patterns or difficulties in their responses. This granular analysis provides deeper insights into individual response strategies and potential inconsistencies. To assess overall reliability, PINES calculates a weighted reliability score for each respondent based on the number of incoherent and highly improbable responses. This score is normalized, with lower values indicating higher reliability, offering a clear and quantifiable measure of response consistency. To demonstrate its practical utility, we applied PINES to the Raven's Colored Progressive Matrices (CPM), using a Bayesian network developed in a prior study involving a sample of 40 first-year primary school children (mean age = 6.68 ± 0.36 years; 52.5% male). We selected three cases with identical total scores, two real respondents and one with random responses, to illustrate how PINES analyzes individual response patterns, detects improbable responses, and assesses reliability. Regarding the three cases PINES identified two highly improbable responses in the first case, yielding a high reliability score (0.172), while in the second with three incoherent responses received a medium reliability score (0.207). In the random response case, eight incoherent and five highly improbable responses were flagged, resulting in a low reliability score (0.534). In conclusion, PINES represents a novel perspective in psychometric methodologies. By explicitly modeling item dependencies and leveraging information theory, it provides a more accurate -at the individual level - and detailed assessment of response reliability. Furthermore, PINES is adaptable to a wide range of psychometric tests and contexts, making it a versatile tool for cognitive testing and beyond.

1.18 STATE OF THE ART: AI: Where Are We and Where Are We Going

Title

STATE OF THE ART: AI: Where Are We and Where Are We Going

Author(s)
Juan Albino Méndez Pérez [1]

[1] University of La Laguna

Abstract

This talk explores the core principles and practical applications of AI. We begin by defining AI as the discipline that imbues machines with human-like intelligence, encompassing reasoning, learning, and creativity. Key characteristics include the ability to perceive, interact, solve problems, act autonomously, and adapt to environments. We will cover the diverse problems AI addresses, such as classification, regression, prediction, clustering, optimization, and Natural Language Processing (NLP), alongside content generation. The presentation traces AI's evolution from symbolic AI to Machine Learning, Deep Learning, and the transformative rise of Generative AI. We will delve into Large Language Models (LLMs) like GPT and GEMINI and the current technologies based on Agentic AI. The global impact of AI is undeniable, with its interdisciplinary nature driving widespread applications across various sectors, significantly improving efficiency and enabling new capabilities worldwide. The presentation will finish analysing the profound influence of this technology on education and research. AI's intrinsic capabilities in learning, reasoning, communication, and creativity are directly applicable, assisting with academic text analysis, content creation, and report generation. In this scenario AI is becoming an indispensable assistant for students, reasearchers and educators alike, with autonomous AI Agents poised to further revolutionize these fields.

1.19 STATE OF THE ART: On Transforming Revisited

Title

STATE OF THE ART: On Transforming Revisited

Author(s)
Anthony Onwuegbuzie [1]

[1] University of Cambridge

Abstract

In this State of the Art Address, I revisit and extend the conceptual boundaries of two core mixed methods transformation techniques: qualitizing and quantitizing. In so doing, I spotlight the expanded methodological and philosophical dimensions that elevate their application in contemporary mixed methods research. The first third of the presentation is dedicated to qualitizing, defined as the transformation of quantitative data into qualitative form that can be analyzed qualitatively. I will outline how qualitizing has evolved to include five major elements: (1) it can yield numerous representations (e.g., narratives, profiles), (2) it can stem from either quantitative or qualitative data, (3) it may involve either qualitative or quantitative analyses, (4) it can be applied as a single or multiple analyses, and (5) it can produce a fully integrated analysis. Special emphasis will be placed on narrative profile formation—such as modal, average, holistic, comparative, and normative profiles—which allows for rich, contextualized interpretations of numerical data.

In the second third of the address, I will introduce the DIME-Driven Model of Quantitizing, which encompasses four core classes of Level 1 quantitizing:

• Descriptive-Based Quantitizing transforms qualitative data into quantitative metrics to summarize patterns using measures such as mean, standard deviation, percentiles, and skewness.

• Inferential-Based Quantitizing involves converting qualitative data into formats suitable for statistical inference, including tests such as analysis of variance (ANOVA), regression, and structural equation modeling.

• Measurement-Based Quantitizing refers to the transformation of qualitative insights into quantifiable constructs for instrument development and validation, often using techniques such as Rasch modeling and Item Response Theory (IRT).

• Exploratory-Based Quantitizing converts qualitative data into numerical formats to explore underlying patterns, relationships, or structures through methods such as factor analysis, cluster analysis, and correspondence analysis.

In this presentation, I will also introduce for the first time a novel concept, which I call Transformatizing. Transformatizing refers to the integrated process of applying both qualitizing and quantitizing techniques within a single analytical framework fully to harness and to interweave the strengths of qualitative and quantitative data transformations. It represents a dynamic, bidirectional approach wherein data are fluidly transformed across paradigms to achieve comprehensive, meta-integrative insights. Major components of transformatizing are QuanQualitizing and QualQuantitizing—both of which will be defined.

To concretize these ideas, I will present a real example from the published literature that illustrates both QuanQualitizing and QualQuantitizing in action.

Throughout the session, I will illustrate these expanded definitions with practical examples from diverse research contexts. Attendees will leave with a clearer understanding of how thoughtfully transforming data across traditions/paradigms not only enriches methodological rigor, but also facilitates deeper, more meaningful meta-inferences. I invite colleagues to consider how these advanced transformation techniques can further democratize evidence, foster integration, and propel mixed methods research into new frontiers.

1.20 Poster Session 2

Title

Combining Prolonged Exposure and Compassion-Focused Therapy for PTSD: A Case of Multilevel Analysis for Single-Experimental Case Designs

Author(s)
Cristina Rodríguez Prada [1] , Mateo Bernal Navas [1]

[1] Universidad Autónoma de Madrid

Abstract

Single-case experimental designs (SCEDs) provide valuable insights into psychological interventions but often require advanced statistical techniques to account for within- and between-subject variability. This study employs a multilevel model to analyse the effects of combining Prolonged Exposure (PE) and Compassion-Focused Therapy (CFT) for PTSD, particularly in individuals with trauma-related shame and guilt. Using a withdrawal crossover SCED (N=4), participants alternated between active listening (A), PE (B), and CFT (C) under two conditions: Condition 1 (A/C/B/C/B) and Condition 2 (A/B/C/B/C). A multilevel approach was applied to model treatment effects over time, capturing intra-individual changes, phase transitions, and cross-condition comparisons. Results indicate that both PE and the combined intervention significantly reduced PTSD symptoms, with the CFT-enhanced approach yielding greater reductions in shame. The multilevel analysis further revealed individual differences in treatment response trajectories, highlighting the importance of guilt, experiential avoidance, and identity-related variables in moderating outcomes. These findings demonstrate the utility of multilevel modelling in SCEDs, offering a robust framework for analysing psychological interventions with small samples.

Title

Methodological Innovations in Forecasting the Global Burden of Antimicrobial Resistance: Integrating Predictive Modeling, Scenario Analysis and Data-Driven Insights

Author(s)
Benn Sartorius [1] , Simon I Hay [2] , Anna Gershberg Hayoon [2] , Eve E Wool [2] , Authia P Gray [2] , Lucien R Swetschinski [2] , Stein Emil Vollset [2] , Rebecca L Hsu [2] , Ben S Cooper [3] , Christopher J L Murray [2] , Daniel T Araki [2] , Mohsen Naghavi [2] , Chieh Han [2] , Erin Chung [2] , Nicole Davis Weaver [2] , Kevin S Ikuta [2] , Andy Stergachis [2] , Tomislav Meštrović [4] , Gisela Robles Aguilar [3] , Christiane Dolecek [3]

[1] The University of Queensland, Australia / Institute for Health Metrics and Evaluation, US; [2] Institute for Health Metrics and Evaluation, US; [3] University of Oxford, UK; [4] University North, Croatia / Institute for Health Metrics and Evaluation, US

Abstract

The global burden of antimicrobial resistance necessitates advanced methodological frameworks in order to capture historical trends and provide reliable future projections. In the Global Research on Antimicrobial Resistance (GRAM) project, which is a joint partnership between the University of Oxford and the Institute for Health Metrics and Evaluation (IHME) at the University of Washington, we used sophisticated epidemiological modeling approach to estimate antimicrobial resistance burden from 1990 to 2021 and forecast outcomes until 2050 across 204 countries and territories. By integrating diverse data sources –including cause-of-death records, hospital discharge data, microbiological surveillance, pharmaceutical sales, antibiotic use surveys, healthcare utilization data and literature reviews –our approach processes 520 million individual records or isolates across 19,513 study-location-years.

The forecasting methodology is built on the Global Burden of Disease (GBD) 2021 study's statistical framework and employs a probabilistic forecasting model that incorporates historical trends, antimicrobial resistance patterns, as well as variations in healthcare systems. This is the first time such extensive methodology was ever employed in the analysis of antimicrobial resistance burden. Such approach employs three distinct scenarios: a reference scenario based on observed trends, a Gram-negative drug development scenario simulating the impact of new antimicrobial agents targeting Gram-negative bacteria, and a better care scenario assuming improvements in healthcare access, infection control and antimicrobial stewardship.

To refine predictions, spatiotemporal Gaussian process regression (ST-GPR) is applied to smooth resistance prevalence estimates across time and geography while addressing data sparsity. Bayesian hierarchical modeling further enhances the robustness of the estimates by integrating historical trends and regional covariate relationships. A decomposition analysis adapted from Das Gupta's framework quantifies the relative contributions of demographic shifts, healthcare improvements and antimicrobial usage changes to the projected antimicrobial resistance burden. Counterfactual scenario modeling is employed to differentiate the burden of resistance, considering both a scenario in which all resistant infections are replaced with susceptible infections and another where infections are entirely eliminated. Monte Carlo simulations are incorporated to quantify uncertainty, generating confidence intervals for projected deaths related to antimicrobial resistance and disability-adjusted life years (or DALYs). Data integration and validation are ensured by benchmarking estimates against historical trends and validating projections using independent epidemiological studies and national surveillance data.

By combining predictive modeling, historical trend analysis and robust uncertainty quantification, this methodological framework establishes a reproducible and scalable approach to global AMR burden estimation. Such forecasting methodology not only informs targeted intervention strategies but also serves as a foundation for broader epidemiological modeling of global health threats, emphasizing the need for continued refinement in predictive epidemiology.

Title

Methodological differences between formative-measured and composite variables: a case study using mixed SEM techniques

Author(s)

Cynthia Maria Delgado Garcia [1] , Daniel Ondé [1] , Jesús Mª Alvarado-Izquierdo [1]

[1] Complutense University of Madrid

Abstract

An essential step in psychometric analysis is the choice of the correct modeling approach (i.e. common factor or composite) and the type of indicators (i.e. reflective and/or causal-formative) of the variable(s) of interest. However, formatively-measured costructs (i.e. variables whose indicators are all or partially causal-formative, FMCs) and composite variables are still commonly confused. In the context of structural equation modeling (SEM), the literature on models in which a FMC occupies a structural endogenous position could serve as a nice example of differences between FMCs and composite variables. Furthermore, a number of different techniques were proposed to simultaneously estimate latent and composite variables. This study compared several models in which: a) indirect (correct) and direct (incorrect) specifications of variables influenciating endogenous FMCs, b) variance-based and covariance-based (CB-SEM) techniques; and c) common factor, composite and mixed (CB-SEM with the Henseler-Ogasawara specification and consistent PLS) approaches are used. The sample consisted of 362 students (12-17 years old) in compulsory secondary education. The results included a) a comparison between loadings, weigths and path values; and b) coefficients to assess the overall fit of the models. The conceptual coherency of each model was also examined. Further studies, specially simulation studies, are needed to analyse the behavior of these models in this context.
Funding: This research was supported by the grant PID2022-136905OB-C22 and an FPU grant (predoctoral contract FPU23/02914), both funded by the Ministry of Science and Innovation –Ministerio de Ciencia e Innovación–.

Title

Prioritization of socio-political values while developing and implementing driverless mobility: a multi-method, multi-step intercultural research approximation

Author(s)

Kerstin Kusch [1] , Miguel Nuñez de Prado Gordillo [2] , Leandro L. Di Stasi [3] , Giovanni Bruno [4] , Sebastian Pannasch [1] , Carolina Diaz-Piedra [3] , Andrea Spoto [4]

[1] Technische Universität Dresden; [2] University of Rijeka; [3] University of Granada; [4] University of Padova

Abstract

Although research on automated driving (AD) dilemmas has helped to focus not only on the AD technical dimension, but also on its socio-political and ethical aspects, it has neglected the study of more realistic scenarios. Decisions regarding more ordinary value trade-offs (e.g., sacrificing personal privacy in favor of enhanced performance functions) will be pervasive in most everyday AD situations. Here, we carried out a multi-method, multi-step research process to gain insight into such value prioritization schemes. First, as a result of several multidisciplinary workshops and an expert-based assessment, we defined a set of socio-political values that represent the framework of this investigation: Privacy, Autonomy, Safety, Security, Performance, and Costs. Such values guided the subsequent interviews with international experts and stakeholders, and focus groups with users, with the aim to define socio-political, legal and ethical issues. Then, a new round of expert assessment was carried out. As a result, the revised initial issues were reorganized around five key thematic areas: Data Management, Legal Protection, New Driver Roles, Decisions on automated vehicle (AV) behaviour, and Security/Hacking. Then, we developed and transformed the issues into a list of must-haves –key recommendations –and functional requirements for AD. Finally, we collected the views of 433 respondents from three different countries (Germany, Italy, and Spain) to compare the value prioritization schemes across different driving cultures. Our interest focused on both other- (i.e., a relative importance order between the system of values when evaluated from the perspective of three potential type of AVs users - a mobile office scenario, a truck driver, an elder - acting in particular traffic conditions) as well as (absolute) self-referred evaluations. The definition of the relative and absolute order of AV's system of values was investigated through a mixed approach, that combined classical experimental material - closed-ended questions - with innovative video-based storytelling for the presentation of the stimuli. Regarding the self-referred evaluations, results indicated different preference levels among different values, confirmed by an Inductive Item Tree Analysis. This analysis showed a partial order having "Safety"at the top, "Autonomy"and "Security"on the same level with "Performance"and "Privacy"below them and on the same level, while "Cost"was at the bottom of this hierarchy, being outperformed by all the other values. Also, we found a higher preference for the value "Performance"in Spain compared to Germany, and the opposite pattern for the value "Security". Regarding other-referred evaluations, while the value "Safety"was the most important, irrespective of the type of user, some differences emerged for the other values. For example, for someone who uses the AV as a mobile office, "Performance"and "Security"were immediately below "Safety", with no order between them. The remaining values were, in relevance decreasing order, "Autonomy", "Privacy" and "Cost". Our analysis of which different socio-political values are at stake in routinary traffic situations (and the interaction between traffic infrastructures, legal frameworks, attitudes and values) and how different cultural groups might prioritize different values may help the standardization and regulation of future AD technologies.

Title

SUPERVISED MACHINE LEARNING MODELS FOR AUTOMATIC DETECTION OF MULTIPLE SCLEROSIS THROUGH ACOUSTIC ANALYSIS OF THE VOICE

Author(s)

Jonathan Delgado Hernández [2] , Miguel Ángel Hernández Pérez [1] , Moisés Betancort Montesinos [2] , Tatiana Romero Arias [3]

[1] Hospital Nuestra Señora de La Candelaria; [2] Universidad de La Laguna; [3] Universidad Europea de Canarias

Abstract

Multiple sclerosis (MS) is an autoimmune and neurodegenerative disease of the central nervous system of unknown etiology and is considered the most common cause of non-traumatic disability in young adults. Early detection of this disease is challenging due to the wide variety of symptoms. Voice biomarkers are particularly relevant to research and clinical practice because they provide objective, naturalistic information about motor function in neurodegenerative diseases such as MS.

The objective of this study was to examine supervised machine learning algorithms that facilitate the automated detection of MS through the analysis of acoustic parameters of the voice.

Two hundred people with a mean age of 47.1 years (SD=12.24) without organic voice disorders were included in the study, 120 diagnosed with MS (pwMS) and 80 neurologically healthy (pNH). Both groups were gender-balanced (⊠2=1.97, p=0.159). Each participant recorded a four-second sustained vowel with the Praat and 15 acoustic parameters were obtained using the Voxplot program. The sample was divided into 80% for training and 20% for testing. To identify the most relevant acoustic parameters for machine learning (ML) models, an elastic net model was applied. A 10-fold cross-validation on the training sample was used to obtain the best alpha and lambda that were then used in the test set was performed to determine optimal alpha and lambda values, which were then applied to testing set to select the final set of acoustic variables. Subsequently, the selected variables were used to train ML models, including Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and eXtreme Gradient Boosting (XGB). To evaluate the robustness of the results, a simulated sample was generated from the original data using the mvrnorm function from the MASS package, and the same procedure was repeated. A balanced sample of 510 (255 pwMS and 255 pHS) was obtained according to the prevalence of the disease in the Canary Islands (255 per 100,000 inhabitants).

In the real sample, the RF model demonstrated the highest level of accuracy (accurate=0.82, sensibility=0.81, specificity=0.84, ROC-AUC= 0.84). The SVM and XGB models also demonstrated satisfactory performance, exhibiting an acceptable level of accuracy and balanced sensitivity and specificity. Conversely, the DT and KNN models exhibited high sensitivity but very low specificity. In the simulated sample, the RF and XGB models demonstrated significant enhancement with perfect sensitivity (accuracy=0.99, sensibility= 1, specificity=0.98, ROC-AUC=1, in both models). The SVM model demonstrated an enhancement in its performance, while the DT and KNN models exhibited an increase in specificity, accompanied by a decrease in sensitivity. The results indicate that increasing sample size improves the performance of bagging and boosting models.

Acoustic voice analysis is a technique that can detect minimal alterations in motor function. The use of ML methodologies, such as RF (or XGB when the data set is sufficiently large), allows high levels of diagnostic accuracy to be achieved.

Title

A comparison of algorithms for tests of variance components in genetics ACE models

Author(s)
Olivier Vivier [2] , Pier-Olivier Caron [1]

[1] Université TÉLUQ; [2] Université du Québec à Montréal

Abstract

Using structural equation modeling, the genetic and environmental similarities (and dissimilarities) between monozygotic and dizygotic twins can be decomposed into three components: additive genetics (A), common environment (C) and unique environment (E). After identifying the ACE model, model fitting is examined to evaluate the improvement of the model after dropping one or more of the components to assess whether they should be retained (Maes, 2014). Three approaches, hereafter referred to as Saturated, Components and Estimates are used : a) comparison of a saturated model to components of the ACE model (i.e., ACE, AE, CE and E models); b) nested comparison of the ACE model's components, and; c) assessing the component based on its parameters estimates in the ACE model. To our knowledge, these three approaches found in the scientific literature were never compared, yet, they may vary in their accuracy and may lead to different conclusions regarding the optimal model. This Monte Carlo simulation study aimed to examine the accuracy of these different algorithms when varying specifications of the ACE model's components and sample sizes. Overall, the results show that the Components algorithm is the best to recover the correct models. However, they also show that the different algorithms struggle to identify ACE models even when parameters are moderate (A = .30, C = .30) and sample sizes are large (ns <= 500) . The Components algorithm approach outperforms the other algorithms when A or C is at zero whereas the Estimates and Components algorithms perform similarly when both A and C are non-zero. The saturated algorithm had the poorest performance overall being only better than Estimates in the A=0 or C=0 conditions, but still much worse than Components. However, in all cases, very large sample sizes are required to reach sufficient accuracy. The current results shed light on the absence of consensus and lack of directives on tests of variance components in ACE models.

Title

Assessing Bias in Non-Randomized Studies: A Systematic Review of Evaluation Tools

Author(s)
María José Cabañero-Martínez [1] , Mar Lozano-Casanova [1] , Néstor Montoro-Pérez [1] , María Rubio-Aparicio [2] , María Sánchez-Marco [1] , Silvia Escribano [1]

[1] University of Alicante; [2] University of Murcia

Abstract

In the domain of health sciences, quasi-experimental studies are extensively utilized due to their feasibility and cost-effectiveness. However, this design type may introduce biases that potentially affect both the validity of results and the decision-making processes of healthcare professionals. Currently, there is an observed proliferation of tools for assessing the risk of bias in quasi-experimental studies, which complicates the selection of the most appropriate instrument, as each possesses distinct psychometric properties that may influence the accuracy and reliability of the evaluation.

The objective of this study is to analyze and compare the psychometric properties of available tools for assessing the risk of bias in quasi-experimental studies, with the aim of identifying which offers superior precision, reliability, and utility to ensure more robust methodological evaluations.

A systematic review was conducted, encompassing searches in PubMed, CINAHL, Web of Science, and Scopus databases to ensure comprehensive coverage of relevant literature. Furthermore, specialized journals and the bibliographies of included studies were examined to identify additional articles. This process facilitated the identification of a corpus of articles for subsequent analysis.

Data on psychometric properties were extracted from individual studies, including measures such as reliability, validity, and measurement error. Meta-analytic computations were performed where applicable to synthesize findings quantitatively. The results were compared with those from existing meta-analyses to evaluate consistency and robustness.

The potential implications of errors and inconsistencies in this process are analyzed.

Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

Title

Bootstrap-F and adjusted F-tests in split-plot designs: The effect of non-sphericity and heterogeneity on Type I error

Author(s)

F. Javier García-Castro [1] , Guillermo Vallejo [2] , Rafael Alarcón [3] , María J. Blanca [3] ,
Roser Bono [4] , Jaume Arnau [4]

[1] Universidad Loyola Andalucía; [2] University of Oviedo; [3] University of Malaga; [4] University of Barcelona

Abstract

Background. Adjusted F-tests are not robust to simultaneous violation of the assumptions of sphericity and homogeneity of covariance matrices when group sizes are unequal. As an alternative, the bootstrap-F (B-F) method has been proposed. Objective. The aim of the study was to analyse the robustness of the Greenhouse-Geisser (F-GG) and Huynh-Feldt (F-HF) tests and the B-F method in split-plot designs under non-sphericity and heterogeneity of covariance matrices between groups. It is hypothesized that the B-F method will outperform the F-GG and F-HF tests. Method. A simulation study was conducted using a split-plot design with two groups and three repeated measures under normal distribution. The manipulated variables were: (a) total sample size, (b) group size, (c) epsilon value (Greenhouse-Geisser estimation), (d) coefficient of sample size variation, (e) homogeneity/heterogeneity of covariance matrices and (f) pairing between group size and covariance matrices (null, positive and negative). Type I error rates for time and interaction effects were computed and results were interpreted according to Bradley's liberal criterion, whereby a procedure is robust if the Type I error rate is between 2.5% and 7.5%. Results. Adjusted F-tests for time and interaction effects were conservative with positive pairing (variance heterogeneity) and a moderate or high coefficient of sample size variation. They were liberal with negative pairing (variance heterogeneity) and a moderate or high coefficient of sample size variation. The B-F method was robust in all conditions when the coefficient of sample size variation was low or moderate. However, with a high coefficient of sample size variation and negative pairing, the B-F method was robust only when N > 20. Non-sphericity had little effect on the Type I error rates of all statistics. Conclusions. The B-F method showed greater robustness than did the F-GG and F-HF tests, making it a valid option for analysing split-plot designs under heterogeneity of covariance matrices in the conditions studied here. Sample size above 20 is recommended. This research was supported by grant PID2020-113191GB-I00 from the MCIN/AEI/10.13039/501100011033.

Title

Learning patterns and reading comprehension in primary education: a mixed methods approach

Author(s)
J. REINALDO MARTÍNEZ-FERNÁNDEZ [1] , Anna Ciraso-Calí [1]

[1] Universitat Autònoma de Barcelona

Abstract

The analysis of learning patterns (Vermunt, 1998) has generated an interesting insight in Higher Education. However, this model has been scarcely analyzed in the field of Primary Education (Martínez-Fernández, et al., 2021). In addition, this line of research has been based almost exclusively on the use of a self-report questionnaire (the Inventory of Learning patterns of Students, ILS). Therefore, in this study we extend the research in this area with a mixed methods approach in the field of Primary Education. Thus, two contributions are oriented to enrich the line of research on learning patterns; on the one hand, Primary Education students who live in a socio-economically vulnerable territory; and on the other hand, we deepen the analysis of learning patterns using the semi-structured interview and the focus group for a joint analysis (meta-reflection) on the results obtained. This paper aims to discuss the contributions of mixed methods to the analysis of learning patterns in relation to reading comprehension as evidence of predictive validity. In this way, we can gain a better understanding of learning processes from the earliest stages of development, based on different points of view (children's and tutors'). A total of 218 primary school students and their tutors (N=3) from 4th, 5th and 6th grades participated. The study was carried out in the periphery of the city of Murcia (Spain) in socially vulnerable areas. Data was obtained from the ILS inventory and interviews with tutors and a focus group with students. This study is part of a larger pretest-posttest design project. However, here we have focused on the analysis of the information provided by the different instruments, and on the role of learning patterns (pretest) in the explanation of reading comprehension (posttest). Patterns of learning are identified from children's self-reported information, like students in Higher Education. However, patterns identified from interviews with tutors were found to be more reliable and correlated with reading comprehension levels. Children with an Undirected pattern have the lowest, even worrying, results. Additionally, the mixed methods approach provides relevant data in understanding learning patterns. We believe that the identification of learning patterns is an excellent tool to provide differentiated learning actions aimed at individual remediation of learning difficulties. In this same sense, we consider that the "remedy" cannot be the same for everyone, nor the same dose, since people are in different learning profiles, and this calls for personalized learning itineraries. Finally, approaching the reality of classrooms, learning processes, and teaching from mixed methods is a potential for educational research and for the design of actions for change that should be encouraged.

Title

An Applied Case of Longitudinal Factorial Invariance: data imputation, limitations and suggestions

Author(s)
Vanessa Elizabeth Da Silva Larez [1] , Daniel Ondé [2]

[1] Universidad Autónoma de Madrid; [2] Complutense University of Madrid

Abstract

This study focuses on longitudinal factorial invariance analysis, emphasizing the need to study equivalence between repeated measurements. The aim is to illustrate the analysis process in R using the Hedonic and Arousal Affect Scale (HAAS) with longitudinal real ordered-categorical data, following this step: (1) the identification of a catalyst and the formulation of hypotheses related to the Response Shift Theory, (2) the treatment of missing values, (3) consideration of prerequisites for longitudinal invariance analysis, and (4) model evaluation. Five levels of factorial invariance (configural, thresholds, metric, scalar, and strict) are established, and it is observed that, although invariance is maintained in the early weeks, potential Response Shift effects begin to emerge. Results indicate recalibration in specific items belonging to both high and low arousal negative affect factors. It is recommended to include the invariance longitudinal analysis in the planning phase from the design stage of longitudinal studies to obtain robust inferences, and reflections are made on the inherent limitations of the procedure.

Title

Normative data on the execution age of action-related sentences in young and older adults

Author(s)
Shivani Daryanani [2] , Maria A. Alonso [1] , Agustina Birba [3]

[1] Universidad de La Laguna, Instituto Universitario de Neurociencia (IUNE), Integración en la Comunidad (INICO); [2] Universidad de La Laguna, Instituto Universitario de Neurociencia (IUNE); [3] Instituto Universitario de Neurociencia (IUNE)

Abstract

The enactment effect is a well-known mnemonic phenomenon that reflects better memory for action-related sentences when participants physically perform the described action compared to when these sentences are processed only at a verbal level, without enactment. To study this effect, it is essential to have appropriate materials that allow for manipulation and control by the experimenter. The aim of the present study was to obtain normative data on the subjective age of execution of action-related sentences in two population groups: young adults and individuals over 60 years old. A total of 536 action-related sentences from the study by Díez-Álamo et al. (2019) were used. Participants were asked to indicate the age at which they believed they had performed the action described in each sentence (e.g., "bite an apple"). The analyses revealed that participants over 60 years old assigned a later execution age (M=9.37) than university students (M=6.28). Additionally, correlational analyses were conducted with other sentence dimensions such as familiarity, emotionality, motor activity, memorability, and vividness. The results showed correlations in both groups between execution age and the dimensions of familiarity, motor activity, memorability, and vividness. Specifically, sentences with an earlier execution age were perceived as more familiar, more memorable, more vivid, and associated with lower motor activity. Emotionality showed a significant relationship in the group over 60 years old, indicating that this factor plays a relevant role in the perceived execution age of actions in this population. These normative data constitute a valuable resource for studying how actions affect cognition, particularly action-event memory, from the perspective of embodied cognition theory.

Title

Robustness of repeated measures ANOVA with non-normal data and very small sample size

Author(s)

F. Javier García-Castro [1] , Rafael Alarcón [2] , María J. Blanca [2] , Roser Bono [3] , Jaume Arnau [3]

[1] Universidad Loyola Andalucía; [2] University of Malaga; [3] University of Barcelona

Abstract

Background. Recent studies have shown that repeated measures ANOVA is generally robust to violations of normality, provided that the assumption of sphericity is satisfied. However, further research is needed as these studies do not consider sample sizes smaller than 10. Objective. To analyse the robustness of the F-statistic in terms of Type I error rate with 19 non-normal distributions (both known and unknown) with skewness values ranging from 0 to 2.31, kurtosis values ranging from 0 to 8 and a very small sample size (N = 5). Method. A Monte Carlo simulation study was performed with data drawn from the aforementioned distributions, using covariance matrices with epsilon values approximately equal to 1 (assumption of sphericity satisfied). The number of repeated measures was also manipulated (K = 3, 4 and 6). Type I error was interpreted according to Bradley's liberal criterion, whereby a procedure is robust if the Type I error rate is between 2.5% and 7.5%. Results. The F-statistic showed robust behaviour, with Type I error rates within the required range [2.5% - 7.5%] in all but three conditions, which corresponded to distributions with the largest deviations from normality (skewness and kurtosis coefficients from 2 and 6). In these conditions the F-statistic was liberal. Conclusions. The F-statistic is generally robust to non-normality with very small sample sizes across a wide variety of distributions. However, if the deviation from normality is extreme, a larger sample size is needed. Researchers are therefore advised to plan for an adequate sample size if the data are expected to show extreme deviation from normality. This research was supported by grant PID2020-113191GB-I00 from the MCIN/AEI/10.13039/501100011033.

Title

Exploring the Social Perception of Cyber-Sexual Exploitation: A Reflexive Thematic Analysis of Reactions toward "Amouranth Case"on X

Author(s)
Rocío Vizcaíno-Cuenca [3] , Alba Sáez-Lumbreras [1] , Jesús L. Megías [2]

[1] Department of Social Psychology, Faculty of Psychology, Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain; [2] Department of Basic Psychology, Faculty of Psychology, Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain; [3] (1) Department of Methodology of Behavioural Sciences, Faculty of Psychology, Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain

Abstract

Introduction: Sexual exploitation experienced by women in online spaces represents an understudied phenomenon. Women are not only victims of this type of violence but also face judgment regarding the incidents they experience. This study analyses the social perception of cyber-sexual exploitation basis on the 'Amouranth case', a cyber-sexual exploitation incident reported by a popular streamer on social media. Specifically, we aim to gain an in-depth understanding of both positive and negative attitudes toward cyber-sexual exploitation to ultimately isolate the content areas that shape the perception of this phenomenon. To achieve this, and following previous research that has highlighted the relevance of information exchanged on social networks, this study conducts a qualitative analysis of social reactions to reports on X (formerly Twitter).
Method: First, a total of 814 posts were extracted using the rtweet data package implemented in the statistical software R. These posts were then analyzed using the reflexive thematic analysis proposed by Clarke and Braun (2018).
Results: The results of the analysis reveal both positive and negative attitudes toward cyber-sexual exploitation. The positive attitudes can be grouped into the following categories: (1) lack of credibility, (2) trolling, (3) counter-stereotypical victim, (4) victim-blaming, and (5) factors related to the dissemination context. Negative attitudes can be grouped into: (1) paternalistic motivations and (2) feminist motivations.
Conclusion: This study applies reflexive thematic analysis to content extracted from social networks, providing a deeper understanding of the social perception of cyber-sexual exploitation. The implications of these findings are discussed.

Title

Recommended open research data repositories for psychology

Author(s)
Ainize Martinez-Soto [1] , Izaskun Ibabe Erostarbe [2] , Mireia Gartzia [3]

[1] University of the Basque Country, UPV/EHU; [2] Universidad del País Vasco, UPV/EHU; [3] Zenit Solar

Abstract

Research data are factual records (numerical, textual data, images and sounds) used as primary source for scientific research. These complex digital resources require effective management and comprehensive descriptions to ensure their standardization, reusability and interoperability. This is achieved through metadata, which consists of structured information describing datasets, providing essential context, enabling retrieval, access, visibility and long-term preservation. The resources for research data are classified as open (freely accessible), mixed (open and restricted data) and complementary tools (not exclusively designed for academic research). Open research data (ORD) foster scientific collaboration and enrich research by giving it greater depth and transparency. Depositing research data in reliable data repositories creates opportunities for future use, extending beyond the initial use and purpose for which the data were originally collected. Scientific data repositories play a key role in science and should adopt systematic data management practices to guarantee the proper collection, curation, preservation, long-term availability, dissemination and accessibility of datasets. Selecting a research data repository is an important decision, both to save data from our research and to reuse data from other researches. In order to be able to reuse research data, researchers also should know the research data quality. Thus, the objective of this presentation was to identify the best ORD repositories (specialist and generalist) with the basic metadata for the description of the data on psychological sciences research. A systematic review was carried out blindly by two independent evaluators exploring two open data search engines (Google Dataset Search and Eudat B2FIND). Additionally, this research explored strategies for discovering publicly available data. Institutional open data were excluded. In order to assess research data quality, the Horizon 2020 Program's Guidelines on FAIR data management were used, referring to how research data should be treated so that they are Findable, Accessible, Interoperable and Reusable. This work offers links to available dataset repositories in the field of psychological science research. Among the most interesting dataset repositories is the list compiled by the American Psychological Association, and the re3data.org project, which is a global registry of research data repositories. Moreover, academic journals such as Scientific Data or Data in Brief publish data papers, that is, articles focused exclusively on datasets, including their description, methodology and purpose. These journals follow a peer review process that guarantees the quality of the data, offering added value to the dataset. Publishing a data paper after depositing the data facilitates the citation, recognition, visibility and monitoring of the datasets. In conclusion, data sharing may determinate the quality of research data. Establishing a robust culture of data sharing requires the adoption of best practices in data documentation and management. Assessing the capabilities and services offered by data repositories provides opportunities for enhancing their functionality and realizing the full potential of data through new applications.

Title

An R package for applying meta-analytical procedures under a mixture model approach

Author(s)
Juan Botella [1] , Manuel Suero [1] , Juan I. Durán [1]

[1] Universidad Autónoma de Madrid

Abstract

Suero et al. (in press) formulated a meta-analytical random effects model (REM) under a mixture model framework intended to resolve some inconsistencies of the classical REM model to the standardized mean difference, g. In this work we present an R package, MixtureREM, that implements the procedures developed so far under the mixture model approach. It provides unbiased point estimates of the mean and variance of the parametric effect size, confidence intervals for the mean and estimated values of the variance of g independent of the values of g itself. The package also includes tests of homogeneity of the parametric effect sizes of the studies with higher statistical power and type I error rates closer to the nominal value than the test developed under the classic model. We will show an example of use of the package, including an additional function for random generation of meta-analytic datasets of g allowing to manipulate several key factors.

Title

USE OF ARTIFICIAL INTELIGENCE IN CONTENT VALIDITY

Author(s)
Xavier G. Ordóñez [2] , Sonia Janeth Romero Martínez [1]

[1] Universidad Nacional de Educación a Distancia; [2] Complutense University of Madrid

Abstract

Artificial intelligence (AI) has significantly evolved across various fields, including psychometrics, where it can enhance the development and validation of measurement instruments. As AI becomes more relevant, literacy in this technology has become essential, driving the creation of educational programs and assessment tools. However, there are still limitations in how AI literacy is measured, prompting the search for new methodologies.

Different authors have proposed models to assess AI literacy. Ng et al. (2021) identified four dimensions: cognitive, metacognitive, affective, and social. Kong et al. (2024) also suggested an approach based on dimensions, including conceptual understanding, real-world application, self-efficacy, and social and ethical awareness. Other studies have developed scales to evaluate AI literacy, such as the AI Literacy Scale by Laupichler et al. (2023) or the MAILS scale by Carolus et al. (2023), which incorporate psychological and educational aspects.

This study proposes an innovative methodology where generative AI is not only used to create items but also to validate their content, replacing the traditional process based on human judges. A total of 720 items were generated using ChatGPT, based on Bloom's Taxonomy and specific AI literacy dimensions. These items were then evaluated by eight generative AI models to analyze their congruence with theoretical dimensions, ensuring content validity. Subsequently, a second group of AI models assessed the clarity, theoretical connection, discrimination, writing quality, and usefulness of the items.

The results showed that the AI-based methodology effectively identifies items with high levels of content validity in various AI application areas, such as social media, virtual assistants, and entertainment. However, challenges were detected in the discrimination of certain items and the content validity of some factors within Bloom's Taxonomy, particularly in the levels of remembering and applying.

Compared to previous studies, this research demonstrates that generative AI can streamline the content validation process, allowing for the evaluation of a large number of items in less time than human expert panels. The methodology of the present study ensures a rigorous statistical analysis, including the Content Validity Ratio (CVR), the Content Validity Coefficient (CVC), and Aiken's V.

Among the study's limitations is the dependency on AI model training and the variability of their responses. However, this methodology opens new possibilities for the creation and validation of assessments in education and psychology. In the future, it is recommended to compare AI-generated judgments with those of human experts to assess reliability and explore AI applications in other fields of knowledge.

The poster will include examples of interactions with AI and the results of the content validity of the proposed items.

Title

Targeting the rSTS with tDCS to Modulate Attentional Bias in Bullied University Students with Low PTG

Author(s)
Yennifer Ravelo González [1] , Hipólito Marrero [1] , Rosaura Gonzalez-Mendez [1] , Olga M. Alegre de la Rosa [2]

[1] Departamento de Psicología Cognitiva, Social y Organizacional; [2] Departamento de Didáctica e Investigación Educativa

Abstract

Abstract
Background & Objectives:
Post-traumatic growth (PTG) refers to positive psychological changes that occur as a result of struggling with adversity. Research suggests that individuals with high PTG exhibit an attentional bias towards positive resilience-related words, which may facilitate coping with trauma. However, those with low PTG may not exhibit this bias. This study explores whether transcranial direct current stimulation (tDCS) can enhance attentional bias towards resilience-related words in previously bullied university students with low PTG.
Methods:
A total of 36 university students who had experienced bullying before entering university participated in the study. Participants completed an emotional Stroop task, where they identified the color of resilience-related and neutral words. The task was administered before and after tDCS stimulation targeting the right Superior Temporal Sulcus (rSTS), an area associated with intentionality processing. Participants were randomly assigned to either an anodal stimulation or a sham (placebo) condition. The study also examined the moderating role of approach motivation in the relationship between PTG and attentional bias.
Results:
The results revealed that anodal tDCS significantly increased attentional bias towards positive resilience-related words in students with low PTG. A moderation analysis showed that this effect was dependent on approach motivation—only participants with medium or high approach motivation benefited from stimulation, demonstrating a stronger attentional bias towards positive resilience-related words after tDCS. No such effects were observed in the sham condition.
Conclusions:
These findings suggest that tDCS targeting the rSTS can enhance attentional bias towards resilience-related stimuli, potentially aiding individuals with low PTG in developing more adaptive cognitive processing patterns. However, approach motivation appears to be a crucial factor in determining the effectiveness of this intervention. Future research should explore the long-term effects of tDCS, potential applications in psychological interventions, and whether combining brain stimulation with cognitive training could further enhance PTG and resilience in trauma-exposed individuals.

Title

Beyond Classical Random Effects Meta-Analysis: Parametric Variance of the Specific Variance Estimator Under a Mixture Model Framework for Standardized Mean Difference

Author(s)
Juan Botella [1] , Alba Lirón León , Manuel Suero [1]

[1] Universidad Autónoma de Madrid

Abstract

The classical meta-analytical random-effects model (REM), when applied to the standardized mean difference, g, is usually computed by taking the conditional (to the parameter) variance of g, instead of the unconditional variance, to estimate the sampling variance in the specific studies involved. This practice introduces dependencies between the effect size (ES) estimates and their variances, which can distort estimation procedures. Additionally, the traditional handling of g's variance in REM leads to biased estimates, as the positive relationship between g and its conditional variance results in underestimation of μ_Δ.

To address these issues, Suero et al. (2025) developed a random-effects model reformulated as a mixture model (MM). This framework is a flexible alternative that encompasses ES indices in which the estimate and its estimated variance are stochastically dependent, and ES indices in which they are independent. It also yields new estimators of the variance of true effects, or specific variance τ^2, and of the mean effect μ_Δ. Nevertheless, deriving an estimator for the variance of $\hat{\tau}^2$ and characterizing its distribution remained outstanding tasks.

In this study, we derive an analytical expression for the parametric variance of the τ^2 estimator under the MM approach, which is crucial for assessing estimation precision. To achieve this, we rely on fundamental theorems from mathematical analysis, key algebraic results, and essential properties of estimators, ensuring a rigorous derivation. An extensive simulation study is conducted to assess the accuracy of the variance formula. This constitutes a step toward understanding the distribution of the estimator, which in turn will allow us to construct a confidence interval.

Title

Bayesian multilevel modeling of visual search trajectories: a simulation study and a hybrid foraging application

Author(s)
Alicia Ferrer Mendieta [1] , Javier Revuelta [2]

[1] Departamento de Psicología Básica, Universidad Autónoma de Madrid, 28049 Madrid; [2] Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, 28049 Madrid

Abstract

Hybrid foraging refers to a visual search task where observers look for multiple instances of several target types. Examining different target types provides insights into the diverse strategies employed during search, such as selecting targets consecutively (runs), alternating between target types (switches), and adjusting the length of runs to meet task demands. Traditionally, search strategies have been analyzed using the proportions of runs and switches; however, these measures are influenced by experimental conditions, including sample size and task characteristics. To address these limitations, Clarke et al. (2022) proposed a Bayesian multilevel model that conceptualizes foraging as generative sampling without replacement, offering a more robust framework for understanding foraging strategies. This model introduces two parameters to evaluate target selection biases: one capturing target preference and another accounting for the spatial arrangement of targets. The spatial parameter incorporates both the Euclidean distance between targets and the angular differences in target search trajectories.

In this study, we examined the accuracy of Bayesian estimators for the spatial location parameters of the multilevel model through a simulation-based approach. The manipulated conditions included the spatial distribution of targets and the characteristics of simulated search trajectories. Model parameters were estimated using an MCMC algorithm implemented in the Stan programming language. Our analysis focused on the RMSE between true and estimated location parameters, as well as the sensitivity of Bayesian model evaluation statistics in identifying misspecified models. Additionally, we applied the multilevel model to real-world data from 30 young adults who completed the FORAGEKID task (Gil-Gómez de Liaño & Wolfe, 2022). This application aimed to evaluate the model's performance under real conditions and to demonstrate the insights provided by Bayesian estimates.

Title

Using an alternative technique of multidimensional scaling to compare three perceptual spaces of animal abuse

Author(s)
Andrea Vera Suárez [1] , Stephany Hess-Medler [1] , Ana M. Martín [1]

[1] Universidad de La Laguna

Abstract

The increasing social and political focus on animal abuse has led to changes in legislation recognizing animals as sentient beings and to research analyzing human-animal relationship. Animal abuse is legally typified as an environmental crime within the category of crimes against the natural environment, such as those that harm flora, fauna, and protected areas. Animal stereotypes influence how they are categorized and treated by humans. The aim of this study is to analyze the similarities and differences among the perceptual spaces that people spontaneously elaborate when representing the abuse of protected animals, pets and farm animals. Participants were 366 men and women aged between 18-82 years, mostly resident in a highly environmentally protected territory. They completed an online questionnaire containing scenarios, based on press releases, of the three categories of environmental crime. Each participant was randomly asked to rate the scenarios from one of these three categories in terms of severity, justification, indignation, intentionality, punishment, and likelihood of personal intervention and calling the police. The questionnaire also included questions on socio-demographic data and a social desirability scale. The data were analyzed with multidimensional scaling using as input matrixes the average of the squared differences of the scores assigned to each pair of scenarios by all participants on each scale, instead of the traditional technique, in which each input matrix corresponds to one participant. The result showed that a three-dimension solution was the best for the three perceptual spaces. However, the content, label and order in which each dimension emerged in the shaping of each space varied. Most pet abuse scenarios were perceived as highly reprehensible and deliberate, with the abuse of dogs and cats being more unjustified and deserving of personal intervention than of other companion animals. Scenarios involving the abuse of protected and of farm animal elicited less consistent reactions, influenced by the perception of their instrumentality for humans, as food or for economic profits. In conclusion, the results suggest that animal abuse is a specific type of environmental crime and that characteristics such as that its victim are specific living beings and that the harm they suffer is observable need to be taken into account for a better understanding. The advantage of the procedure used for the multidimensional scaling was that, in addition to providing the weights of each scenario in the scaling dimensions, it also facilitated the weights of the scales in relation to these dimensions. These weights provided very useful quantitative information for the interpretation of the dimensions that would otherwise had to be based exclusively on the relative proximities of the scenarios. Future research should use alternative methodologies and techniques, in different samples and settings, to explore the key variables for effective interventions to prevent and control the social problem of animal abuse.

Title

Effectiveness of Online Psychological and Psychoeducational Interventions in Preventing Maternal Perinatal Anxiety: A Preliminary Meta-Analysis of Randomized Controlled Trials

Author(s)

Carlos Barquero-Jiménez [1] , Sergio Castellanos-Luna [1] , Patricia Moreno-Peral [2] , Cristina Garcia-Huércano [2] , Alessia Caffieri [3] , Emma Motrico , Irene Gómez-Gómez [1]

[1] Universidad Loyola Andalucía, Spain; [2] Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina (IBIMA Plataforma BIONAND), Malaga; [3] Department of Humanistic Studies, University of Naples Federico II, Naples, Italy

Abstract

Introduction: The perinatal period can be a challenging time for women, often associated with mental health issues such as anxiety. However, there is limited evidence on the effectiveness of online interventions aimed at preventing anxiety disorders during this period. This study aims to conduct a meta-analytic synthesis of randomized controlled trials (RCTs) to assess the effectiveness of online psychological and/or psychoeducational interventions in preventing maternal perinatal anxiety.
Methods: A meta-analysis of RCTs was performed by searching the most relevant databases in this field. The risk of bias in the included studies was assessed using the Cochrane Risk of Bias Tool version 2. Data extraction focused on key variables, including: the target population (sample characteristics, pregnant or postpartum women, parity), the type of prevention (universal, selective, or indicated), the intervention (timing, orientation, guidance), and the outcomes (instruments, constructs assessed, primary or secondary outcomes, time points of assessment, and effectiveness). Statistical analysis was conducted using the Comprehensive Meta-Analysis software. The standardized mean difference (SMD) was calculated using Hedges' g to obtain the pooled effect size, applying a random effects model for the calculation of the pooled SMD. Heterogeneity was assessed using the Q statistic and its p-value, along with the I^2 index and its 95% confidence interval (95% CI). Publication bias was evaluated using Begg's and Mazumdar' s rank correlation test and Duval and Tweedie's trim-and-fill procedure. Sensitivity analyses examined variations in the pooled SMD using a fixed effect model, Cohen's d as the effect size, exclusion of the most heterogeneous study, and risk of bias. Subgroup analyses were conducted using a mixed-effects model based on the aforementioned categorical variables. The Knapp and Hartung procedure was applied to calculate beta coefficients, standard errors, p-values, and the 95% CIs for the meta-regression model.
Results: Thirteen RCTs were included in the meta-analysis. The pooled SMD was small (g= -0.231, 95% CI: -0.444 to -0.017) but statistically significant (p = 0.034). The effect became non-significant when studies at high risk of bias were excluded. Ten RCTs were rated as high risk of bias. The Q index, its p-value and the I^2 revealed significant (Q_{15} = 72.643; p = 0.069) and substantial (I^2 = 79.351%; 95% CI) heterogeneity between the studies. Sensitivity analyses showed no change in the pooled effect size. No publication bias was detected based on the Begg's and Mazumdar's rank correlation test and the Duval and Tweedie's trim-and-fill procedure. Significant differences in effectiveness trends were observed according to prevention type and guided intervention type (p < 0.05). The final meta-regression model explained 31% of the variance. However, no significant associations were found between intervention effectiveness and the covariates included in the final model, such as continent, year of publication, intervention guidance and orientation, and type of prevention.
Conclusions: These preliminary findings suggest that online psychological and/or psychoeducational interventions may be effective in preventing maternal perinatal anxiety, although further research is needed to confirm these results.

Title

Measuring the dynamic structure of affect using a stepwise exploratory structural equational approach

Author(s)
Shannon Dickson [1] , Eva Ceulemans [1] , Kim De Roover [1]

[1] KU Leuven

Abstract

The proliferation of experience sampling methodology (ESM) has advanced the study of affect dynamics. In ESM, participants respond to multiple questions about the presence and intensity of positive an negative emotions at random moments throughout the day for many consecutive days or weeks. These items are commonly thought to represent two latent constructs, positive affect (PA) and negative affect(NA), and the dynamics of the summarized PA and NA scores are subsequently modelled using vector autoregressive modelling (VAR). In a VAR model, momentary PA and NA are predicted based on the preceding PA and NA scores. The computation of PA and NA implicitly relies on a so-called measurement model (MM) that specifies how observed items relate to latent constructs. However, in ESM this MM is unlikely to hold over time and the misspecification can lead to incorrect VAR parameter estimates. Therefore, the MM must be explicitly studied before fitting the VAR. The recently proposed three step latent vector autoregressive modelling (3S-LVAR) is a stepwise approach for estimating VAR models for latent constructs. In this approach, the MM is estimated, then factor scores are computed, and finally the auto- and cross-regressive relations are estimated while accounting for the uncertainty in the factor scores. Up to now, an important limitation of 3S-LVAR is that prior specification of the MM is required, which is notoriously difficult with ESM, as the commonly assumed PA-NA distinction is not systematically found. To address this, we extend 3S-LVAR by incorporating exploratory factor analysis to infer the MM from the data in the first step. One challenge of this exploratory approach is the choice of rotation, as different rotation criteria can lead to differences in the MM and in the dynamics of the latent variables, potentially altering conclusions regarding the auto- and cross-regressive parameters. I will present on the impact of different rotation criteria on the interpretation of the auto- and cross regressive relations for existing ESM data.

1.21 Session 13 : ”Longitudinal models and Individual variability”

Title

Longitudinal diagnostic models for small sample size contexts

Author(s)
Hyunjee Oh [1] , Chia-Yi Chiu [2] , Pablo Nájera [3] , Miguel A. Sorrel [4]

[1] Columbia University; [2] Teachers College, Columbia University; [3] Departamento de Psicología, Universidad Pontificia Comillas; [4] Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid

Abstract
Cognitive diagnostic models (CDMs) constitute a family of confirmatory, restricted latent class models in which individuals are classified into different profiles based on their mastery or non-mastery of the measured attributes (e.g., skills, competences, psychological processes). Beyond their use in clinical or organizational psychology, CDMs have been widely applied in educational settings, where they directly impact students'learning by providing diagnostic feedback that guides remedial instruction based on their strengths and weaknesses. However, two main challenges hinder the applicability of CDMs for these purposes. First, most CDMs require large sample sizes (N > 500) for accurate estimation. Second, learning (or improvement, in a psychological context) can only be assessed through longitudinal designs. The first challenge was addressed by the development of the R-DINA model (Nájera et al., 2023), a parametric CDM specifically designed for small-scale assessments. The second challenge was tackled with the TDCM (Madison & Bradshaw, 2018), a general longitudinal diagnostic model. The present study combines these two advancements by proposing the integration of the R-DINA model into the TDCM to enable longitudinal diagnostic assessments in classroom-level settings. The performance of the longitudinal R-DINA model is tested and compared to traditional CDMs through an exhaustive simulation study focused on small sample size conditions (25 < N < 200). Different test lengths, numbers of attributes, item discrimination levels, and attribute correlations are explored. Results demonstrate the viability of the longitudinal R-DINA model for small-scale longitudinal assessments to track students'learning or patients'improvement. Implications and practical recommendations will be discussed.

Title

A Proposal to Test Approximate Measurement Invariance of Multi-Item Self-Reports Across Intense Longitudinal Assessments

Author(s)
Oliver Schilling [1]

[1] Universität Heidelberg

Abstract

Intensive longitudinal studies - commonly referred to as experience sampling methods, ecological momentary assessments, ambulatory assessments, or daily diary studies - have become a prominent domain of psychological research focused on short-term within-person changes of mental attributes, which are typically measured by composite scores of the participants'self-reported endorsements of several questionnaire items. However, frequent self-reports within a short observation period might impact alertness towards and perception of one's respective experiences, hence changing item discrimination between levels of the latent construct and/or item severity (difficulty). Thus, there is a need to examine measurement invariance (MI) across the frequently repeated ambulatory assessments. To do so, recent research proposed to model approximate MI of intensive longitudinal measures by means of multilevel cross-classified (assessments nested within individuals and within occasions) structural equation models, with random effects of item intercepts and loadings and the respective between-occasion random variances signalling differential item functioning across repeated assessments (e.g., McNeish et al., 2021). The current study presents and adds to this approach, proposing a rationale to determine tolerable amounts of non-invariance, which could serve to test for approximate MI in the absence of significance testing options for the respective random variances (as the procedure needs Bayes estimation). Results from a simulation study will be presented to demonstrate the application and the feasibility of the approximate MI strategy proposed.

Title

The Invariance Partial Pruning Approach to The Network Comparison in Longitudinal Data

Author(s)

Xinkai Du [1] , Sacha Epskamp [2] , Sverre Urnes Johnson [1]

[1] University of Oslo; [2] National University of Singapore

Abstract

Network models from time-series and panel data have been powerful tools to investigate the dynamical relations among variables. A common goal of empirical research is to compare the networks of different groups, such as treatment and control, to understand how inter-variable relations are shaped by the grouping variable. However, existing methods to compare idiographic networks are merely global tests that cannot tell specific location of edge difference and equality. Furthermore, there is a lack of easily applicable methods to compare networks from panel data where just a few time-points are available per person. We therefore present the invariance partial pruning (IVPP) approach, which first evaluate the presence of heterogeneity with the network invariance test, and then determine the exact locus of edge equality and difference with partial pruning. Through simulation studies, we discovered that network invariance test based on AIC and BIC performed well, but LRT was prone to false discovery. Comparison with the fully constrained model revealed superior performance than comparison with the fully unconstrained model. Partial pruning successfully uncovered specific edge difference with high sensitivity and specificity. We conclude that IVPP is an essential supplement to the existing network methodology by allowing the comparison of networks from time-series and panel data, and also allowing the test of specific edge difference. The method permits the network comparison of both different groups/persons, or different time periods of the same group/person. We implement the algorithm in the R-package IVPP.

Title

Commensurable indicators - finding potentially metric invariant indicators

Author(s)
Eric Klopp [1] , Stefan Klößner [1]

[1] Saarland University

Abstract

In multi-group confirmatory factor analysis, it is important to test for metric measurement invariance (MMI), and searching for invariant indicators may be demanding. We present a method to find metric invariant indicators using some mathematical properties of invariant indicators.

We start with introducing an example provided by Yoon and Millsap (2007, Structural Equation Modeling, 14, 435-463) and introduce the notion of proportional factor loadings and the concept of change of scale (Klopp and Klößner, Methodology, 19, 192-227). From this, we derive and formally define the concept of commensurable indicators. The loadings of commensurable indicators are multiples of each other and conform to an equivalence relation. We show that when two indicators are commensurable, either both or none fulfill metric invariance. Each equivalence class with respect to commensurability forms a commensurable indicator subset (CIS), and correspondingly, the set of CIS will be called a partition of commensurable indicator subsets (PCIS).

Using the notion of a loading profile, i.e., a vector containing the loadings of a factor's indicator over the groups, it is possible to define an average loading profile and a static indicator loading for a given CIS. We derive the proposition that the loading profile can be multiplicatively decomposed into a static component and a second component, which within a CIS does not depend on the particular indicator but only on the corresponding CIS's average loadings profile. Metric invariance of the indicators belonging to a CIS can be read off the average loadings profile: If and only if this loadings profile is constant, i.e., if and only if it is a non-zero multiple of the vector of ones, then metric invariance of all indicators belonging to the CIS is fulfilled. Thus, the average loadings profile of CIS reveals whether its indicators fulfill MMI.
This mathematical framework provides the possibility of inferring the CIS structure from data using a second-order factor model with certain constraints to implement the decomposition of the loadings profiles from our proposition and find the partition with invariant indicators by systematically considering all partitions of the indicator set. The framework, therefore, enables finding sets with possible metric invariant indicators. The optimal partition can be inferred from estimating these models and using information criteria like the AIC, BIC, or sBIC.

Using the initial example, we conducted a Monte Carlo simulation with 10.000 repetitions with sample sizes of N=150, N=300, and N=500. For N=150, most of the time, the AIC and sBIC detected the correct partition, whereas the BIC failed to find the correct partition. However, for N=300 and N=500, the method selected the correct partition in most cases. In particular, the AIC and sBIC performed the best.

Although considering all partitions is a brute-force method, the simulation demonstrated that the method derived from the mathematical properties of the CIS structure has the potential for applied MMI analysis. Lastly, we want to mention that the same considerations and the method to find the partitions also apply in longitudinal invariance settings.

1.22 Session 17 : ”Sampling and Responses in experience sampling studies”

Title

Prediction Intervals for the Target Sample Size during Information-Based Monitoring

Author(s)
Tom Loeys [1] , Ole Schacht [1] , Beatrijs Moerkerke [1]

[1] Department of Data-Analysis, Ghent University

Abstract
A key feature of reproducible research in psychology is having a sufficiently large set of observations to support that reliable conclusions can be drawn from the data, which includes having enough statistical power. The current practice for sample size calculation in psychology typically entails specifying the a priori power probability with which a presumed effect is to be detected at a prespecified significance level. This also requires information on the nuisance parameter(s) of the test statistic at hand (such as the variance). Unfortunately, due to often limited information on these nuisance parameters at the study design stage, the sample size may easily be misjudged.

The last two decades, more flexible frequentist alternatives have been developed that enable reestimation of the target sample size based on interim inspections of the variability (i.e., during ongoing data collection) without inflating the type I error. More precisely, nuisance parameters are monitored as new data comes in, and data collection is terminated only when the effect of interest can be sufficiently precisely estimated. A standard statistical test is then performed with the desired power at the prespecified significance level. Moreover, effect estimators remain unbiased, and no loss in efficiency is observed. Clearly, using such flexible methods may be attractive to applied researchers as accurate information on the variability in the data is no longer needed when planning a study or experiment.

While such approaches offer an adjustment for incorrect assessment of necessary study resources, it is also often criticized because researchers do not know the final sample size at the start of their study. It would therefore be desirable to have an estimator for the target sample size along with a measure of its uncertainty that can updated while data is accumulating. The target sample size that is estimated during the course of the study is a random variable that has not been studied yet. This unknown dispersion of the target sample size is often regarded as the main impediment of information-based monitoring.

In this talk we present a new, intuitively easy, and general approach to make interval predictions for the target sample size of a study. These prediction intervals are based on distributional properties of the Fisher information and can be made at arbitrary points throughout the sampling process. We demonstrate the approach in a typical setting where interest lies in the effect of a focal predictor in a regression model while adjusting for other covariates. We provide a user-friendly Shiny app to facilitate the usability of the prediction intervals.

Title

Tracking micro-minorities: methodological solutions to the challenge of small numbers

Author(s)

Camille Kandiko Howson [1]

[1] Imperial College London

Abstract

This paper aims to raise awareness, problematise and look for solutions to the challenge of small numbers in quantitative analysis of inequalities in education. Gathering data and evaluating progress are key aspects of regulatory reporting in higher education. However, when working with some hyper-marginalised groups of people (or micro-marginalised people), we run into the problem of small numbers—which can present methodological, legal, ethical and practical concerns. The biggest danger is that small numbers of individuals means that a group, a specific characteristic or a set of intersectional factors get ignored as it was not clear how to account for them in the data.

Examples include exploring progression and completion rates for a university department which had 1 transgender student, and a 10-year analysis of the progression of another department's undergraduate students which identified only 5 students of a targeted underrepresented group. This paper draws on frameworks of quantitative criticalism and critical quantitative inquiries and explores methodological solutions from a range of fields, including psychology, epidemiology, public health, neuroscience and medical research. Research drawing on critical quantitative approaches, or CritQuant, focuses on the use of numerical data to uncover power inequities, the formulation of rigorous models to better represent minoritised people and social groups and prioritising social justice-oriented research. Critical race quantitative intersectionality, or QuantCrit, more specifically draws on critical race theory. However, neither approach provides specific methodological solutions to the problem of small numbers in an existing dataset.

How researchers use statistical analyses shapes their research toward or away from social justice agendas. Points include how proactively using social category data can uncover and address discriminatory practices, and how analytic approaches may not fully support equity efforts. While research over the past five years has raised awareness of the need for criticality in quantitative research, specifically that of race and ethnicity, less has been done on the specific methodological approaches for doing so or for broadening efforts more widely across socio-demographic characteristics. Integrations of intersectionality theory and moderated general linear modelling (MGLM) offer some possibilities for methodological approaches using a social justice perspective.

The paper presents on principles to guide more equitable research practices including balancing privacy and transparency. It also offers methodological solutions and considerations for different analytical approaches, including: small sample sizes, comparing across groups, intersectional analysis, interpreting binary outcome variables and eliminating outliers.

Title

The distracted participant? Experience sampling response behavior and participant burden in social settings

Author(s)
Inez Myin-Germeys [1] , Gudrun Eisele , Robin Achterhof [2]

[1] KU Leuven; [2] Erasmus Universiteit Rotterdam

Abstract

Background: Theoretical accounts of careless responding name environmental distractions as a key contributing factor. In experience sampling method (ESM) studies, participants receive questionnaires across a range of potentially distracting situations. Previous research suggests that participants find ESM questionnaires particularly disturbing in social situations, especially when engaging in social interactions. Social situations also represent highly distracting environments. Yet, the effects of these environmental distractions on response behavior remain poorly understood.
Methods: We investigated the effects of responding to questionnaires in distracting environments by comparing disturbance and response behavior across various social and non-social situations. Data from three young adult samples (combined N = 293) and a general population youth sample (N = 1903) was analyzed with multilevel (logistic) regressions.
Results: In line with previous research, adults were significantly more disturbed by assessments when in company compared to when alone, especially when also interacting with their company. In addition, we found small but significant differences in response behavior between social settings in adults, with changes pointing towards lower data quality when participants are in company. Interestingly, patterns were different, in some cases even reversed, in school-going adolescents.
Conclusions: While our findings suggest that the distraction of social settings affects participant burden and response behavior, the influence on data quality was minor. Differences across samples suggest that the setting of the social experience (in vs outside school) needs to be considered. Preparing participants for sampling in distracting (social) environments may help safeguard data quality and reduce participant burden.

Title

An extensive evaluation of the stochastic countdown

Author(s)
Niels Smits [1]

[1] University of Amsterdam

Abstract

The countdown method (CM, Ben-Porath et al., 1989) and its stochastic extension (SC, Finkelman et al., 2012), have been shown to be very cheap but valuable additions to the field of variable-length classification testing, which is dominated by methods based on psychometric models (Thompson, 2007). In this simulation study CM and SC are extensively studied under a series of varying factors such as calibration sample size, number of items, distributional shape of sum scores, location of decision threshold, inter-item correlations and sum score reliability. Preliminary rules for minimal requirements are suggested.

References

Ben-Porath, Y. S., Slutske, W. S., & Butcher, J. N. (1989). A real-data simulation of computerized adaptive administration of the MMPI. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1(1), 18.

Finkelman, M. D., Smits, N., Kim, W., & Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. Applied Psychological Measurement, 36, 632–658.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. Practical Assessment, Research, and Evaluation, 12(1), 1–12.

Title

Defining ratio effects in randomized controlled trials using a stochastic theory of causal effects

Author(s)
Axel Mayer , Christoph Kiefer [1]

[1] Bielefeld University

Abstract

In cases in which the outcome variable is binary (e.g., success/no success) or a count variable (e.g, number of depressive symptoms), the effect of a treatment or intervention is often expressed as ratio (e.g., risk ratio, odds ratio). While it is relatively straightforward to estimate some kind of ratio effect based on a logistic regression or Poisson regression, it is a non-trivial question whether ratio effect measures should be considered and if yes, how they can be interpreted and which assumptions need to be fulfilled in order for them to have a causal interpretation. For example, it is somewhat counter-intuitive in the context of ratio effects that an effect measure based on group averages does not necessarily resemble an average over individual effect measures, not even in randomized controlled trials. This phenomenon is known as (non-)collapsibility and has received quite a lot of attention in the biostatistics and epidemiology literature. In this talk, we discuss the usefulness of a stochastic theory of causal effects for defining different types of ratio effects and for clarifying the necessary assumptions for their identification. We briefly introduce the core aspects of the stochastic theory of causal effects before showing how to define ratio effects either as individual ratio effects or as average ratio effects. The different types of effects require different causality assumptions and have a different meaning, which only becomes clear when building on theories of causal effects. In addition, we present new features in the R package EffectLiteR that allow to estimate the various types of causally defined ratio effects based on the generalized linear model. The approach is illustrated by a simulated example from a psychological randomized controlled trial.

1.23 Session 21 : ”Psychometric Innovations and Diagnostic Methodologies”

Title

Hospital Pedagogy and health in children and families. Contributions from observational methodology and mixed methods.

Author(s)
Manuel J. Rodríguez Allué [1] , Sarah Muñoz-Violant [2] , Verónica Violant Holz [3] , Mariona Portell [4]

[1] Department Biomedical Sciences, Institute of Neurosciences, School of Medicine and Health Sciences, Universitat de Barcelona, Barcelona, Spain; [2] The University of British-Columbia, Canada; Research Group and Innovation in Designs (GRID). Technology, multimedia, and digital application to observational designs; [3] Universitat de Barcelona, Spain; Research Group and Innovation in Designs (GRID). Technology, multimedia, and digital application to observational designs; [4] Universitat Autònoma de Barcelona, Spain; Research Group and Innovation in Designs (GRID). Technology, multimedia, and digital application to observational designs

Abstract

Developing intervention research in the field of hospital pedagogy and health requires defining the concept in a way that helps to make the involved variables visible. Violant defined in 2017 hospital pedagogy as the integral action that assures ethical and bioethical principles and the right and duties of a person with the aim of improving the individual, the family, and the social well-being during the person's lifetime. The integral action is the key even before one' s life with illness and convalescence. Interventions designed within the framework of hospital pedagogy and health often align with the concept of complex intervention (involving multiple dimensions, actors, and levels of action). To evaluate the implementation process of this type of intervention, it is essential to use appropriate methodologies that capture its multifaceted nature. From this perspective, the design of the intervention and its evaluation are conceived as interrelated processes that evolve dynamically based on observed realities and specific contexts. The aim of this presentation is to show low intervention evaluation designs to obtain data and the possibilities with indirect observational methodology and mixed methods analysis from the quality of life (QoL) and coping strategies from complex medical conditions (CMC) pediatric sample in Spain (n=11, 3 to 17 years) and their caregivers (n=24). We conducted descriptive analyses of the perception of QoL and well-being (using validated KINDLR questionnaire) and of the participants'coping strategies (using open-ended question following the three-level hierarchical structure model of the Coping Strategies Inventory based on Folkman and Lazarus' model) and performed comparisons between the cohorts and transformed qualitative data from coping strategies into quantitative data. The results show that children aged between 3 and 6 years and their caregivers scored physical well-being the lowest out of all dimensions of well-being, and they scored family well-being the highest. Moreover, youth between the ages of 7 and 17 years and their caregivers scored school-related well-being the lowest. These results support the conceptualisation of mixed methods analysis as an appropriate approach to the inherent complexity of data obtained through indirect observation, and due to the undeniable need in the development of effective, holistic, and relevant strategies in intervention programs applied in the hospital pedagogy and health field.

Title

Cross-Validating Thematic Networks: An Explanatory Sequential Mixed Methods Study with Chronically Ill Students

Author(s)
Carmen Jerez Molina [1] , Francisca Jiliberto [2] , Verónica Violant Holz [3]

[1] Campus Docent Sant Joan de Déu. Universitat de Vic-Universitat Central de Catalunya, Barcelona, Spain; [2] Doctoral Programme Education and Society, Universitat de Barcelona, Spain; [3] Universitat de Barcelona

Abstract

Mixed methods research entails integrated analysis combining data and techniques, enhancing validity and depth by leveraging strengths of each method.
To understand the experiences and meaning that adolescents and young adults with chronic illnesses assign to their student life, a multicentre study using a sequential explanatory mixed methods design (quantitative ⟶ qualitative) and a phenomenological approach was conducted with a purposive sample of 32 chronically ill students aged 14–24. The WHOQOL-Bref questionnaire was administered, followed by ten focus groups. The study involved multi-method (selective and indirect observational methodology), and analysis integrated both data sets.
Five recurrent themes emerged from the conversation analysis and guided the construction of corresponding networks. These networks were corroborated through statistical analysis, which revealed significant relationships within several network connections.
Data should be explored using diverse techniques to cross-validate findings and interpretation. Methodological integration—particularly across qualitative and quantitative approaches in network analysis—serves as a quality criterion.

Title

Automated Scoring of Open-Ended Responses: Evaluating LLMs and Prompting Strategies

Author(s)
Daniil Talov [1]

[1] HSE University

Abstract

Before the rapid development of artificial intelligence, standardized tests mainly relied on multiple-choice questions because evaluating open-ended tasks required significant resources. Modern large language models (LLMs), such as ChatGPT, Gemini, and Llama, now enable automated assessment of open-ended tasks. Unlike traditional machine learning or deep learning methods, foundational LLMs do not require labeled datasets or extensive expertise in data science and programming. Users only need to create a well-structured prompt and verify alignment with human raters on a small sample.

This study aims to assess the feasibility of using LLMs to score open-ended tasks through prompting. Additionally, it explores effective prompting strategies. Several LLMs were compared, with ChatGPT-4o serving as the baseline model. It was evaluated against ChatGPT-o3-mini, which features advanced latent reasoning. YandexGPT4 was also included, as it has been specifically trained in Russian, the language used by respondents in this study. To explore alternatives to cloud-based solutions, the DeepSeek r1 8B and Llama 3.1 8B models were tested, as they can run locally on a computer, reducing evaluation costs and ensuring confidentiality.

To evaluate the effectiveness of different prompting strategies, both analytical and holistic rubrics were used. Techniques included zero-shot (no examples), one-shot (one example), and few-shot (multiple examples). Additional strategies included chain-of-thought prompting (reasoning before making a decision), tree-of-thought reasoning (deliberation among “multiple evaluators” before deciding), and enhancing LLM “motivation” using specific phrases.

The models were tested on three types of tasks: (1) short-answer reading comprehension tasks for children aged 10-11 (1-2 sentences); (2) a prompt engineering task for undergraduate students (up to 150 words); and (3) an economics essay task for economics students (up to 144 words). LLM performance was measured using Weighted Quadratic Cohen’s Kappa (WQK) and Mean Absolute Error (MAE). Scores were compared with human raters, who evaluated 50 randomly selected responses for each task type.

The presentation will compare five LLMs across three types of open-ended tasks. Results show that LLMs can achieve sufficient accuracy for use in low-stakes assessment scenarios. Additionally, findings on effective prompting strategies will be shared, identifying best practices. Specifically, analytical rubrics combined with multiple examples provided the most accurate assessments. Interestingly, phrases like “focus and take a deep breath before answering” or “I will pay you for good work” resulted in accuracy comparable to example-based prompting.

These results highlight the potential of LLMs for scoring open-ended responses. At this stage, they can be used in low-stakes assessments or as an assistant to human raters. Furthermore, LLMs can support data annotation and the creation of training datasets for other machine learning models.

Title

Rasch-Based Unidimensional Integration of the Most Widely Used Scales for Assessing Belief in a Just World

Author(s)
Javier Mayoral López [1] , Marta Godoy Giménez [1] , María Casado Sánchez [1] , José Fornieles Alonso [1] , Pablo Sayans Jiménez [1] , Andrés Soler Martínez [1]

[1] University of Almería

Abstract

The Belief in a Just World has traditionally been approached from a multidimensional perspective, distinguishing between the recipient of justice and the type of justice involved. However, recent research suggests exploring a unidimensional conceptualization. After confirming a bifactor structure capable of integrating the three most widely recognized scales in the literature (Dalbert, 1999; Lipkus et al., 1996; Lucas et al., 2011), Rasch models were applied to examine the psychometric properties of the items along a single dimension. This study was conducted with a sample of 515 individuals (59.8% women). Unidimensionality was confirmed through a principal component analysis of the residuals, complemented by a simulation based on the same parameters for items and respondents, with replacement, to ensure the stability of the results. The psychometric properties were adequate. All items showed good fit indices. The item separation index was [10.07, 10.69] and the person separation index was [4.40, 4.99] ([4.31, 4.80] including extreme scores). The item characteristic curves were also adequate. Moreover, due to high redundancy in severity and content, factor loadings from a bifactor model were used to propose a parsimonious evaluation of global BJW without compromising content representativeness. The psychometric properties of the short version and person scores were compared with those of the initial set of items to ensure that the reduction of redundant items did not affect either the psychometric properties or the estimation of the scores.

Title

Using Cognitive Diagnostic Models for Criterion-Referenced Standard Setting in Legal Literacy Assessment

Author(s)

Daria Gracheva [1] , Sergei Tarasov [1]

[1] HSE University

Abstract

Cognitive diagnostic models (CDMs) serve as an effective approach to diagnostic assessment in education. This study explores how CDMs can be applied to criterion-referenced standard setting. The relevance of the study lies in the increasing demand for more detailed assessment in education. Modern education systems require not only the sorting of students on the basis of their performance, but also a detailed analysis of their specific knowledge gaps in relation to their level of proficiency.

The study employed the Legal Literacy Test, developed for 9th-grade students. Based on their test results, students are assigned to one of three proficiency levels of legal literacy: developing, basic, or advanced. According to the theoretical framework, the test encompasses three broad subject areas, which are further divided into seven specific topics. The test consists of 30 items, including 11 dichotomous and 19 polytomous items (from 2 to 7 items per topic). The study sample included 767 9th grade students.

As in the Bookmark method of standard setting, experts review the descriptors of achievement levels and determine which items students at different levels are likely to answer correctly. For some dichotomous items, the experts assume that students at basic level or above would be able to answer correctly and receive 1 point; for other dichotomous items, advanced students would be more likely to answer correctly, or a developing level would be sufficient. Polytomous items provide more detailed diagnostic information. For some items, experts suggest that advanced students are more likely to score 2 points and basic students are more likely to score 1 point. In other cases, developing students might get 1 point while basic or advanced students might get 2 points.

The Q-matrix for CDM analysis combines information about specific topics for each item and achievement level for each item category according to expert opinion. We assume that the probability of answering the item correctly is a combination of specific topic and achievement level. Two different types of CDM models were tested. The sequential attribute CDM model (Ma, W., & de la Torre, J., 2016) included dichotomous attributes for levels, while the polytomous CDM model (de la Torre, J., Qiu, X.-L., & Santos, K. C., 2021) used a polytomous attribute for levels. Different cognitive diagnostic models (DINA, RRUM, LLM, ACDM, GDINA) were calibrated and compared using the GDINA package in R (Ma, W., & de la Torre, J., 2020).

CDMs allow the integration of information about item content and proficiency levels, providing a more nuanced analysis of student performance. The applied approach to criterion-referenced standard setting using CDM models provides precise insights into which specific components of literacy students have mastered and which require further development, and also allows students to be placed into proficiency levels for the overall test.

Title

Validation of an Italian work-adapted Technostress Scale for older employees

Author(s)
Anna Comotti [2] , Cristina Di Tecco [1] , Matteo Bonzini [2] , Alice Fattori [3] , Teresa Barnini [4]

[1] INAIL, Rome, Italy; [2] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan (Italy); [3] University of Milan, Italy; [4] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

Abstract

Introduction
In today's rapidly evolving technological landscape, the integration of Information and Communication Technology (ICT) in the workplace has transformed work processes, offering numerous benefits while also introducing new challenges. One such challenge is technostress, a phenomenon describing the strain caused by the pervasive use of ICT. Although technostress affects workers across all age groups, research suggests that older employees may experience it more intensely due to digital literacy gaps and difficulties adapting to constantly changing technologies. This study aims to evaluate the psychometric properties of a workplace-adapted version of Nimrod's Technostress Scale, originally developed for older adults.
Methods
A sample of 470 Italian full-time workers aged 50 and above from three different sectors - finance, packaging, and steel- participated in the study. The questionnaire was administered during occupational health surveillance, ensuring a high response rate. The original scale was translated into Italian and tailored to assess work-related stress. The adapted scale measured five dimensions of technostress: overload, invasion, complexity, privacy, and inclusion.
Results
Confirmatory Factor Analysis (CFA) supported the five-dimensional structure, with a bifactor model providing the best fit. The scale demonstrated good reliability (Cronbach's alpha = 0.75; McDonald's omega = 0.76) and validity, with technostress significantly correlating with increased perceived stress ($r = 0.32$, $p < 0.001$), decreased well-being ($r = -0.17$, $p < 0.001$), and reduced workability ($r = -0.24$, $p < 0.001$). Significant differences emerged across occupational groups and gender. Blue-collar workers reported higher overall technostress, particularly in the dimensions of overload, complexity, and inclusion, while white-collar employees experienced more invasion and privacy-related concerns. Women reported higher technostress scores than men, especially in invasion and inclusion.
Conclusion
This study validated a workplace-specific technostress scale for older workers, offering a reliable tool for assessing the impact of ICT-related stress and guiding organizational policies aimed at fostering a healthier work environment. Given the aging workforce and the increasing reliance on digital tools, addressing technostress is crucial for promoting employee well-being, productivity, and job sustainability.

Title

Assessing Professionalism in Spanish Healthcare Contexts: Cultural Adaptation and Validation of the Profes- sionalism Mini-Evaluation Exercise (P-MEX)

Author(s)
Raul Castañeda-Vozmediano [3] , Niels Smits [1] , Emilio Cervera-Barba [2] , Santiago Álvarez-Montero [3] , Miguel A.Sorrel [4] , Diana Monge [3] , Cristina Cisterna [3] , Valle Coronado-Vázquez [3]

[1] Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands; [2] Faculty of Medicine, Universidad Francisco de Vitoria; [3] Faculty of Medicine, Universidad Francisco de Vitoria, Madrid, Spain; [4] Faculty of Psychology, Universidad Autónoma de Madrid, Madrid, Spain

Abstract

Medical professionalism is defined as the commitment of physicians to the health of patients and society, the profession, and themselves. Measuring medical professionalism is crucial as it directly impacts the quality of patient care. The Professionalism Mini-Evaluation Exercise (P-MEX), consisting of 24 items across 4 domains, measures the professional behavior of medical professionals or students reported by observers. Although it has been adapted in several countries, a validated version for the Spanish population does not yet exist. Objective: The aim of this study is to translate, culturally adapt, and validate the original version into Spanish for residents evaluated by their medical supervisors. Method: After direct and reverse translation, each of the obtained versions was discussed by a committee of experts. The feasibility, comprehension, and appropriateness of the questionnaire were tested with a pilot sample, followed by validation with a larger sample. Reliability analyses (internal consistency, test-retest), dimensionality analyses (parallel analysis, exploratory graph analysis), and other sources of validity (convergent, criterion, internal structure) were explored. Results: The four-factor model proposed in the literature was replicated, obtaining similar values of internal consistency for each subdomain. 9.77% of the 276 participants that participated in this study completed the questionnaire a second time, and adequate temporal stability was found. The total score was negatively related to the informal complaints received by patients regarding the residents. Conclusions: The Spanish version of the P-MEX has been found to have reliability and validity indices similar to those reported in the literature, despite differences between national health systems and the clinical context of each country. This supports its use in Spanish samples, allowing for further exploration of this construct with significant implications for society.

1.24 Symposium : ”Research methodologies in social cognition: A measurement approach from the Social Neurosciences.”

Title

A systematic review of interventions designed to reduce alterations in social cognition

Author(s)
Cristina Martell Siqueiros , Ma de la Cruz Bernarda TELLEZ ALANIS [1] ,
Sandra Meza Cavazos , Gabriela Ramírez Alvarado

[1] CITPSI UAEM

Abstract

Introduction. Social cognition allows understanding and predicting both one's own actions and those of others. It includes processes such as the perception of emotions, the theory of the mind, empathy and social judgment. The alterations in these processes have been broadly studied in people with Schizophrenia (Ef) and Autistic Spectrum Disorders (ASD). Currently, both diagnostic evaluation and interventions have been expanded to include a variety of conditions. These treatments are a fundamental basis in neuropsychological rehabilitation programs given their crucial impact on the quality of life and social integration of affected people. Objective. Through a systematic review, to detect therapeutic interventions that diminish alterations in social cognition processes. Method. The PRISMA procedure was implemented, and the PICO question was defined: Patients, any type with alterations in social cognition; Interventions, any type of treatment; Comparison with control groups or repeated measures; Outcomes, improvements in social cognition tasks or in daily life. The search was carried out on August 6, 2024, in PUBMED. Original articles, no revisions, no metanalysis, no gray literature, in English and Spanish were requested. The terms for the research were social cognition AND treatment, rehabilitation, intervention, stimulation, therapy. The search was only in the title and included papers published from 2000 to August 2024. 94 elements were detected, 6 repeated. Therefore, 88 summaries were reviewed, of which 50 were excluded for being systematic reviews, states of the art, comments, opinions, editorials, conceptual analyses, correlational or predictive diagnostic studies, chapters, conference discussions, protocols or case studies. 38 were accepted, which were controlled studies with and without random assignment and pilot studies. Results. It was found several groups of treated patients, being the studies with Ef (16), other types of psychosis (7) and ASD (6) more abundant. The rest of participants were patients with brain damage (4) multiple sclerosis (1), mild cognitive impairment (1), children with cerebral palsy (1), children with neuromuscular diseases (1) and young offenders (1). The types of interventions were functional (magnetic and direct current), pharmacological (olanzapine, risperidone, haloperidol, clozapine), virtual and in-person cognitive and emotional stimulation programs (free and commercial), group therapies, theater, yoga and community-based psychosocial interventions, in some cases combined. Most studies (27) found improvements in some of the processes of social cognition, mainly in the recognition of emotions, and a little less in theory of the mind and empathy. Eight studies did not show improvements, 5 were with functional techniques, and 3 with training sessions (2 cognitive and 1 theater) applied to people with psychosis (6) ASD (1) and multiple sclerosis (1). Conclusions. Most interventions presented improvements in social cognition processes -although the gains were partial- being pharmacological and stimulation of social cognitive processes interventions more effective. However, research must be continued to guarantee improvement in various processes of social cognition and its transfer to everyday life activities.

Title

Exploring the Relationship between Fear of Public Speaking, Social Cognition, and Communication Skills in University Students

Author(s)

África Borges del Rosal [1] , Ernesto Pereda de Pablo , Ricardo Quintero Rodríguez

[1] Universidad de La Laguna

Abstract

Introduction. Fear of Public Speaking or Public Speaking Anxiety is a specific manifestation of Social Anxiety Disorder that can significantly interfere with personal, academic and professional performance. On the other hand, Social Cognition, which includes skills such as emotion recognition, theory of mind, empathy and attributional styles, is fundamental for interpreting the intentions and emotional states of others, facilitating effective communication and adjusting behaviours in different social contexts. In addition, communication skills, which encompass the clear and persuasive expression of ideas and the ability to adapt to social cues, are essential for successful interactions. Objectives. The present study aims to analyse the relationship between Fear of Public Speaking, Social Cognition and Communication Skills. Methodology. A sample of 436 university students (26% male, 74% female; mean age = 21.1 ± 3.46 years) was obtained through convenience sampling. The tests used were: Social Anxiety Questionnaire for Adults (SAQ-A30), Reading the Mind in the Eyes Test (RMET), Interpersonal Reactivity Index (IRI), Penn Emotion Recognition Task (ER-40), Attribution Style Questionnaire (ASQ), Communication Skills Questionnaire (HABICOM). Multiple regression analyses were performed with SPSS v.29 to assess how different measures of Social Cognition (RMET, IRI, ER-40, ASQ) and Communication Skills (HABICOM) predict the Fear of Public Speaking (SAQ-A30; 'Public Speaking'factor). Results. Findings revealed significant relationships between predictor variables and Public Speaking Anxiety. First, a positive association was identified between empathy levels and anxiety, whereas a negative attributional style (particularly internality and globality in negative situations) was linked to a higher propensity to experience this fear. On the other hand, Communication Skills are presented as a relevant protective factor, given that their presence is related to lower levels of Public Speaking Anxiety. Finally, gender is recognised as a significant factor influencing levels of Fear of Public Speaking and some of the processes of Social Cognition. Discussion. The results suggest that people with greater empathy may be more vulnerable to social evaluations, while a negative attributional style may intensify the perception of threat. On the other hand, the development of communication skills may decrease barriers to expression, thereby reducing anxiety. These findings underline the importance of working on empathy, strengthening communication skills and modifying attributional style as a strategy to address Fear of Public Speaking in personal, academic and professional contexts.

Title

Analysis of the factor structure of the Yoni Task instrument for its cross-cultural validation in the Spanish- speaking population

Author(s)
África Borges del Rosal [1] , Elena Rodríguez Naveiras [1] , Leire Aperribai Unamuno [2]

[1] Universidad de La Laguna; [2] University of the Basque Country UPV/EHU

Abstract

Introduction: The assessment of social cognition through the Theory of Mind can contribute to the study of this construct. One of the instruments proposed to assess the Theory of Mind is the Yoni Task, with which a cross-cultural validation is being carried out for the Spanish-speaking population (Argentina, Mexico and Spain). Objective: This study aims to determine the factor structure of the Yoni Task instrument. Method: A first version of the instrument consisting of 97 items has been applied to a sample of 596 participants between 18 and 65 years old of the three countries, which has shown a good reliability index. In a first exploratory factor analysis, with all the items, the instrument has shown an inadequate factorization, with a solution difficult to interpret due to the high number of factors obtained, when based on the theory only two are expected. In the second step, a procedure for the assignment of items was followed to allow the selection of those items with the best psychometric properties and the factorization of the second version of the instrument was analyzed by means of a robust factor analysis using the Factor program. Results: a short version of the instrument was created, consisting of 17 items. The results of the factor analysis performed with the short version show 2 factors (Affective and Cognitive) to which the items fit adequately and have obtained good reliability indices.
Conclusion: based on the results, it has been verified that the short version works adequately in this sample, and it is proposed to administer the test to a new sample to ensure that the results are replicable and thus, to be able to validate the conclusions.

Title

The Internal Structure of Theory of Mind: Factorial Analysis of Its Evaluation Instruments

Author(s)

Anthony Millán , Edith Aristizaba , Wilmar Fernando Pineda Alhucema [1] ,
Johana Escudero Cabarcas

[1] Universidad Simón Bolívar

Abstract

Introduction: Theory of Mind (ToM) is a fundamental neurocognitive function for Social Cognition. However, there are still not enough validated and standardized instruments to assess this function in the Latin American population, and even fewer in Colombia, which limits its clinical analysis.

Objective: Analyze the internal structure of instruments for the assessment Theory of Mind in children and adolescents through factorial analyses.

Method: The analyzed instruments were the Theory of Mind Battery (ToMB), the Reading then Mind in the Eyes Test (RMET), the Faux Pas Test (FPT), and the Theory of Mind Inventory (ToMI). A heterogeneous sample of 531 participants aged 3 to 17 from the Atlántico Department, Colombia, was used. The analysis was conducted using a non-restrictive exploratory approach with confirmatory aims through structural equation models adjusted by refined regression and evaluated using Tukey's hinges. Normative data were generated from linear regressions and standard deviations of the residuals from the models.

Results: Ten factors were identified for the ToMB, three factors for the RMET, two factors for the FPT, and two factors for the ToMI. All instruments showed adequate psychometric properties.

Conclusions: The factorial analyses confirm that each of the instruments assess different dimensions of Theory of Mind, indicating that ToM is multidimensional. Additionally, the instruments presented good reliability indicators, allowing their inclusion in a unified protocol for clinical use, being a key component in neuropsychological assessment.

Title

Challenges and limitations in the evaluation of theory of mind in Latin America: Methodological and contex- tual challenges.

Author(s)
Paula López , Maria del Pilar Villa , Marcela López , Cristian García Bauza ,
Maria Jose Aguilar [1] , Verónica Zabaletta

[1] Consejo Nacional de investigaciones cientificas y tecnicas (CONICET) y Universidad Nacional de Mar del PLata (UNMdP)

Abstract
The concept of social cognition includes a series of processes that allow people to understand the social world.

The theory of mind as the capacity of people to ascribe mental entities such as desires, beliefs, intentions and emotions has had a great development. The classic tasks that allowed the evaluation of the process were limited to all/nothing tasks, that is, the person evaluated presented or not difficulties in the capacity. Over the years, evaluation techniques were developed mainly in English with adaptations in other languages. A relevant aspect to consider was the effect of cultural differences on the tasks that were developed. On the other hand, new models of the functioning of the theory of mind have approached it as a two-dimensional process (cognitive and affective theory of mind) that presents levels and indicators of development, so it would stop functioning as a unique capacity and with an all/or nothing operation. The objective of the work is to analyze the challenges and limitations presented by the classic and most used tasks in the evaluation of the theory of mind considering their psychometric properties and contextual limitations.

1.25 Symposium : ”Statistical Learning Approaches to Psychometric Modeling Challenges”

Title

Regularized Estimation of the Latent Space Item Response Theory Model

Author(s)
Dylan Molenaar [1]

[1] University of Amsterdam

Abstract
In latent space item response theory (IRT) modelling, both subjects and items are positioned in R dimensional Euclidian latent space. This framework allows for detailed modelling of local dependences among items and subjects, which are assumed to be absent in conventional IRT models. Latent space IRT has demonstrated its value in diverse fields, including intelligence assessment (Kang & Jeon, 2025; Kim et al., 2014), developmental psychology (Go et al., 2022), mental health (Jeon & Schweinberger, 2024), social influence (Park et al., 2023), national school policy evaluation (Jin et al., 2022), and student monitoring (Jeon et al., 2021). However, its broader application is limited by the computational challenges posed by the Bayesian algorithms commonly used for model estimation.

Therefore, in this presentation, a novel estimation procedure is proposed based on regularized joint maximum likelihood estimation. This approach significantly reduces computational demands making it feasible to conduct more robust model evaluations using K-fold cross-validation. The advantages of this method are illustrated in a simulation study and a real data analysis.

Title

“Factor analysis”of Process Data via Psychology-Informed Variational Recurrent Autoencoders for the Anal- ysis of Critical Online Reasoning

Author(s)
Denis Federiakin [1] , Lidia Dobria [2] , Olga Zlatin-Troitschanskaia

[1] Johannes Gutenberg University Mainz; [2] Wilbur Wright College

Abstract

The cornerstone of psychometrics –factor analytical methods –is designed for the interpretable dimensional reduction of response accuracy vector data. This approach can be likened to Variational AutoEncoders (VAEs) with shallow decoders (Urban & Bauer, 2021). However, it is not suitable for analyzing raw process data due to its inability to account for autoregressive dependencies within sequential data. To address such dependencies, various Recurrent Neural Network (RNN) architectures have been proposed, including Variational Recurrent AutoEncoders (VRAEs; Fabius & Van Amersfoort, 2014). This type of RNN creates vector representations of sequential data and reconstructs sequences while preserving autoregressive dependencies.

In this presentation, we propose two custom, interpretation-based recurrent units –one for encoding and one for decoding sequences –tailored for analyzing behavioral data. Both units utilize gating mechanisms to mitigate the vanishing and exploding gradient problem (Hochreiter et al., 2001) while preserving interpretability.

The Recurrent Encoding Behavioral Unit (REBU) is inspired by the Long Short-Term Memory unit (Hochreiter & Schmidhuber, 1997), whereas the Recurrent Decoding Behavioral Unit (RDBU) is developed from contemporary psychological theories on Person-Situation Interactions (Furr & Funder, 2018). The RDBU accounts for situational strength (environmental cues regarding the desirability of potential behaviors) and situational affordances (contextual features enabling the expression of specific traits), resulting in an interpretable decoder structure similar to the VAE-based approach to factor analysis. The architecture consists of two information channels: Long-Term Memory (LTM) and Short-Term Memory (STM). The LTM channel stores information about the vector representation throughout the entire sequence, while the STM channel is responsible for capturing first-order autoregressive dependencies. In RDBU, LTM satisfies the principle of “factor scores”by remaining constant throughout the sequence. This ensures that the learned latent representations are independent of the reconstruction process.

For our analysis, we used log data from 315 higher education students who participated in the Critical Online Reasoning assessment (Molerov et al., 2020). In this scenario-based assessment, students were presented with a problem that lacked a clearly definitive correct answer and were instructed to search online for relevant information. Over the course of 20 minutes, they conducted a brief Internet search and compiled a short essay based on the arguments they discovered. During this process, their search history and actions were tracked.

In our analysis, websites are treated as situational contexts, and clickstream data as actions. The results suggest that it is possible to both generate and interpret vector representations of students’action sequences with acceptable model quality metrics (ROUGE-L and BLEU scores of approximately 0.7).

We also discuss common challenges associated with VRAEs (and their potential solutions). These challenges include longer training times compared to vector-to-vector VAEs; the rare token problem (Yu et al., 2021) which can be addressed through the introduction of “unknown” tokens and balanced reconstruction loss; and posterior collapse, which can be mitigated using input and output STM dropout (Gal & Ghahramani, 2016) combined with KLD annealing (Bowman et al., 2015). Finally, we outline directions for future research.

Title

Variational Autoencoders for Models with Latent Classes

Author(s)
Dylan Molenaar [1] , Karel Veldkamp [2] , Raoul Grasman

[1] University of Amsterdam; [2] Universiteit van Amsterdam

Abstract
Amortized variational inference (AVI) has recently been proposed in the field of Item response theory as a computationally efficient alternative to marginal maximum likelihood estimation (MML). The current study investigates if the computational advantages of AVI for large, high dimensional data carry over to discrete latent variable models. We adapt three techniques from the machine learning literature to the estimation of discrete latent variable models. In separate simulations, we compare the different approaches for latent class analysis, cognitive diagnostic models and mixture IRT models respectively. Results show that AVI is much faster than MML for mixture IRT models. AVI is also slightly faster than MML for LCA models with a large number of classes and items, and is less likely to end up in local minima. Overall we conclude that AVI provides accurate parameter estimates for all three models discussed, but that the computational advantages are most significant for models that have a mixture of discrete and continuous latent variables, such as mixture IRT.

Title

Automated Essay Scoring Using Generative Artificial Intelligence: Illustration of a Systematic Evaluation Framework

Author(s)

Laura Stahlhut , Rudolf Debelak [1] , Kyle Matoba

[1] University of Zurich, EPFL

Abstract

Automated essay scoring systems can support teachers by providing rapid, cost-effective verbal and numerical feedback on student writing. In recent years, these systems have improved significantly with the rise of generative artificial intelligence models based on the transformer architecture. Research consistently shows that these models outperform traditional machine learning approaches across a wide range of natural language processing tasks, including essay scoring.
Despite these advancements, the application of this technology in psychology and education presents several risks, including: a) biased, inconsistent, or inappropriate verbal feedback, b) numerical scores that are highly arbitrary or systematically deviate from human ratings, and c) potential discrimination against specific groups
in scoring and feedback.
In this talk, we illustrate these risks using empirical findings from an ongoing pilot study on automated essay scoring in Switzerland, drawing on examples from several widely used generative models. We then introduce a framework for evaluating the reliability, validity, and fairness of automated essay scoring systems. This framework integrates psychometric principles, such as item response theory and probabilistic test theory, with benchmarking standards from computer science to systematically identify problematic model behaviour. We demonstrate the framework's practical application using anonymized data from our pilot study and summarize main takeaways and challenges.

2 Thursday, 24 July 2025

2.1 Session 14 : ”Dynamic and temporal models in psychology”

Title

Towards a Clearer Understanding of Causal Estimands: The Importance of Joint Effects in Longitudinal Designs with Time-Varying Treatments

Author(s)
Lukas Junker , Ramona Schoedel [1] , Florian Pargent [1]

[1] LMU Munich

Abstract
Longitudinal study designs present unique challenges for causal reasoning. In longitudinal designs, the potential outcomes framework leads to joint effects, which extend average treatment effects to effects of repeated treatments and thus provide a practical measure of cumulative intervention effects over time. Besides explaining the concept of joint effects and how they relate to mediation, we discuss their applicability to psychological research. We focus on their interpretation and whether they can realistically be identified in longitudinal observational studies in psychology. In this context, addressing unmeasured confounding is a crucial aspect of causal inference and mediation analyses, yet it is insufficiently discussed in the psychological literature. To bridge this gap, we propose a class of research designs for psychological studies where treatment assignment is driven by observable covariates so that joint effects can be identified under more reasonable assumptions.

Title

Exploring heterogeneity in temporal dynamics with different extensions of time-varying coefficient models

Author(s)
Esther Ulitzsch [1] , Therese Snuggerud [1] , Steffen Nestler [2] , Sverre Johnson [1] , Oliver Lüdtke [3]

[1] University of Oslo; [2] University of Münster; [3] IPN - Leibniz Institute for Science and Mathematics Education

Abstract

Many psychological interventions aim to uncouple aversive stimuli and negative emotions or cognitions, e.g., the connection between negative triggers and rumination in treatments for anxiety disorders. Understanding whether people differ in when, how effectively, and how enduringly an intervention breaks such links is crucial for its evaluation. Time-varying coefficient models (TVCMs) provide flexible tools for exploring dynamic associations between constructs, approximated by continuous, non-parametric coefficient functions. TVCMs are limited, however, in that they assume coefficient functions to be the same for all persons. We propose and evaluate a flexible, yet parsimonious TVCM extension that allows gauging and quantifying between-person heterogeneity in coefficient functions. To this end, we introduce function-specific latent variables that modulate the coefficient functions, buffering or amplifying them depending on the person's location on the latent variable and the time segment. We illustrate this model extension using intensive longitudinal data collected from 19 patients with anxiety disorders over six weeks —two weeks each before, during, and after an attention training intervention —and explore heterogeneity in the evolving relationship between rumination and nervousness across this period. Our analysis reveals stable rumination-nervousness relationships pre-intervention, varying in strength across individuals. During therapy, the relationship weakens for patients with initially weaker associations but strengthens unexpectedly for those with stronger initial links. Post-intervention, relationships stabilize with minimal rebound effects. To explore how well individual coefficient functions can be approximated by a single latent variable in real data, we contrast model-implied conclusions on individual trajectories against results from case-wise applications of TVCMs.

Title

Time-varying continuous-time models: Extending the framework for dynamic parameters that change over time

Author(s)
Steffen Zitzmann [1] , Martin Hecht [2] , Julian Lohmann [3]

[1] MSH Hamburg; [2] HSU Hamburg; [3] CAU Kiel

Abstract

Continuous-time (CT) modeling has become a widely used approach for analyzing longitudinal psychological data, particularly in ecological momentary assessment (EMA) studies. Traditional CT models assume stationarity—i.e., stable process means and (co)variances over time—which may not adequately capture real-world psychological dynamics. Nonstationarity, which can appear due to time-varying auto- and cross-effects, is often expected in psychological processes but remains underexplored in CT modeling.

In this study, we extend an established CT modeling framework for dynamic parameters that change over time, enabling the analysis of nonstationary processes. We implemented this model in Stan and conducted a proof-of-concept simulation study, demonstrating satisfactory parameter recovery under a very limited set of conditions. An empirical illustration using EMA data on depression and loneliness further highlights the practical applicability of the approach.

While this work provides an initial methodological contribution, further refinement is required, particularly in estimation stability, model complexity, and usability. Future research should explore extensions such as measurement error models, individual differences in dynamic change, and nonlinear time dependencies. This study advances the methodological toolbox for studying psychological processes in their full temporal complexity.

Title

An investigation of the temporal dynamics of careless responding across different populations in experience sampling data

Author(s)

Gudrun Eisele , Inez Myin-Germeys , Ginette Lafit , Lisa Peeters , Olivia Kirtley , Milla Pihlajamaki

Abstract

Background: Recent technological advancements have contributed to the growing popularity of the experience sampling method (ESM) across various fields. However, the intensive nature of ESM raises concerns about careless responding, where participants provide responses without paying sufficient attention to the questionnaire. To better understand careless responding, this study investigated its temporal dynamics in ESM data across three commonly used sample types (community, student, clinical).
Methods: We leveraged four careless responding indicators from previous research: response time, within-beep standard deviation, occasion-person correlation, and inconsistency index. We used multivariate and univariate multilevel models and analyzed the trajectories of the indicators over time, both over the course of the study and over the course of a day.
Results: Our results showed that careless responding is not a stable phenomenon, and the indicators differ in their ability to capture it. Specifically, response time and within-beep standard deviation decreased over the course of the study, suggesting increased carelessness, although these trends were also likely influenced by habituation effects. Inconsistency index remained largely stable, indicating that it might not capture temporal changes in carelessness effectively. Occasion-person correlation showed mixed trends, raising questions about its ability to detect carelessness. The presence of few and small associations among the indicators implies that they flag distinct kinds of carelessness and thus complement each other.
Conclusions: Overall, these findings highlight the importance of accounting for carelessness in ESM studies and demonstrate that these indicators, despite their individual limitations, provide valuable tools for identifying different patterns of careless responding.

Title

Mapping methodological variation in experience sampling research from design to data analysis: A systematic review

Author(s)
Lisa Peeters [1] , Ginette Lafit [1] , Richard Artner [1] , Olivia Kirtley [1] , M. Annelise Blanchard [2] , Gudrun Eisele [1] , Wim Van Den Noortgate [1]

[1] KU Leuven; [2] UCLouvain

Abstract

The Experience Sampling Method (ESM) has become a widespread tool to study time-varying constructs across many subfields of psychological and psychiatric research. This large variety in subfields of research and constructs of interest has contributed to considerable methodological variation. Despite the importance of the methodological choices made by ESM researchers for the quality and veracity of current ESM research and potential future innovation, few have attempted to systematically assess these choices, and to explore their justification. Therefore, the two aims of the current work are to 1) describe the methodological variation in ESM study designs in the recent psychological literature and 2) assess the transparency (i.e., reporting and open science practices) of these studies. These aims are a first step towards a broader goal to improve the methodological quality of ESM research in psychology, contributing to a more rigorous, credible science of daily life.

A broad systematic review of methodological practices in the recent psychological and psychiatric ESM literature addresses these aims. For this review, we developed an extensive list of data extraction items covering the entire workflow of an ESM study, from conception of the research question to reporting of the results. This data is extracted from 150 recently published articles applying ESM in the field of psychology and psychiatry (Stage 1 Registered Report at https://doi.org/10.17605/OSF.IO/ZTVN3). A descriptive and narrative synthesis provides a broad overview of current methodological and reporting practices. These include conceptualization and operationalization of research questions, sampling and design (including within-study adaptivity in these decisions), data pre-processing and analysis, open science practices, and reporting. In this discussion, description of the variation is prioritized over mere identification of the most ‘common’practices. This broad overview can be a starting point for 1) anyone interested in learning about experience sampling, 2) development of guidelines for specific aspects of the method, and 3) future development of experience sampling methodology.

Title

Beyond 'Accuracy': AI vs. Humans as Raters

Author(s)

Aaron Petrasch [1]

[1] University of Munich (LMU)

Abstract

Artificial intelligence is becoming increasingly prevalent in social science research, raising critical questions about its role as a complement or substitute for human raters or judges. While AI-based judgments offer new possibilities, their validity and comparability to human judgments still need to undergo careful examination. This talk presents a framework for evaluating AI as raters or judges, emphasizing the need for assessments that go beyond simple ‘accuracy’ (often measured as correlation with some criterion). First, I will introduce several psychometric methods to compare AI and human judgments systematically. Second, I will present an extension of the Brunswikian Lens model that enables the examination of which textual or visual cues (units of information) are used by humans versus AI in forming judgments. Drawing on empirical examples of text- and image-based evaluations, I will demonstrate how these methods reveal meaningful differences in how judgments are made. Ultimately, I argue that integrating AI into social science research requires at least the same level of methodological rigor as human-based evaluations, ensuring that AI-driven assessments are both valid and reliable.

2.2 Session 19 : ”Advanced statistical models and trust in Science”

Title

On The Interpretation of Vector Autoregressive Models

Author(s)
Ai Ye [3] , Sy-Miin Chow [1] , Charles Driver [2]

[1] Penn State University; [2] University of Zurich; [3] Ludwig Maximillian University in Munich

Abstract
Understanding dynamic processes—whether in psychology, economics, or neuroscience—often requires models that can capture both the evolution of variables over time and the intricate, sometimes instantaneous, interactions between them. Traditionally, the discrete-time vector autoregression (VAR) model has been the workhorse for analyzing time series data, capturing how past values influence current outcomes. The conventional discrete-time VAR is prized for its simplicity and ease of estimation, making it a popular choice for forecasting and exploratory analysis. However, its inability to disentangle mixed effects or account for instantaneous interactions has spurred interest in more refined approaches. SVAR models address this limitation by imposing identification restrictions that allow for the separation of instantaneous and lagged effects, thereby offering a closer approximation to the causal dynamics that may be present in an underlying continuous-time system.

More recently, continuous-time VAR (CTVAR) models have emerged as a promising alternative, providing an even more natural representation of systems that evolve continuously, albeit observed only at discrete intervals. Continuous-time VAR models directly model the underlying process as evolving continuously over time. This approach is particularly appealing when the data sampling rate is insufficient to capture fast-acting influences adequately. CTVAR not only offers a conceptual match to the true dynamics of many natural systems and clearer causal interpretations, but also presents distinct estimation challenges and opportunities.

Previous studies by Demeshko et al. (2015) have shown the mathematical transformations between VAR, SVAR, and CTVAR. In theory, one can obtain the parameter values of one version of the VAR model by using the set of parameter estimates of another VAR variant from a given software. Interestingly, a different VAR model represents an alternative dynamic process, carrying a distinct interpretation. However, few studies fit empirical data using different VAR models provided by corresponding software. Therefore, there is a lack of knowledge about whether the empirical result of an alternative VAR model would match the theoretical parameters from model transformation under an empirical setting; further, how would interpretations from the alternative result change, especially from a causal inference perspective.

In this paper, we systematically compare the three representations—regular VAR, SVAR, and CTVAR—with a dual focus on their interpretation and estimation performance in existing software. We review the relevant software implementations that support these methodologies and conduct both simulation studies and empirical analyses drawn from psychological data. We are particularly interested in investigating whether the causal interpretation from the results will differ as a result of model choice. Our goal is to provide a comprehensive evaluation that not only highlights the strengths and limitations of each approach but also offers practical guidance on selecting the most appropriate modeling framework based on research objectives, data characteristics, and theoretical considerations. Through this comparative study, we aim to contribute to the methodological toolbox available to researchers seeking to unravel the complex, dynamic interplay of variables that characterize many real-world systems.

Title

A sequence sensitive model of encoding precision

Author(s)
Michael Aristodemou [1] , Rogier Kievit [1] , Anna-Lena Schubert , José García Alanis

[1] Radboud University Medical Center

Abstract

Canonical visual working memory models do not incorporate the temporal structure of tasks, despite memory performance in real-life contexts almost invariably operating within sequentially structured activities. To address this gap, we developed a Sequence Sensitive model of working memory and compared its ability to explain the structure of fluctuations in encoding precision and recognition speed to two established resource-based models of working memory, the Population Coding model and the Variable Precision model. We show how Dynamic Structural Equation Modeling can be used to formalize the trial-level dynamic relationship between a neural proxy of encoding precision and recognition speed as predicted by all three models. We compare the three resource-based models by fitting them to data from a large sample of 142 participants who completed 100 trials of a working memory task, the Sternberg task, while their neural activity was recorded using an electroencephalogram. Our results show that the Sequence Sensitive model outperforms canonical candidates in the context of understanding performance on a sequential trial task. However, a visual comparison of model implied and observed values, shows that all three resource-based models only explain a smaller portion of the observed variance. Thus, we should aim to integrate sequential effects into richer, more comprehensive, models that incorporate additional features of working memory, such as the interaction between encoding and retrieval processes.

Title

Extending MPT Models for Continuous Variables: A Comparison of Parametric and Non-Parametric Approaches

Author(s)

Anahí Gutkin [2] , Daniel W. Heck [1]

[1] University of Marburg - Philipps-Universität Marburg; [2] Universidad Francisco de Vitoria

Abstract

To jointly model continuous and discrete variables, parametric (Heck et al., 2018) and non-parametric (Heck & Erdfelder, 2016) extended multinomial processing tree (Extended-MPT) models have been proposed, but they have never been systematically compared. This study compares Extended-MPT procedures in terms of power and robustness using three simulations based on the Weapon Identification Task (WIT). In this context, two statistically equivalent MPT models have been proposed, namely, the preemptive-conflict-resolution model (PCRM) and the default-interventionist model (DIM), which differ solely in their underlying assumptions about the temporal sequence of latent cognitive processes, specifically regarding response times (RTs). The first simulation evaluates the calibration and statistical power of the nonstandard goodness-of-fit test for the parametric approach (i.e., the Dzhaparidze–Nikulin statistic), as well as the ability of different distributional assumptions to fit simulated RT data. The second simulation compares nested models to study the power for testing hypotheses about RTs within each model. The third one focuses on model-recovery performance for the two non-nested models. In all three simulations, we manipulated the size and nature of discrepancies (location/scale or shape) between latent RT distributions, sample size, and parametric assumptions. Our results show that the parametric approach is powerful but highly sensitive to incorrect assumptions about data distribution. In contrast, the non-parametric approach is more robust but less powerful, especially with small samples. Results of proper specification and selection of extended MPT models show that the parametric approach has higher statistical power but is also sensitive to misspecifications of distributional assumptions. The study provides recommendations on when to use each procedure and highlights the importance of appropriate Extended-MPT procedure selection for validating underlying cognitive processes and model selection.

Title

Towards a Global Predictive Model of Trust in Science

Author(s)
Oscar Lecuona [1] , Tobias Wingen [2]

[1] Complutense University of Madrid; [2] FerUniversitët in Hagen

Abstract
Trust in science plays a crucial role in modern societies, shaping individual behaviors related to climate and health. However, research on the factors driving trust in science has been hindered by two key limitations: reliance on verbally specified and thus vague theories that lack numerical precision and an overemphasis on data from the Global North (e.g., WEIRD).
To address these gaps, we aim to develop a quantitative and generalizable predictive model of trust in science using the novel TISP dataset (”Trust in Science and Science-Related Populism”; Mede et al., 2025). This global research project (n = 71,922 in 68 countries) surveyed public perceptions of science. We will use machine learning (e.g., elastic net regression) to create an interpretable predictive model of trust in science using previously published models as reference (Hehman & Neel, 2024). We expect that this model, which we call Trust in Science Model 1.0, will be capable of robustly predicting trust in science across different cultural contexts. We further anticipate that this model will identify core drivers of trust in science worldwide. Nevertheless, the developed model is naturally limited by the predictors available in the TISP dataset. Thus, our main goal is that this model can serve as a valuable benchmark for future theoretical advancements (e.g., using other predictors), allowing researchers to numerically test their alternative models against our Trust in Science Model 1.0.
To illustrate how such model comparisons can advance our understanding, we will take the first step by collecting new data with additional, theory-driven predictors to test potential increments in predicting capacity into a new model (Trust in Science Model 1.1). This comparison will demonstrate the value of predictive modeling in refining theories of trust in science and encourage further research to build upon our findings.

Title

Integrative Pipelines for Preprocessing Mobile Sensing Data

Author(s)
Markus Bühner [1] , Ramona Schoedel [2] , Larissa Sust [1] , David Goretzko , Philipp Sterner [3]

[1] LMU Munich; [2] Charlotte Fresenius University; [3] Ruhr University Bochum

Abstract

Research in the social sciences has traditionally emphasized questionnaire-based assessments, often overlooking the study of real-world behavior. However, with the proliferation of smartphones and digital platforms, researchers now have access to vast amounts of behavioral data generated in people's everyday lives. This shift offers unprecedented opportunities to model diverse phenomena from psychology and beyond, but it also introduces challenges as digital behavioral data are often high-dimensional and low-density, requiring sophisticated preprocessing techniques to extract meaningful variables for formal analysis. Addressing these challenges is essential to ensure the integration of such data into behavioral research in a replicable and sustainable manner.

In this presentation, we introduce a conceptual framework for systematically and transparently reporting preprocessing strategies for mobile-sensing data. Drawing from extensive analyses of smartphone-generated data, including, for example, high-resolution app usage events or GPS logs, our framework focuses on two key dimensions of preprocessing. The first dimension, data enrichment, involves transforming raw data into meaningful variables by adding context, such as integrating multiple data sources or creating new labels. The second dimension, data aggregation, refers to summarizing data at various levels of complexity, ranging from basic descriptive statistics to advanced machine learning models.

To illustrate the application of this framework, we present several preprocessing cases. In the simplest scenario, raw logging data are meaningful enough to be aggregated directly. For instance, app usage events from a specific app, such as TikTok, can be grouped into sessions to derive behavioral indicators like daily usage duration using straightforward algorithms. However, most raw data require enrichment before aggregation. This can involve manually categorizing app usage events into broader categories, such as social media apps, to generate more general variables. Advanced enrichment methods, such as natural language processing, can also be applied, for example, by automatically creating app labels based on their commercial descriptions. Further complexity arises when integrating multiple data sources to provide richer context. For instance, app usage data can be combined with GPS logs to explore patterns, such as how social media usage differs when being at home versus elsewhere. Clustering algorithms, like DBSCAN, can reduce raw GPS coordinates into location categories. Once enriched, these data can be aggregated using statistical models to extract variables that capture relationships across sources. For example, integrating app usage data with ecological momentary assessments (EMAs) can reveal person-level parameters, such as how smartphone-mediated communication relates to social experiences. Again, more complex approaches like machine learning models may also be applied to establish association between data, which may, in turn serve for formal statistical modeling afterwards.

These cases highlight how the complexity of preprocessing affects both the computational demands and the interpretability of the resulting variables. By systematically addressing these challenges, the proposed framework aims to enhance the replicability and interdisciplinary integration of mobile-sensing research. Ultimately, this approach provides practical guidance for leveraging high-dimensional digital data to explore behavior more effectively.

Title

A family of within-test operation-specific learning models

Author(s)
José Héctor Lozano Bleda [1] , Javier Revuelta

[1] Universidad Autónoma de Madrid

Abstract

A family of within-test operation-specific learning models is presented, characterized by fixed-effect versus random-effect learning parameters and by modeling learning from all responses versus only from correct responses. The models, therefore, result from combining the estimation of contingent or non-contingent learning with the consideration or non-consideration of inter-individual variability in learning effects. A simulation study examines parameter recovery and model evaluation. The estimation was conducted by means of Markov chain Monte Carlo using the NUTS algorithm. Model evaluation was based on posterior predictive model checking, while model comparison and selection was based on WAIC and LOOIC. The results show good performance in parameter recovery and model evaluation. An empirical study illustrates the applicability of the models.

2.3 Symposium : "Correlation of cognitive variables with brain activity measured by EEG and fRMI"

Title

Theory of mind and high abilities, EEG analysis

Author(s)
Jesús del Pino Relwani Moreno , África Borges del Rosal [1] , Ernesto Pereda de Pablo , Jesús del Pino Relwani Moreno

[1] Universidad de La Laguna

Abstract
Introduction
Theory of Mind (ToM) refers to the ability to understand and represent both one's own mental states and those of others, enabling the process of mentalizing (Happé et al., 2017). This study posits that individuals with high cognitive abilities may exhibit distinct neural processing patterns during ToM tasks, reflecting a potentially more efficient or elaborate engagement of brain regions associated with social cognition. By employing electroencephalography (EEG) to examine brain activity in individuals with varying intelligence levels during a ToM task, this research aims to shed light on the neural correlates of intelligence-related differences in social cognitive processing, contributing to a more nuanced understanding of the intersection between cognitive
ability and social understanding.
Objective
To explore whether brain processing, measured through electroencephalography (EEG), differs according to intelligence levels during Theory of Mind (ToM) tasks. The study compares individuals with high cognitive
abilities and those with a normal IQ.
Method
EEG measures were used to analyze brain processing. Participants were classified into two groups using the MATRICES-TAI test: 36 university students, 18 with high cognitive abilities and 18 with normal IQ, aged between 18 and 55 years. EEG data were obtained through a reduced version of the Yoni Task to measure responses to ToM stimuli. A time-frequency analysis has been carried out
Results
The study identified differences in brain activity across cognitive, affective, and physical conditions. Cognitive Condition: Differences were observed in beta and gamma frequency bands (23-32 Hz, 36-40 Hz) in prefrontal and frontocentral regions, particularly around 100-300 ms. Affective Condition: Theta (6-7 Hz) and beta (16-29 Hz) frequency differences were detected in anterior regions, particularly around 100-200 ms. Physical Condition: Differences were present in gamma (36-40 Hz) and alpha (9-12 Hz) activity, particularly at 250-550 ms.
Conclusions
The findings highlight differences in brain activity and connectivity across cognitive, affective, and physical conditions.
● Differences in high-frequency activity were found in cognitive tasks.
● Differences in theta and beta frequencies were observed in affective conditions.
● Variability in gamma and alpha activity was present in physical conditions.

Title

Emotional Synergy in Music-Color Combinations: A Neurophysiological Study

Author(s)
Bruma Palacios Hernández , Diana Platas Neri , Ma de la Cruz Bernarda TELLEZ ALANIS [1] ,
Pablo Valdés-Alemán [2]

[1] CITPSI UAEM; [2] Universidad Nacional Autónoma de México

Abstract

This study investigated whether combining musical and chromatic stimuli with congruent emotions produces a synergistic effect on emotional responses, measured through subjective self-reports and electroencephalography (EEG). The sample consisted of 33 participants (20 females; M = 20.3 years, SD = 2.4), all free of moderate to severe depressive symptoms (BDI-II: M = 5.5, SD = 5). Professional musicians were excluded to avoid potential biases in neurophysiological responses. Emotionally validated stimuli (n = 32) were selected, including light cyan (positive) and dark yellow (negative) from the Berkeley Color Project, along with musical excerpts with high and low emotional valence and arousal, based on prior evaluations (n = 85). Stimuli were presented in randomized order, either individually or as music-color combinations. EEG data were recorded using electrodes placed at F3, F4, P3, and P4 according to the 10/20 system, maintaining impedance levels below 5 kΩ. Preprocessing included bandpass filtering (1–30 Hz), independent component analysis (ICA) to remove ocular artifacts, and normalization of absolute power (μV^2) in theta (4–8 Hz) and alpha (8–13 Hz) bands, subtracting resting-state activity and applying natural logarithmic transformations to enhance signal-to-noise ratio. Subjective responses were collected using continuous slider scales to assess valence, arousal, pleasure, and stimulus predominance. Repeated-measures ANOVAs and nonparametric tests (W of Kendall) analyzed emotional and neurophysiological differences across experimental conditions, with corrections applied for sphericity (Greenhouse-Geisser) and effect sizes calculated (η^2 and Cohen's d). Results showed no synergistic effect between congruent music-color pairs. Subjective data indicated music as the dominant emotional stimulus, independent of its combination with congruent or incongruent colors, likely due to its greater perceptual and emotional complexity. EEG findings corroborated this, with the sad color evoking lower theta activity in the parietal region compared to more emotionally activating stimuli. This highlights the influence of stimulus complexity on emotional processing. Methodologically, this research underscores the value of integrating subjective reports with neurophysiological measures to investigate multimodal interactions. These findings have implications for affective neuroscience, virtual environments, and therapeutic applications, providing a framework for developing tools that leverage multisensory integration for emotional regulation and
engagement.

Title

Electroencephalography as a Recording Method in Visual Photosensitivity

Author(s)
Ma de la Cruz Bernarda TELLEZ ALANIS [1] , María Graciela Cano Celestino [2]

[1] CITPSI UAEM; [2] Universidad Autónoma del Estado de Morelos

Abstract

Electroencephalography is a harmless recording technique (Rivera et al., 2023) employed in both clinical and research settings to obtain an electroencephalogram (EEG). It has been recognized as a gold method of brain electrical activity to discover structural or functional damage in people with or without a diagnosis of neurological disease such as epilepsy (Guerrero Aranda, 2020). As a result, methodologies for EEG recording are periodically updated and reviewed to ensure best practices (Kasteleijn-Nolst Trenité, 2012). Addressing the demands of emerging interdisciplinary research, this study details the methodology employed and the data analysis processes used to examine a group of young university students without epilepsy, aiming to identify brain activation responses triggered by graphic images in visual photosensitivity which is mainly related to a high perceptual sensitivity to lights (Fisher, 2022). The international 10-20 system for EEG electrode placement was used to record brain electrical activity from 21 electrodes, incorporating a Vision Test, Baseline Recording (BR), and Pattern Sensitivity Test (PST). Changes in brain electrical activity were analyzed using clinical and psychological approaches, focusing on detecting biomarkers of abnormalities during the recording process (BR, PST). Additionally, frequency analysis (Hz) and band power (μV^2/Hz) were evaluated, with special attention to delta (0.2-3.5 Hz), alpha (8-12.5 Hz), and gamma (30 90 Hz) bands following the PST period. The fast Fourier transform method was employed for this analysis. This work hypothesizes that graphic images with specific structural features may modify the normal brain electrical activity in young people with undiagnosed visual photosensitivity.

Title

Resting-State Brain Activity and Connectivity in Individuals with High Cognitive Abilities

Author(s)
Juan Manuel Plata Bello [1] , África Borges del Rosal [1] , Dante Noah Jorge Sálamo ,
Jesús del Pino Relwani Moreno

[1] Universidad de La Laguna

Abstract
Introduction
Recent investigations point to a link between intelligence and more efficient neural processing, suggesting that people with higher cognitive performance tend to have stronger integration among key brain areas and reduced redundant activity (Jung & Haier, 2007). Grounded in this concept of neural efficiency, the current study examines resting-state functional activity and connectivity in individuals with above-average cognitive abilities compared to those with average IQ, aiming to shed light on how brain organization differs according to intelligence level.
Objective
To explore whether resting-state brain activity, as measured by amplitude of low-frequency fluctuations (ALFF) and functional connectivity, differs according to intelligence levels. Specifically, the study examines differences in brain connectivity between individuals with high cognitive abilities and those with an average
IQ.
Participants
The sample consists of 10 women and 10 men, with an average age of 20.1 years. All of them students of the ULL
Method
Resting-state functional magnetic resonance imaging (rs-fMRI) was used to analyze brain activity and connectivity.
Participants were classified into two groups based on their IQ scores:
● High cognitive ability group (IQ ≥ 120)
● Average cognitive ability group (IQ 90–119) The sample consisted of 10 participants per group,all university students from Universidad de La Laguna. ALFF was measured in key brain regions, including the ACC, left frontal pole (lFP), and subcallosal cortex, to assess spontaneous neural activity. Functional connectivity analyses were conducted to examine relationships between these regions and other cortical and subcortical structures.
Results
The study revealed differences in brain activity and connectivity between individuals with high and average intelligence.
● Higher intelligence was associated with increased ALFF in key brain areas, indicating greater neural efficiency.
● Differences in connectivity patterns were observed, suggesting variations in the way brain networks communicate and integrate information.
● Regions linked to executive functioning and emotional regulation showed notable distinctions between the groups, reinforcing the idea that intelligence influences brain organization.
Conclusion
The findings highlight key differences in spontaneous brain activity and connectivity between individuals with different intelligence levels.
● Higher IQ individuals exhibited stronger brain activity in cognitive control regions.
● They demonstrated more efficient functional connectivity, particularly in prefrontal networks.
● The results support the neural efficiency hypothesis, suggesting that intelligence is associated with optimized
brain function.

These findings contribute to a better understanding of how intelligence shapes brain organization, emphasizing that intelligence is not simply about higher activity in specific areas but about the efficient integration of multiple networks. Future research could further explore how these neural differences relate to cognitive performance in complex tasks.

Title

Study of Brain Activity at Resting State by Functional Magnetic Resonance Imaging in People with High and Low Sensitivity

Author(s)

Juan Manuel Plata Bello [1] , África Borges del Rosal [1] , Julio Manuel Plata Bello ,
Michelle Padrón

[1] Universidad de La Laguna

Abstract

Introduction: Sensory processing sensitivity (SPS) is an inherited personality trait that determines people to feel, think and interact with others differently from others. Several research studies have shown these differences through studies on brain processing. Objective: To analyse the differences in resting brain activity, as determined by functional magnetic resonance imaging, between people with high and low sensitivity, in order to test their neural processing. Method: Two study groups of 10 participants were selected by sensitivity condition (with a mean of 78.3 in high sensitivity and 39.3 in low sensitivity) according to their response to the Spanish version of the HSC (Higly Sensitive Child) scale developed by Pluess, where the mean age of the participants was 23.8 years, with 13 women and 7 men, in order to subsequently record the basal functional image of the brain, through fMRI, calculating the Fractional Amplitude of the Low Frequency Fluctuations signal (fALFF). Results: In a first analysis, the results showed a positive relationship between fALFF and PAS levels (high and low) at the level of the left parietal lobe and the left cerebellar hemisphere, corresponding to the posterior lobe, and a negative relationship at the level of the left thalamus, bilateral medial frontal lobe (including the anterior cingulum) and the left superior temporal lobe. In a second analysis, a study was conducted between each subject's fALFF and their individual score on each of the PAS factors (AES, EOE, LST). In the AES analyses, negative relationships were only observed for both hemispheres located in the right temporal lobe (medial temporal gyrus, fusiform and parahippocampal), left temporal lobe (left superior temporal gyrus) and at the level of the cuneus. In the correlation analysis between fALFF and the EOE factor, a positive relationship was observed at the level of the left parietal lobe (including the postcentral gyrus, supramarginal and inferior parietal lobe), and a negative relationship in the left thalamus and bilateral medial regions at both parietal (cuneus and precuneus) and frontal levels. Finally, correlation analysis between fALFF and LST showed a positive relationship at the left parietal level and with the posterior lobe of the left cerebellum, and a negative relationship in the left thalamus, some regions of the left temporal lobe (medial and superior temporal gyri), medial frontal and parietal regions. Discussion: It has been observed that most of the regions that show some kind of significant result belong to regions related to the somatosensory system, such as the regions that form part of the parietal lobe, as well as the left postcentral gyrus, which correspond to the somatosensory cortex. On the other hand, the thalamus is considered one of the most important sensory neural regions, being the main intermediary, together with the cerebral cortex, in the processing of emotional stimuli, with the exception of olfaction. Research has also shown that the prefrontal and temporal cortex and some limbic structures, such as the cingulate, play a fundamental role in the development of empathic quality.

2.4 Symposium : ”Standards and Guidelines for Educational and Psychological Assessment: Continuing the Conversation”

Title

Difficult Conversations in Revising the Standards for Educational and Psychological Testing

Author(s)
Stephen Sireci [1]

[1] University of Massachusetts Amherst

Abstract

Psychological tests are essential tools that help psychologists make decisions about people. The Board of Assessment (BoA) of the European Federation of Psychologists' Associations (EFPA) has various projects aimed at improving tests and testing practices across Europe and beyond. In this presentation, we share two BoA projects. The first is the BoA's flagship project, which focuses on tests: the Test Review Model. This model provides a systematic framework for reviewing and assessing psychological tests based on several criteria (materials, reliability, validity, norms, etc.). It has recently been updated to incorporate aspects of digital assessments, inclusivity, and diversity, among others. The second project focuses on test user standards, defining the competencies and skills required to ensure proper test use. In this presentation, we will highlight the main features of these projects, emphasizing their impact and the challenges of implementation.

Title

The Role of the New PISA Quality Standards to Promote Fairness

Author(s)
Javier Suárez-Álvarez [1] , Mario Piacentini

[1] University of Massachusetts Amherst

Abstract
The value of the Programme for International Student Assessment (PISA) in informing evidence-based policymaking relies on the degree of precision with which population-level statistics are estimated and reported. But also, the degree to which those aggregate statistics can be meaningfully compared (e.g., country-level mean scores) and the interpretations made based on those comparisons are valid for the intended purposes. Although validity, comparability, and reliability are important components of fairness, they do not address all issues of fairness in assessment. Fairness provides an important additional lens for ensuring the validity of research, policies, and all other aspects of a testing program to promote positive, intended outcomes and minimize negative ones. PISA Technical Standards serve as a set of criteria for post hoc data adjudication (decisions on whether the data for a specific country are of sufficient quality for inclusion in the international reports), but do not address aspects of assessment quality like fairness, validity, reliability, and comparability. This presentation will describe the principles behind the new PISA Quality Standards to deliver relevant, rigorous, and transparent information to policymakers through assessment instruments that provide fair, valid, and reliable data comparable across cultural settings, time, and groups. The presentation will focus on major threats and suggested guidelines for ensuring fairness, validity, comparability, and reliability. The presentation will stress the similarities and differences with other professional standards.

Title

The Role of the EFPA Board of Assessment in Promoting Testing Standards

Author(s)

Urszula Brzezinska , Nigel Evans , Mark Schittekatte , Ian Florence , Helen Baron , Dragos Iliescu , Ana Hernández

Abstract

Standards play a crucial role in guiding practices in Educational and Psychological Assessment. Various professional associations continuously update guidelines to support practitioners in assessment-related processes, leading to the emergence of different approaches. However, how do these associations gather information to propose new standards? To what extent do they consider users'experiences? Do their ultimate goals differ? Moreover, how can professionals navigate and integrate the diverse guidelines available? This discussion will explore the complementarity between different approaches and examine ways to support users in effectively applying multiple standards.

Title

Discussion about Current Standards and Guidelines for Educational and Psychological Assessment

Author(s)

Isabel Benítez Baena [1]

[1] University of Granada. Mind, Brain and Behavior Research Center (CIMCYC)

Abstract

The Standards for Educational and Psychological Testing have been published by the American Psychological Association, the American Educational Research Association, and the National Council on Education since the 1950s. They are currently under revision, and the forthcoming version, is expected to be published in 2026. In this presentation, a member of the Joint Committee revising the Standards will discuss the difficult questions and issues that are being considered for the revision. These issues include whether the five sources of validity evidence are sufficient and appropriate for guiding practitioners, the distinction and overlap between validity and fairness, the desire to make the Standards more authoritative, and the desire to make the Standards more understandable to lay audiences. The presenter will take questions from the audience to understand what other issues members of EAM think are important to consider in revising the Standards.

2.5 Session 20: ”IA y M learning”

Title

Artificial Intelligence and learning: Contributions to the state of the art in Education

Author(s)
Ana Pereira Antunes [1] , Carlota Fernandes [2]

[1] University of Madeira & CUIP & CIEC; [2] University of Madeira

Abstract

Introduction: Nowadays, artificial intelligence is a topic that evolves speculation, questioning and usability. Also in Education is an emergent, fabulous, and intriguing topic that deserves attention by teachers, students, politicians, and researchers. This paper is constructed based on the activities of MathIA project, an Erasmus project (ID KA220-SCH-0486161A) to promote de development of students'Mathematics skills.
Objectives: The main goal of this paper is to analyze the publications on the topic of artificial intelligence in the field of Education.
Methods: To achieve the goal a documentary analysis was conducted. The data was collected on b-On and Google platforms according to previous selected descriptors such as: Education AND Artificial intelligence AND Teaching Learning. The search was conducted to select texts published in the last five years (2025) and the language of the publications were Portuguese, English, Spanish and Italian. The selected publications were 107 after a reading the title and the abstract of a larger number of publications that emerged in the search.
Results: This is an ongoing work, and data analysis is not finished yet. However, preliminary results indicate some research has already been made in the topic and the number of publications differs according to the language publication. In the oral presentation we will present complete findings according to several categories including the studies'specific themes and the research methods used.
Conclusions: The topic of AI and Education is a current and demanding topic and researchers seem to pay attention to the (re)evolution of AI in Education.

Title

Machine Learning for Psychological Research: Benchmarking Forecasting Performance of Deep Learning Models for Longitudinal Data

Author(s)
Andre Nedderhoff , Steffen Zitzmann [1] , Martin Hecht

[1] MSH Medical School Hamburg

Abstract

The study of longitudinal data has been a cornerstone of psychological research, shaped significantly by the foundational work of Baltes and Nesselroade (1979). With the rise of mobile IT tools, interest in modeling intensive longitudinal data has surged. Traditionally, psychologists have relied on time-series analysis methods such as the Random Intercept Cross-Lagged Panel Model (RI-CLPM) and continuous-time models to analyze (intensive) longitudinal data and examine temporal relationships between variables.
In recent years, advances in Machine Learning (ML) have opened new possibilities for analyzing longitudinal data, particularly in forecasting future values. Our study explores the potential of ML models for this purpose by simulating time series based on a discrete-time model, varying key design factors such as the number of input and forecasted time points, sample size, and temporal drift.
We evaluated these simulated time series using two established psychological models, the discrete time RI-CLPM (Hamaker & Grasman, 2015) and the continuous-time RI-CLPM (Oud & Delsing, 2010) - and compare their forecasting performance with two widely used ML models: Long Short-term memory networks (LSTM; Hochreiter & Schmidhuber, 1997) and Transformer neural networks (Vaswani et al., 2017). Our study assesses each model's forecasting capabilities in a univariate forecasting scenario, using bias and root mean square error (RMSE) as performance metrics.

References

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), Longitudinal Research in the Study of Behavior and Development (pp. 1-39). Academic Press.

Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. Frontiers in Psychology, 5. https://doi.org/10.3389/fpsyg.2014.01492

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Oud, J. H. L., & Delsing, M. J. M. H. (2010). Continuous Time Modeling of Panel Data by means of SEM. In K. Van Montfort, J. H. L. Oud, & A. Satorra (Eds.), Longitudinal Research with Latent Variables (S. 201–244). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11760-2_7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Title

AI and students learning of Mathematics: A bibliometric study

Author(s)

Carlota Fernandes [1] , Ana Antunes [2]

[1] University of Madeira; [2] University of Madeira & CUIP & CIEC

Abstract

Introduction: Mathematics skills are essential for academic curricula despite the difficulties some students reveal on the topic. To help the development of adolescents' mathematical skills is being implemented the Erasmus project: MathIA (Artificial Intelligence Model to enrich and improve mathematical skills in adolescent students) and this study is framed on this project.
Objectives: In this paper we want to present the publications that are being made concerning artificial intelligence and Mathematics teaching and learning and, consequently, discuss the pertinence of this topic in Education.
Methods: A bibliometric study was conducted to achieve the defined goal. An online search was made on b-On and Google platforms searching publications between 2020 and 2025 and using descriptors such as Mathematics AND Artificial intelligence AND Teaching Learning. The four languages of the MathIA project were considered in the articles search: Portuguese, English, Spanish and Italian. From the search 31 articles were selected after title and abstract reading. Currently, full texts are being analyzed.
Results: As the data analysis is being made, we do not have results yet. We expect to present them in the congress grouped by categories like the number of studies, the specific topic studied, the population, and the research methodology.
Conclusions: At the end of the data analysis, we hope findings will make it possible to understand the current trends in IA research in mathematics teaching as well as the way this topic is being studied.

Title

Model IA European MathIA Project

Author(s)

Ana Fuensanta Hernández Ortiz [2] , Javier salazar [1] , FERNANDO JIMÉNEZ [1] , José Tomás Palma [1] , Andrej Franulic [3]

[1] universidad de murcia; [2] Profesora Universidad de Murcia; [3] EAM2025

Abstract

This paper presents the methodology and preliminary results of an Erasmus+ KA2 project that aims to leverage artificial intelligence to improve mathematics skills in secondary school students. The project fosters equity and diversity by addressing the needs of high-achieving students, those with learning difficulties, and those with special educational needs. Furthermore, it aims to support STEM educators by developing a resource bank for lesson planning and reducing students'test anxiety. We propose a methodology to build, evaluate, and validate machine learning classification models capable of predicting learning difficulties in mathematics among secondary school students. Data are being collected from secondary schools in the Region of Murcia, Colegio Vicente Medina, and from Tenerife, Colegio Mayco (Spain), charter schools with two secondary lines in each grade, as well as from partner institutions in Italy, such as Instituto Via Angelini, from Pavia, a public secondary school. The procedure for the design and development of the model is being carried out through an exam design with exercises designed by the secondary school teachers of each of the centres and with inter-judge validation. Once validated, it is being inserted into a database to be carried out through the Moodle platform and for the students of all the centres involved to have access to it. Once the data has been collected, it will be fed into the machine to obtain the first results and validate them, taking into account the variables of gender, response time, level of difficulty of each content and evaluation criteria according to the curriculum of this stage. Meanwhile, synthetic data sets have been generated with characteristics similar to the real data to allow a preliminary evaluation of the proposed model. The machine learning methodology covers data preprocessing, feature selection, imbalanced classification and rigorous evaluation and validation processes. Preliminary findings suggest that the proposed approach is promising for accurately identifying at-risk students and for obtaining results on the level of mathematical competence of each student, both to determine whether they are above, below, or above average for their age, which will ultimately contribute to personalized interventions and to improving educational outcomes.

Title

Incorporating longitudinal variability in prediction models: a comparison of machine learning and logistic regression

Author(s)

Liza de Groot [1] , Martijn Heymans , Jos Twisk , Almar Kok

[1] Amsterdam UMC

Abstract

Background: Clinical prediction models estimate health outcome probabilities, aiding decision-making. Incorporating longitudinal data can improve predictive accuracy, but complexity and interpretability challenges often limit its use. While the predictive value of a predictor's mean and change over time is well-documented, the variability around this change remains underexplored. Traditional regression analyses, though interpretable, struggle with repeated-measurements data, e.g., collinearity and numerous potential predictors. Machine Learning (ML) methods, i.e., Random Forest and Lasso regression, may better handle repeated-measurements data. This study evaluated the predictive value of three longitudinal parameters: mean, change, and variability for a time-independent binary outcome and compared ML methods with logistic regression.
Methods: Random Forest, Lasso regression, and logistic regression were compared regarding selected predictors, interpretability of results (predictor-outcome relationships), and predictive performance (AUC and calibration curves). Depression (clinically significant symptoms) was the binary outcome, with 81 longitudinal parameters (mean, change, variability) as predictors. Models were trained on 70% of data and internally validated on 30% using the Longitudinal Aging Study Amsterdam (LASA).
Results: All methods identified similar important predictors, including variability parameters. Analyses incorporating variability parameters achieved slightly higher AUCs than those without. Regression coefficients of Lasso regression and logistic regression were consistent with the predictor-outcome relationships reflected in Partial Dependency Plots from Random Forest. Predictive performance was comparable across methods (test, AUC: 0.768–0.775). Calibration curves revealed overestimation, with predictions remaining low (<0.6).
Discussion: The high false-negative rates and low predictions, probably are a result of the imbalanced dataset (13.05% depression prevalence). Sensitivity analyses with Random Under Sampling of the majority class resulted in predicted prevalences closer to the observed, a greater variation in predicted probabilities in calibration curves; however, AUCs did not improve.
Conclusion: Advanced ML techniques did not outperform logistic regression in predictive performance. However, incorporating longitudinal predictors'variability around change over time is critical for improving clinical prediction models. This is particularly relevant in contexts with long follow-up periods where predictors are likely to change on average over time, e.g., when following an aging population.

2.6 Session 3 : ”Measurement and Analysis of violence and discrimination”

Title

Measurement structure and invariance of intimate partner violence against women in lower- and middle-income countries

Author(s)
Angie Bengtson [1] , Regine Haardoerfer [1] , Irina Bergenfeld [1] , Cari Jo Clark [1]

[1] Emory University

Abstract
Intimate partner violence (IPV) is a pressing international issue affecting at least 27% of women and girls. However, accurate global assessment of IPV prevalence is limited by a lack of consensus around the domains of IPV and sparse evidence on cross-country comparability. We aimed to assess the measurement structure and regional invariance of IPV scales in large, population-based surveys using the same standard item sets to measure physical, sexual, and emotional IPV, as well as controlling behaviors. Using confirmatory factor analysis (CFA), we tested unidimensional, multifactorial, hierarchical, bifactor, and bifactor S-1 models for lifetime and past year IPV across 46 lower-and- middle-income countries. We then assessed the invariance of the best-fitting model across countries within the same world region using multiple-group CFA. Although other models showed good fit in most countries, bifactor and bifactor S-1 models had the best fit across all countries and showed strong or strict invariance within most regions. Most bifactor models, especially without controlling behaviors, were primarily unidimensional; IPV can therefore be conceptualized as a single construct with nuanced facets. Researchers seeking to model IPV should consider the bifactor/bifactor S-1 model, unidimensional model, or simple summative measures incorporating physical, sexual and emotional domains. In alignment with prior research, controlling behaviors should be modeled separately to avoid parameter bias. Finally, the interrelated nature of physical and emotional IPV underscores the need to incorporate emotional IPV into global monitoring and reporting structures

Title

Understanding Perceived Vulnerability to Intimate Partner Violence: A Bifactor(S-1) Model Exploring the Role of Sexism and Violence Myths Against Women

Author(s)

Rocío Vizcaíno-Cuenca [1] , Hugo Carretero-Dios [1] , Michael Eid [2] , Mario Lawes [2] , Mónica Romero-Sánchez [3]

[1] Department of Methodology of Behavioural Sciences, Faculty of Psychology, Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain; [2] Department of Methods and Evaluation, Faculty of Education and Psychology, Free University of Berlin, Berlin, Germany; [3] Department of Social Psychology, Faculty of Psychology, Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain

Abstract

Introduction:
Intimate partner violence against women is a significant social and public issue, with myths that justify and minimize intimate partner violence playing key roles in its perception and perpetration. Although a measure to assess myths about intimate partner violence against women (AMIVAW) has been developed and validated, the empirical relationship between AMIVAW and other measures of sexist attitudes is strong, with some correlations approaching levels that suggest a common construct. Therefore, we aimed to explore whether myths about violence against women in intimate partner relationships (AMIVAW) differ from other constructs related to feminist attitudes, sexism and violence against women (i.e., hostile sexism, benevolent sexism, rape myths), or whether they represent different manifestations of the same underlying construct.
Method:
A total of 485 participants (199 men and 286 women) from the United States completed a survey that included measures of feminist attitudes, sexist attitudes and myths about intimate partner violence. First, we examined the relationships among these variables using a first-order Confirmatory Factor Analysis (CFA) model. We hypothesized positive relationships between AMIVAW and sexist attitudes (both hostile and benevolent sexism) and rape myths, and a negative relationship between feminist attitudes and these constructs. Second, we used a bifactor (S-1) model to analyze the specific contributions of each variable while controlling for AMIVAW, hypothesizing that these attitudes are distinct but partially overlapping with AMIVAW. Finally, we explored how AMIVAW predicts the perception of vulnerability to intimate partner violence, expecting AMIVAW to predict vulnerability in women, but not in men.
Results:
The tested models showed an acceptable fit (CFI > .95, SRMR < .06, RMSEA < .08). The results confirmed our hypotheses and further revealed that AMIVAW was the strongest predictor of vulnerability to intimate partner violence.
Conclusion:
This study provides a deeper understanding of the relationship between measures of feminism, sexist attitudes and myths about violence against women.

Title

Advancing Quantitative Methodologies to Achieve Equity

Author(s)
Regine Haardoerfer [1]

[1] Emory University

Abstract

Despite great efforts, inequities in life expectancy, health, and quality of life are persistent. Scientific revolutions are preceded by paradigm shifts (Kuhn, 1997). To achieve equity for all, we need to understand the current quantitative inquiry paradigms. To do so, quantitative methodological literacy is needed. The proposed presentation Advancing Quantitative Methodologies to Achieve Equity will provide quantitative researchers with the tools to interrogate and advance their quantitative methodologies. In quantitative research, the focus has been almost exclusively on methods. Methodologies, where the term is used, focus on techniques and procedures for specific approaches, such as survey methodology. This limited definition misses the philosophical and societal underpinnings of those approaches, specifically, ontology, epistemology, and axiology as well as social position and how those have limited both theoretical frameworks and research methods deemed useful. The aims of this presentation are: 1) Define research methodologies, present their colonial history, impact on research and persisting inequities, and non-positivist methodologies developed to-date. 2) Introduce a framework that can be used to advance quantitative researchers'methodologies through a) interrogation of research standpoints, theoretical frameworks, and research methods b) assessing the influences of -isms on research c) practicing inclusive team science and d) Using participatory research methods and 3) Outline steps on how to develop a praxis of advancing quantitative methodologies through practices such as researching back, advancing contextually relevant and theory-driven research driven by liberatory and transformative intent. The long-term goal of this work is to contribute to a paradigm shift in quantitative methodologies by making methodological literacy a key component of researcher education which will subsequently strengthen quantitative research and advance knowledge needed to achieve equity. It will advance methodological and methods pluralism allowing for broadening scientific knowledge. Thus, the proposed presentation aims to present tools to advance methodological literacy needed to achieve equity.

Title

What is a 'resilient' symptom network? Assessing multiple response trajectories to stressful events using network theory

Author(s)
Gaby Lunansky [1]

[1] Amsterdam UMC

Abstract

The network theory of psychopathology proposes that individuals demonstrating resilient responses to adverse events are characterized by symptom networks with low connectivity, indicating that symptoms do not easily co-occur despite facing stressful events. We investigated this hypothesis using the US Health & Retirement Study, a nationally representative longitudinal survey monitoring older adults'physical and mental health. We included 7023 older adults, who experienced severely stressful events, including cancer diagnoses, bereavement or divorce. Using Growth Mixture Modeling, we identified four response trajectories before, during, and after the event, consistent with response patterns found across many studies. We then estimated Ising model networks of depressive symptoms within each time point and response trajectory class. Contrary to expectations from the network theory of psychopathology, symptom networks from the resilient trajectory class were characterized by high connectivity. However, threshold parameters in this group were high, indicating low likelihood of symptom activation. Furthermore, we estimated stability landscapes for all networks. Stability landscapes summarize the most likely symptom dynamics given network parameters, quantifying the stability of symptom states. We found that the stability landscapes for the resilient networks aligned with the response trajectories observed in the data. This alignment suggests that the combination of high connectivity and high thresholds reflects a resilient state characterized by a low probability of symptom activation. We conclude that resilient responses to adverse events can be characterized by different combinations of network connectivity and threshold parameters. Studies using symptom networks should consider both type of parameters to interpret their results.

Title

Investigating Differential Item Functioning of the Reading Comprehension Section of a High-Stakes Test across Booklet and Gender: An Analysis with Recursive Partitioning Rasch Tree

Author(s)
Farshad Effatpanah Hesari [1]

[1] TU Dortmund University, Dortmund, Germany

Abstract

Differential item functioning (DIF) is a critical psychometric feature in educational testing, evaluating whether distinct subgroups respond differently to test items despite possessing the same latent ability. DIF occurs when examinees with equivalent underlying traits have varying probabilities of correctly responding to an item, influenced by subgroup membership. This study investigated whether the position of response options in the reading comprehension section of a high-stakes test contributes to DIF across subgroups defined by booklet and gender. Unlike the conventional DIF methods, the Rasch tree does not require a pre-specification of groups for detecting DIF. To investigate DIF of the test, the study analyzed item responses from 10,000 examinees on 20 multiple-choice items presented across four booklets using the ‘psychotree’package in R. The test content and item positioning were consistent across booklets, with only the positions of the answer options varying. Following a Rasch model fit assessment, the Rasch tree analysis identified three non-predefined nodes, highlighting varying patterns of item difficulties. Four items were flagged as exhibiting DIF, with findings revealing that a combination of booklet and gender influenced test performance. Notably, booklet emerged as a primary source of DIF, particularly in interaction with gender, as DIF patterns varied between females and males, with some items displaying more pronounced deviations. A content analysis explored potential sources of DIF. One hypothesis considered was the positional effects of answer choices, such as the primacy and recency effects, where examinees might favor the first or last options. If correct answers were positioned accordingly in certain booklets, this could alter accuracy rates. It turned out that the performance of examinees in Node 2 was affected by the position of response options, specifically due to the close proximity between the correct answers and the distractors. Another hypothesis examined examinees’strategic guessing based on perceived patterns in correct answer positions. However, this was refuted, as no discernible pattern emerged across booklets. The content of flagged items was also scrutinized to identify inherent features requiring varying levels of cognitive load. Items demanding deeper comprehension or subtle interpretation, such as identifying the main idea or understanding specific details, were more likely to show DIF. Variations in cognitive styles, educational backgrounds, and test-taking strategies among subgroups likely amplified these effects. For instance, methodical examinees considering all options may be less susceptible to positional biases, while those relying on heuristics or elimination strategies might be more affected by answer placement. This study underscores the potential for DIF arising from subtle interactions between booklet design, answer positioning, and subgroup characteristics, particularly gender. The Rasch tree effectively captured these interactions, providing a detailed understanding of how covariates influence test performance. The findings emphasize the importance of investigating non-obvious sources of DIF, such as response option positioning and its interaction with cognitive processes. This research contributes to the field by demonstrating the utility of the Rasch tree model in detecting and understanding complex DIF patterns, ultimately enhancing the fairness and validity of high-stakes assessments.

Title

Enhancing Model Visualization in Statistical Analysis: Introducing the R Package MoPlot

Author(s)

Stefano Dalla Bona [1] , Marianna Musa [1] , Lorenzo Atzeni [1] , Umberto Granziol [1]

[1] University of Padova

Abstract

In many statistical analysis methods, interpreting model coefficients solely from the coefficient table can be difficult or counterintuitive (i.e., the coefficient of polynomial contrasts or in generalized additive models). To facilitate understanding, researchers often rely on effect plots. However, most statistical software generates plots that primarily display trends/curve for continuous variables or raw/predicted means for categorical variables, without providin key details such as the contrast types used, effect sizes of each effect, or the information about statistical significance.

We introduce MoPlot, a novel R package designed to address these limitations. By simply inputting a fitted model, MoPlot generates plots for both single and interaction effects while explicitly accounting for the contrast coding scheme applied in the model. The package automatically visualizes planned comparisons, providing a clearer representation of the related categorical predictors. Additionally, it generates comprehensive figure captions detailing the effects displayed, including significance information.

Furthermore, MoPlot includes a “coefficient mode,”allowing users to visualize model coefficients alongside their corresponding effect sizes—a crucial feature for meta-analytic studies. The package supports a wide range of contrast coding schemes available in R, such as treatment, sum, sliding difference, polynomial, Helmert, reverse Helmert, and custom-defined contrasts. Real-world data examples will illustrate how MoPlot enhances interpretability in models with both continuous and categorical predictors.

Finally, we discuss the implications of using MoPlot across various statistical models and highlight its potential contributions to reproducibility and open science. By improving the clarity of effect visualizations, MoPlot provides researchers with a more informative and transparent approach to statistical analysis.

2.7 Session 4 : ”Structural equation models and model evaluation”

Title

Chasing Normal Linear Unicorns – More Realistic Fit Indices for SEM

Author(s)
Charles Driver [1]

[1] University of Zurich

Abstract

Structural Equation Modeling (SEM) remains a cornerstone of psychological research, providing a means of testing complex theoretical models. However, traditional model fit indices assume that data can be sufficiently summarized by a single covariance matrix and mean vector, an assumption rarely scrutinized despite psychological data frequently exhibiting nonlinear dependencies and subgroup heterogeneity. We compare standard SEM fit indices with alternative approaches that assess misspecification directly from raw data residuals. Specifically, we explore the Hilbert-Schmidt Independence Criterion (dHSIC) in both multiple bivariate and multivariate contexts, the Heteroscedasticity Fit Index (HFI), and multivariate normal energy tests as means of detecting missing dependencies in SEM models. These techniques allow for targeted misspecification checks that distinguish between general model assumption violations and the more crucial issue of unmodeled variable dependencies. While traditional fit indices fail to detect nonlinear misfit, approaches such as dHSIC focus on missing nonlinear dependencies while remaining insensitive to deviations from univariate normality – important, as SEM models primarily concern inter-variable relationships rather than marginal distributions. Recognizing the distinct roles of different fit checks is crucial for advancing SEM beyond its reliance on global covariance structures, enabling more robust and realistic model assessment in psychological research and theory development.

Title

Abandon All Thumbs Ye Who Model: Model Fit Evaluation in SEM for a New Century

Author(s)
Edita Chvojka [1]

[1] Universiteit Utrecht

Abstract

Structural equation modeling (SEM) is a central statistical technique for evaluating intricate relationships between latent constructs. A crucial ingredient of SEM is model fit assessment, which often relies on universal cutoffs, such as RMSEA ≤ .05 or CFI ≥ .95. These thresholds stem from past simulations and experience but fail to account for variations in model and data characteristics. In this talk, I will discuss the limitations of these cutoffs, particularly those proposed by Hu and Bentler, and highlight how model complexity, reliability, localization of misspecification and other model and data properties influence fit indices. I will present the results of a systematic review and demonstrate how fit measures behave under different conditions and why seemingly universal cutoffs often lead to misleading conclusions. Finally, I will propose a more nuanced approach to model fit evaluation, emphasizing conceptual understanding and situational assessment. This talk aims to encourage a more thoughtful approach to SEM, sparking discussion on improving fit evaluation practices and reducing the prevalence of poorly fitting models in social science research.

Title

SEMtrees in Longitudinal studies: The utility of goodness-of-fit indices for building theory

Author(s)
Elisabeth Valeriano-Lorenzo [1] , Teodoro del Ser [2] , Carmen Ximénez [1]

[1] Universidad Autonoma de Madrid; [2] CIEN Foundation, Queen Sofia Foundation Alzheimer Centre. Madrid, Spain

Abstract

Longitudinal data allow us to examine both simple and complex types of change over time, but also require defining the expected types of change based on a theoretical background. In this context, SEMtrees (Brandmaier et al., 2013; Zeileis et al. 2008) is a powerful statistical method for building models to examine parameters of: (1) intraindividual change, and (2) interindividual differences, and also understand how and why both parameters are articulated, revealing patterns and interactions in data. In other words, SEMtrees advance the understanding of the relationships and mechanisms between variables, through the tree-based structure, providing a deeper, more theory-driven analysis.

The current work illustrates the application of SEMtrees to empirical data in the context of Alzheimer's disease, and discusses the substantive interpretation of the results. In particular, it highlights how and why aspects of cognitive evolution are linked to the biology of the disease, neuronal activation, and the evolution of plasma biomarkers.

Additionally, this work analyzes the performance of some of the most well-known goodness-of-fit indices for SEM models: CFI, TLI, RMSEA, SRMR, and SRMRu within each latent class grouping, using both empirical data and a simulation study.

The results align with previous SEM research, indicating that not all indices are capable of detecting specification errors in the model.

Finally, we discuss the implications of this work and provide practical recommendations regarding advanced new methods for accurately tracking individual changes over time, and identifying the sources of variance that contribute to observed age-related changes. Also, we highlight the importance of using goodness-of-fit indices in applied research based on longitudinal designs.

Title

Regularised SEM-Based Out-of-Sample Predictions

Author(s)
Julian D. Karch [1] , Mark de Rooij [1] , Aditi M. Bhangale [1] , Zsuzsa Bakk [1]

[1] Leiden University

Abstract

Predictive modelling—which applies model parameters from one data sample to generate predicted values for new observations beyond that sample—can play a critical role in psychological research. Until recently, prediction mechanisms were limited to the traditional linear regression and machine learning frameworks. However, these approaches assume that psychological variables are measured without error, which is often not the case. Structural equation models (SEMs) do consider measurement error, and a prediction rule for SEMs with normally distributed, continuous data was recently proposed. Although the SEM-based prediction rule outperforms predictions based on (regularised) linear regression models in most cases, it is sensitive to model misspecification—specifically when additional direct effects between indicators on the predictor side and the latent response variable are included in the data-generating SEM. Regularising the SEM-based prediction rule—using methods like ridge regression or regularised discriminant analysis—can help address this issue. In this study, we propose using regularisation to achieve a data-driven compromise between a restricted SEM and a linear regression model fit to the same data, thereby producing regularised SEM-based out-of-sample predictions. The combination of regularisation and SEM indirectly accounts for the degree of model misspecification to produce more accurate and precise predictions by weighting the influence of the linear regression and the SEM. We hypothesise that the regularised SEM-based prediction rule will perform at least as well as the SEM-based prediction rule when the model is misspecified.

Title

Residual Dynamic Structural Equation Modeling for Analyzing Interindividual Variability in Intensive Data from Factorial Experiments

Author(s)
Benedikt Langenberg [1]

[1] Maastricht University

Abstract

Introduction. The collection of intensive longitudinal data in experimental factorial designs has become more common in social and behavioral sciences. This creates the need for more advanced analytical methods that can model complex within-person processes and between-person differences. Residual Dynamic Structural Equation Modeling (RDSEM) provides a strong framework by combining multilevel, time-series, and latent variable modeling, which is estimated using Bayesian methods. RDSEM allows for interindividual differences in custom contrasts, dynamic autoregressive processes, and explains interindividual differences interindividual variability by modeling within residual variance as a function of covariates, rendering RDSEM a useful method for analyzing experimental data. This presentation demonstrates the application of RDSEM to an intensive eye-tracking dataset, showing its advantages in analyzing fine-grained longitudinal experimental data.

Methods and Results. We used RDSEM to analyze eye-tracking data from a reading 2x3 within-subject experiment with children, investigating the development of reading efficiency in elementary school. The outcome was reading time at a word level. Custom contrasts were used to specify reading efficiency as differences in reading time between experimental conditions. We estimated and compared three models: ANOVA, which aggregated data within conditions to estimate average effects and interindividual variability in contrasts; LMM, which accounted for within- and between-person as well as time-varying covariates, but lacked interindividual differences in residual variances; and RDSEM, which included time-varying covariates such as landing position within words, modeled autoregressive effects between residuals, interindividual differences in autoregressive effects, and allowed within residual variance to be predicted by time-invariant covariates.
The results showed that RDSEM detected a significant autoregressive effect, meaning that viewing times for words were influenced by the difficulty of previous words. The inclusion of time-varying covariates showed that landing position was an important predictor of word reading time, confirming prior findings on skilled reading patterns. The model also showed that reading difficulties of children (RD) strongly predicted the within residual variance, meaning that children with RD had larger fluctuations in their residual variability over trials. Lastly, unlike LMM, RDSEM also made it possible to estimate random effects in residual variances and autoregressive effects, further explaining interindividual differences.

Discussion and Conclusion. The findings confirm that RDSEM is a valuable tool for modeling experimental factorial designs with intensive repeated measurements. RDSEM is particularly useful because it models dynamic within-person processes while also capturing interindividual differences in within residual variance and autoregressive effects. The ability to model within residual variance as a function of covariates provides a better understanding of variability in reading efficiency.

2.8 Symposium : ”Innovative Approaches to Address Validity”

Title

Building and Validating Culturally Responsive Assessments

Author(s)
Stephen Sireci [1] , Omaya Horton

[1] University of Massachusetts Amherst

Abstract

An exciting development in educational and psychological testing is culturally responsive assessment, which is assessment that is “mindful of student differences and employs assessment methods appropriate for different student groups”(Montenegro & Jankowski, 2017, p. 9). Although the call for culturally responsive assessment is strong, there are few examples of good practices in this area and even fewer examples of studies validating that assessments are truly culturally responsive. In this presentation, we briefly describe the goals of culturally responsive assessments and its design principles. Next, we describe studies that could be done to evaluate the degree to which assessments that strive to be culturally responsive are achieving that goal. We end with suggestions for future research and practice in this area, including suggestions for test development and validation.

Title

Mapping construct representations: Integrative approaches to Content Validity and Latent structures

Author(s)
Albert Sesé [1]

[1] University of the Balearic Islands

Abstract

This oral communication examines evidence of content validity in psychometrics, focusing on established procedures and potential complementary approaches. We review traditional validation methods, acknowledging their foundational importance. Our discussion emphasizes the need for isomorphic relationships between construct definitions and operational representations, a cornerstone of measurement theory. We explore how classical methods aim to achieve this isomorphism in capturing construct domains, considering both item content and latent structure. As a complementary approach, we introduce psychometric network analysis, suggesting its potential to offer additional insights into construct representation and item interrelationships. This method may enhance understanding of content coverage and construct isomorphism, working alongside established techniques. We propose an integrative approach that combines expert judgments, quantitative indices, and latent structure techniques to improve evidence of content validity and construct isomorphism.

By presenting a balanced examination of current practices and emerging perspectives, we aim to contribute to the ongoing methodological dialogue in measurement science, encouraging collaborative exploration of content validity assessment strategies. We discuss the implications of these integrative approaches for improving the accuracy and comprehensiveness of psychological measurements, highlighting their potential to refine our understanding of complex psychological constructs and their underlying structures.

Title

Collecting Validity Evidence Through the Measurement of Eye Movement

Author(s)
Elena Riol , Isabel Benítez Baena [1] , Patricia Román

[1] University of Granada. Mind, Brain and Behavior Research Center (CIMCYC)

Abstract

Item functioning is typically evaluated through pilot studies to identify problematic items and assess their performance. However, such analyses often fail to provide insights into the underlying causes of these problems. To address this gap, alternative strategies such as psychophysiological measures, including eye movements, may offer valuable insights into participants' response processes. This study investigates the potential of eye-tracking data to inform item functioning and provide validity evidence. Two studies were conducted. The first study had two phases: Phase 1 involved creating sentences with varying levels of legibility and examining corresponding eye movement patterns. In Phase 2, eye movements were analyzed in relation to three sentence features: syntactic complexity, lexical frequency, and sentence length. Results from both phases established criteria linking specific eye movement patterns to potentially problematic items. The second study compared two versions of an instrument administered to different groups. The first group responded to the original version, while the second group received a modified version with adjustments to reduce problematic elements.
Differences in responses between the groups demonstrated the extent to which eye-tracking data can guide the development of improved items. Conclusions include practical recommendations for researchers designing educational and psychological assessments. The study also highlights the utility of eye movements in providing validity evidence.

Title

Improving web probing method to obtain validity evidence of response processes by an AI generative app

Author(s)
David Sánchez Casasola [1] , Jose-Luis Padilla Garcia [2]

[1] Universidad de Granada; [2] University of Granada Faculty of Psychology: Universidad de Granada Facultad de Psicologia

Abstract

Practitioners and researchers use open-ended questions when designing survey questions to obtain validity evidence of response processes to survey items and questions. Data cleaning and response coding pose significant challenges, particularly for “web probes,”given the self-administered nature of “web probes”and the large number of participants compared to the smaller number of people interviewed in the cognitive interview method. The integration of generative AI, especially models based on GPT architectures, offers new opportunities to automate these processes efficiently and accurately. This proposal focuses on the development of a data post-processing solution for automated debugging and coding of responses to web probes. This solution could be implemented by using advanced prompting techniques, in an application of the GPTs store, a standardised data processing procedure with these AI-tools; or by an application that uses the OpenAI API to offer advanced features, depending on the results or performance of each option.
The objective of the paper is twofold: a) to present the development and validation of a generative AI-based data post-processing application and procedure that allows; and b) to illustrate how such a procedure, depending on how and in which model it is applied, deductively codes themes and sub-themes in the substantive responses, and automatically detects indicators of low involvement in the response process, such as mismatches and motivational losses.
Textual data from the CAS questionnaire on climate change anxiety and other questionnaires on “quality-of-life”will be categorized into substantive (1) and non-substantive (0) by coders. Subsequently, this coding will be compared with the coding generated by four different AI models (4th, 4th custom, O1, and Deepseek) and by the state-of-the-art OpenAI model using the API. A one-shot approach will be applied, calculating the correlation between manual and automated ratings. The application is expected to demonstrate high accuracy in response debugging and coding, significantly reducing the time and effort required in manual data processing.
This project aims to set a new standard in the automated processing of open-ended responses in psychometrics and contribute to fostering “web probing”to obtain validity evidence of response processes. Future developments of AI generative for improving validation of response processes “in vivo”will be also discussed.

2.9 Symposium : ”Measurement and Machine Learning”

Title

About the (Non-) Invariance of Sensing Data

Author(s)
Clemens Stachl , David Goretzko

Abstract

In recent years, psychological research has increasingly utilized novel (often digital) data sources. Sensing data, such as those collected from smartphones, enable researchers to monitor human behavior across diverse, ecologically valid contexts and extended periods with relative ease. These rich datasets offer great potential for predicting psychological traits, such as personality facets, through approaches like personality computing and machine learning. While previous research shows promising results, the quality and comparability of sensing data —both within and across studies —remain challenges. Smartphone sensing data, for example, are not only influenced by different preprocessing steps, but can also depend on the used hardware, the respective operating system and to some degree even on the version of a specific app. Consequently, measurements derived from sensing data may contain systematic biases unrelated to the intended behavioral constructs. To address these issues, this project adopts a measurement invariance perspective for analyzing sensing data. We adapt and apply methods from latent variable modeling to ensure comparability between data from different devices. Additionally, we explore potential biases introduced by “non-invariant”sensing variables and discuss their implications for subsequent statistical modeling.

Title

The Impact of Measurement Non-invariance in Target Variables on Machine Learning Prediction

Author(s)

David Goretzko , Eunsook Kim , Philipp Sterner [1]

[1] Ruhr University Bochum

Abstract

Psychology is increasingly interested in the prediction of psychological constructs via machine learning (ML) models, for example, predicting a person's personality or intelligence. To measure these psychological constructs, psychologists often draw on questionnaire data. In supervised ML, these measurements are then used as target variables (i.e., the "ground truth") for model training. Recently, Tay et al. (2022) introduced a conceptual framework that outlines various sources of bias throughout the ML modeling process. One potential bias is non-invariance of the questionnaire data across groups that is used as target values for supervised learning. As Tay and colleagues state, if the questionnaire used to collect the target data produces different expected scores between two groups with the same true score, this might bias the predictions of the final ML model. Specifically, two groups with the same underlying true score on the construct of interest might receive different predicted scores by the ML model. The goal of this work is to assess the actual impact of a lack of measurement invariance in target variables on the predictive performance of ML models. We address and investigate the impact of non-invariance in three different ways: empirically, semi-empirically, and simulation-based. We also discuss possible solutions to counter the impact of non-invariance in target variables.

Title

Addressing Measurement Error in Machine Learning-Assisted Social Science Modeling

Author(s)
Erik-Jan van Kesteren , Javier García Bernardo , Qixiang Fang [1]

[1] Utrecht University

Abstract

With the advent of machine learning tools and large language models (LLMs), the collection of measurements related to social science constructs (e.g., personality traits, political attitudes, human values) has become easier, faster and more affordable. These measurements are subsequently used for modelling of societal and group processes that social scientists typically engage in, where inferences from samples to populations are also made. Valid modelling and inferences, however, requires high-quality measurements or at the very least, methods to deal with the presence of measurement error. Just like traditional questionnaire-based measurements, machine learning- and LLM-based measurements have been shown to suffer from validity and reliability issues.

While there is an abundance of research literature in dealing with measurement error, they focus on questionnaire-based measurement error. It is unclear yet how measurement issues arising from machine learning tools and LLMs should be handled in social science modelling research.

This study has two primary objectives. First, we review existing literature to identify practices for addressing machine learning- and LLM-related measurement error, both in computer science and in social sciences.

Second, we synthesise these findings with existing measurement modelling literature to propose a practical framework for making valid inferences using machine learning- and LLM-based measurements in social sciences. By bridging the gap between modern machine prediction capabilities and social science inference requirements, our framework aims to enhance the reliability and validity of social science research outcomes in the era of machine learning and LLMs.

Title

A Machine Learning-Based Workflow for Model Evaluation and Revision in SEM

Author(s)
Melanie Viola Partsch [1] , David Goretzko

[1] Utrecht University

Abstract

Despite the popularity of structural equation modeling (SEM), investigating the fit of SEM models is still challenging—especially, if the global model fit evaluation implies non-negligible misfit, and researchers need to further investigate the type and severity of the misspecification in their model. Being overwhelmed by poorly fitting models, researchers sometimes strain the interpretation of their global model test (e.g., the χ2-test or model fit indices, such as the CFI and the RMSEA, in combination with cutoff values) and attest acceptable model fit, even though they would be well advised to reject or revise their model. To counteract this questionable research practice, we developed a method that guides researchers through a more thorough process of model fit evaluation and, if necessary, revision.

Based on a proof-of-concept study, in which we have previously shown that a pre-trained machine learning (ML) model can detect misfit in multifactorial measurement models with a high accuracy, we developed an automated ML-based workflow for SEM evaluation and revision. This workflow involves several ML models that we trained based on a maximum of 173 model and data features extracted from more than 1 million simulated data sets and multifactorial models fitted by means of confirmatory factor analysis. In the first step of the workflow, the researcher's model is classified as either (a) correctly specified or misspecified by neglecting (b) a factor, (c) factor correlations, (d) cross-loadings, or (e) residual correlations. For classes a–c, we, in summary, give the following recommendations: (a) accept the model, (b) reject the model and revise the underlying theory or operationalization, (c) free the factor correlations, if willing to lift orthogonality constraints, or revise model by including method factor(s). For classes d–e, the second step of the workflow is initiated that determines the number of cross-loadings or residual correlations. Based on the severity of the misspecification, we, in summary, recommend the following: In case of a mild misspecification, researchers might freely estimate the concerned parameter(s), scrutinize their operationalization to understand the misspecification, and cross-validate it based on new data. In case of a moderate misspecification, researchers might revise their operationalization. In case of severe misspecification, researchers might reject the model and revise the underlying theory.

While this ML-based workflow for SEM evaluation and revision is not without limitations (e.g., it cannot identify a mix of misspecifications, it is only applicable for multifactorial measurement models so far), it provides applied researchers with unprecedented guidance in the complex, often iterative process of measurement and theory development, thereby hopefully encouraging them to face up to model misfit instead of neglecting it.

Keywords: Structural Equation Modeling (SEM), Latent Measurement Models, Model Misspecifications, Model Fit Evaluation, Model Revision, Machine Learning

2.10 Poster Session 3

Title

Dilemmas and Methodological Challenges in Research on Older Adults

Author(s)
Jakub Fabiś [1]

[1] Uniwersytet Łódzki

Abstract

Social research on older adults constitutes a significant area of exploration in the context of aging societies, yet its implementation is accompanied by numerous methodological challenges. This poster presentation focuses on analyzing key dilemmas faced by researchers in this field.

The discussion will address ethical aspects of research, including the protection of participants' privacy, obtaining informed consent, and accounting for potential barriers arising from physical or cognitive limitations of older individuals. Additionally, challenges related to sampling, particularly in terms of representativeness and access to hard-to-reach groups, will be examined.

The session will also explore the difficulties associated with designing research tools tailored to the specific characteristics of this population and with interpreting research findings in a way that avoids stereotyping.

Examples of good practices that can contribute to enhancing the quality of social research on older adults will be presented.

Title

Assessing Fear of Heterosexism: Spanish Adaptation and Validation of a Psychological Measure

Author(s)

Laura Vozmediano [1] , Jone Aliri [1] , Maite Azabal [1] , Alexander Trinidad [2]

[1] University of the Basque Country UPV/EHU; [2] University of Cologne

Abstract

Heterosexism is a form of socioculturally located violence that affects gender and sexually diverse communities, resulting in widespread discrimination. The Fear of Heterosexism Scale (Fox & Asquit, 2015) is a self-report measure consisting of 15 items designed to assess the fear of heterosexism experienced by these communities. Although the scale has established psychometric properties in English, no Spanish version currently exists. This study performed the cross-cultural adaptation of the Fear of Heterosexism Scale for the Spanish population and examined its psychometric properties, including reliability, structural validity, and concurrent validity. The adaptation process followed the guidelines outlined by the International Test Commission, involving translation, back-translation, and expert review. The sample consisted of young adults from Spain who self-identify as non-heterosexist, with data collected at baseline, alongside measures of general fear of crime, anxiety, depression, stress, experiences of heterosexist violence, level of identity concealment, and connections to queer or general communities. The analysis focused on evaluating the internal consistency and factorial structure of the Spanish version, as well as its relationships with the other measures to assess concurrent validity. The results suggest that the Spanish version of the scale demonstrates adequate reliability and a factorial structure that aligns with expectations. Furthermore, significant correlations with the other measures support its concurrent validity. These findings indicate that the Spanish version of the Fear of Heterosexism Scale is a promising tool for measuring fear of heterosexism in Spanish-speaking populations. This adaptation fills a critical gap in the measurement of heterosexism within Spanish-speaking communities and contributes to a better understanding of its psychological impact. Furthermore, the refined scale offers valuable insights for research on the consequences of heterosexism and informs clinical interventions aimed at supporting gender and sexually diverse populations. By improving the scale's wording and validating its psychometric properties, this research advances the study and intervention of heterosexism in diverse cultural contexts.

Title

Employing item response theory and factor analysis to purify latent trait estimates

Author(s)
Ján Pavlech [2] , Patrícia Martinková [1]

[1] Institute of Computer Science of the Czech Academy of Sciences; Faculty of Education, Charles University; [2] Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

Abstract

The binary factor analysis (FA) model and the 2-parameter item response theory (IRT) model impose different structures and assumptions, but the models are well known to be equivalent (Martinkov´a & Hladk´a, 2023, Section 7.5). In this work, we investigate a generalization of the FA model to three- and four-parameters using the mixture-dichotomized model, and its relationship with the three- and four-parameter IRT model. We then focus on the estimation of the ability scores cleaned of guessing and inattention in the context of educational assessment, or of pretending and dissimulation in the context of psychological assessment, which we term the NGI (non-guessing and non-inattention) scores. We propose a Bayesian estimation method, and compare it with other alternatives. We discuss the advantages of NGI scores over total scores and other estimates of latent scores, as well as the strengths and limitations of various estimation methods. The methods are illustrated using real data examples from well-being and educational assessment.

Title

An Exploration of R-squared Effect Size Measures in Mediation

Author(s)

Alexander Miles [1] , Amanda Fairchild [2]

[1] University of Southern California; [2] University of South Carolina

Abstract

Effect size (ES) measures offer insight into the magnitude of an effect. It has been 26 years since the 1999 APA Task Force on Statistical Inference recommended reporting ES (and complementary metrics like confidence intervals, CIs) alongside results of significance tests, and the availability of routine ES measures for a variety of statistical models has increased their uptake in the literature. ES options for some multivariate models remain understudied, however. One such model, mediation analysis, has been increasingly used over the past 25 years to understand the potential pathway(s) through which a predictor affects an outcome. Despite widespread application of the mediation model, less progress has been made in developing suitable mediation ES measures. Among the measures available, R-squared measures are particularly attractive - as they give researchers an easily interpretable portion of variance explained in the outcome by the indirect effect. At least three R-squared measures for mediation have been proposed, with each measure having unique advantages and limitations. Only one measure has offered associated CIs. As such, we develop and evaluate analogous CIs for the other two measures and then compare the widths and coverage of each. We additionally compare expected values for the sample estimators and relative percent bias under a variety of parameter combinations, shedding light on the strengths and weaknesses of each measure in different scenarios to inform their application.

Title

Psychometric Properties of the Interpersonal Rejection Sensitivity Scale (IRSS) in a Spanish Sample: Factor Structure, Reliability, and Measurement Invariance

Author(s)
Irene Caro Cañizares [1] , María José González Calderón [1] , Garazi Laseca Zaballa [1] , Desirée Blázquez-Rincón [1] , María Elena Brenlla [1] , Eva Izquierdo-Sotorrío [1] , Alba Lirón León [2]

[1] Universidad a Distancia de Madrid; [2] Universidad Autónoma de Madrid

Abstract

The Interpersonal Rejection Sensitivity Scale (IRSS) is an 11-item scale aimed to measure sensitivity to rejection in interpersonal relationships based on the Interpersonal Acceptance-Rejection Theory. In a multicultural study, published in 2023 and involving 3,083 participants from eight countries, evidence of partial measurement invariance of the IRSS, indicating the existence of some cultural differences in the interpretation of the construct. However, the scale has never been validated in the Spanish population. Therefore, the aim of the present study was to assess the psychometric properties of the IRSS in the Spanish sample of 388 people who took part in the multicultural study. First, its internal structure was analyzed by means of exploratory factor analysis in the randomly selected first half of the sample. After this, the fit of the model suggested in the previous analyses was studied in a confirmatory factor analysis with the second half of the sample. From these results, the reliability index most suitable for the internal structure of the IRSS was computed. Likewise, the convergent and criterion validity were also studied by analyzing the degree of correlation between the IRSS scores and those of other theoretically relevant scales. Finally, data are provided on the measurement invariance (configural, metric and scalar) achieved between those who responded online and those who responded on paper.

Title

Reevaluating Factor Analysis: Violations of the Reflective Measurement Model and a Proposed Correction for Item Selection Bias.

Author(s)

Miguel Bosch Francisco

Abstract

Factor analysis is widely used in psychological measurement under the assumption that it adheres to a reflective measurement model, where observed variables are manifestations of an underlying latent construct. This approach aligns with a scientific realist perspective, where factors represent independent causal structures. However, we demonstrate that standard factor analytic procedures violate key assumptions of reflectivity, particularly due to item selection biases that impose a range restriction on factor loadings. This restriction inflates factor variance estimates, distorts latent correlations, and introduces systematic bias in fit indices.

Through theoretical derivation and extensive simulations, we show that selecting items based on their observed loadings alters the empirical distribution of factor loadings, leading to overestimation of factor strength and underestimation of error variance. Furthermore, this bias propagates into structural analyses, affecting the validity of latent correlations and goodness-of-fit statistics. Our work extends previous discussions on the impact of item selection, but crucially distinguishes range restriction at the variable level from traditional sample-level restrictions.

To address this issue, we propose a novel correction method that adjusts for the distributional distortion of factor loadings, preserving the intended properties of the reflective measurement model. We compare our approach with existing corrections, including SRMR-unbiased adjusted by communalities, and evaluate its effectiveness across various sample sizes and item selection criteria.

Beyond its technical implications, this study raises fundamental epistemological concerns about the alignment between latent variable modeling and the philosophical assumptions of psychological measurement. By refining factor analytic methods, we enhance both their theoretical coherence and practical utility.

Title

Validation of the Multidimensional Scale of Perceived Social Support (MSPSS) in Mental Health Service Users

Author(s)
Ángela I. Berrío [1] , Georgina Guilera [1] , Maite Barrios [1] , Hernán M. Sampietro [2]

[1] University of Barcelona; [2] ActivaMent Catalunya Associació

Abstract

Perceived social support is a key factor in the recovery and well-being of individuals with mental health disorders. The Multidimensional Scale of Perceived Social Support (MSPSS) is a widely used instrument to measure social support, but its psychometric properties in clinical populations in Spain have not been explored. This study evaluates the factorial structure, internal consistency, measurement invariance, and convergent evidence of the MSPSS in a sample of community mental health service users.

A sample of 345 adults (M = 46.6 years, SD = 11.9), users of community mental health rehabilitation services (55.4% men), diagnosed with a severe mental disorder, participated in the study. The Spanish version of the MSPSS was administered alongside the Maryland Assessment of Recovery Scale (MARS-12) and the Dispositional Hope Scale (DHS) to gather convergent evidence. Confirmatory factor analyses were conducted to examine the scale's dimensionality and test factor invariance across gender. Additionally, reliability was assessed using McDonald's omega (ω) and Cronbach's alpha (α).

Factor analysis confirmed the three-factor structure of the MSPSS (significant others, family support, and friend support), with factor loadings above .83 and a good model fit. Reliability was excellent for both the total scale (ω = .92, α = .92) and its subscales. Factor invariance across gender groups was confirmed, allowing for valid comparisons. Additionally, the MSPSS showed significant correlations with the MARS-12 ($r = .44$, $p < .001$) and DHS ($r = .44$, $p < .001$), providing convergent evidence.

The findings support the MSPSS scores as valid and reliable for assessing perceived social support in a clinical sample of individuals with mental health diagnoses in Spain. Their application can help identify support needs and enhance mental health care.

Title

Development and Content Validation of the Advance Care Planning Attitudes Scale for Mental Health: Incorporating Mental Health Service Users'Perspectives

Author(s)
Hernán Sampietro [1] , Chao Zhang [1] , Georgina Guilera [1] , Maite Barrios [1]

[1] University of Barcelona

Abstract

Advance care planning (ACP) is increasingly recognized as a vital component in medical care due to its emphasis on individuals'autonomy and dignity across diverse health conditions, including mental health. Prior research has produced numerous measurement instruments to assess ACP-related aspects, with attitudes towards ACP remaining the most frequently measured construct. However, existing tools are not tailored to the context of mental health. Therefore, this study aimed to develop a Spanish-language scale assessing attitudes towards ACP for mental health and gather content-based validity evidence incorporating mental health service users'perspectives.

Guided by the Theory of Planned Behavior, preliminary items were drafted to align with four theoretical domains measured using a 5-point Likert scale: attitudes towards ACP, subjective norms, perceived behavioral control, and intention to engage in ACP. Three rounds of review and refinement were conducted by four researchers, one of whom was a mental health service user. This process yielded an initial pool of 42 items for further analysis. The readability of the items was assessed using a readability tool. The 37 items that met the readability criteria were then evaluated for relevance and clarity by an expert panel via Qualtrics. Content validity indices were computed, including item-level (I-CVI) and scale-level (S-CVI). Values were considered adequate if ≥ .78 for I-CVI and ≥ .90 for S-CVI.

Readability scores ranged from 61.99 to 102.83, indicating that the language used in the instrument is generally accessible. Feedback was received from 13 experts (81.25% participation rate), including 4 users. I-CVI scores ranged from .69 to 1 for relevance, with three items considered candidates for revision or deletion, and from .77 to 1 for clarity, with two items prompting revision or deletion. S-CVI scores were .93 and .94 for relevance and clarity, respectively, indicating excellent content validity. Based on these results, the Advance Care Planning Attitudes Scale for Mental Health is proposed.

This scale incorporates the mental health service users' perspective from its development to its content validation and addresses the gap in attitudes assessment regarding ACP in mental health. By understanding these attitudes, more effective interventions can be designed to make the implementation of ACP a reality, empowering individuals to take an active role in shaping their future care.

Title

Psychometric properties of Multidimensional Scale of Perceived Social Support in Spanish university students

Author(s)
Júlia Gisbert-Pérez [1] , Manuel Martí-Vilar [1] , Laura Badenes-Ribera [1] , Elena Cejalvo [1] , Laura Galiana [1]

[1] Universitat de València

Abstract
Perceived social support (PSS) refers to the individual’s perception of the availability and quality of emotional, instrumental, and social support that he/she can receive from his/her network of relationships. PSS is linked to physical and psychological well-being. The Multidimensional Scale of Perceived Social Support (MSPSS) is one of the most used instruments to measure PSS. The MSPSS includes 12 items that measure three dimensions of PSS: support from family, peers, and other significant. This study aimed to analyze the psychometric properties of the MSPSS in Spanish university students. The sample included 795 university students (M = 21.5 years; SD = 5.01; 68% male). Confirmatory factor analysis (CFA) and internal consistency were assessed. Results of CFA supported the factorial validity of the MSPSS including 1 general factor and 3 specific factors (χ2(51)111.139; TLI = .999, RMSEA = .039, CFI = .999, SRMR = .017). Cronbach’s alpha and McDonald’s omega of the total scale were .88 and .85. Regarding subscales, Cronbach’s alpha and McDonald’s omega were family support (.92, .92), peer support (.92, .82), and other significant support (.97, .97), respectively. The MSPSS provides valid and reliable scores for measuring perceived social support in the Spanish university population.

Title

A multilevel Ornstein–Uhlenbeck process with crossed random effects for multivariate time series

Author(s)
José Ángel Martínez-Huertas [1] , Emilio Ferrer [2]

[1] UNED; [2] University of California, Davis

Abstract

Stochastic differential equation (SDE) models are very useful for analyzing time series. In Psychology, some SDE models have been used for the study of affect dynamics. For example, the Ornstein–Uhlenbeck (OU) process is a simple but very interesting model for this purpose. In this context, some authors have extended that SDE model to a multilevel framework using Bayesian estimations to study individual differences in their parameters. Here, we propose the OU process, a SDE model, to analyze multivariate time series by means of crossed random effects for individuals and variables. Our extension allows to estimate the variability of different parameters of the process, such as the mean (μ) or the drift parameter (φ), across individuals and variables of the multivariate system using a Bayesian framework. In this presentation, we illustrate the estimations and the interpretability of the parameters of this multilevel OU process in an empirical study of affect dynamics, and also conduct a simulation study to evaluate whether the model can recover the population parameters generating the OU process. Our results support the use of this model. Thus, we conclude that it can be a useful approach to analyze multivariate time series of affect dynamics.

Title

Type I error of repeated measures ANOVA with non-sphericity and very extreme deviation from normality

Author(s)

F. Javier García-Castro [1] , Rafael Alarcón [2] , María J. Blanca [2] , Roser Bono [3] , Jaume Arnau [3]

[1] Universidad Loyola Andalucía; [2] University of Malaga; [3] University of Barcelona

Abstract

Background. Recent studies have shown that repeated measures analysis of variance (RM-ANOVA) is generally robust to violation of normality provided the sphericity assumption is fulfilled. However, violation of sphericity has an important impact in terms of Type I error. In this scenario, the Greenhouse-Geisser (F-GG) and Huynh-Feldt (F-HF) adjustments have been widely used as alternatives to the F-statistic. However, the performance of both F-GG and F-HF remains unclear when sphericity is violated under very extreme violation of normality. Objective. The aim of this study was to analyse the performance of the F-statistic, F-GG and F-HF in terms of Type I error, with designs including three repeated measures, very extreme violation of normality (i.e. $\gamma 1 = 3$, $\gamma 2 = 21$), epsilon values ranging from the lower to its upper limit (from .50 to 1), and a wide range of sample sizes (from 10 to 300). Method. Monte Carlo simulation was performed, with results being interpreted according to Bradley's liberal criterion. Results. F-GG and F-HF are generally robust when normality is violated, provided that there is no extreme violation of sphericity (i.e. epsilon values ≤ .60). In this case, their robustness depends on the sample size, and they are liberal with small sample sizes. Conclusions. The more severe the violation of both normality and sphericity, the larger the sample size needed to achieve robustness of F-GG and F-HF. Further studies with a larger number of repeated measures are needed to analyse robustness of these statistics with extreme violation of both normality and sphericity. This research was supported by grant PID2020-113191GB-I00 from the MCIN/AEI/10.13039/501100011033.

Title

Assessing false recognition with ad hoc categorical, associative and taxonomic lists.

Author(s)
Verónica Benítez , Maria A. Alonso [1]

[1] Instituto Universitario de Neurociencia (IUNE)

Abstract

The DRM paradigm for studying false memories involves presenting lists of words that are semantically related to a critical non-presented word (CNW). Numerous experiments have demonstrated robust rates of false recall and false recognition of the CNW under various conditions and with different types of relationships between the presented words and the CNW (associative, taxonomic, phonological, etc.). However, research using lists based on ad hoc categorical relationships is very limited. Ad hoc categories are spontaneously generated based on a specific goal within a given context and group elements from different taxonomic categories (e.g., "things people take to a wedding"). To explore the contribution of different types of semantic relationships (ad hoc categorical, associative, and taxonomic) to false recognition, 70 CNWs were selected, and 210 lists were constructed (70 for each type of relationship). A total of 365 participants studied lists corresponding to each type of relationship and subsequently completed a recognition test. The d'prime parameter (signal detection theory) was used to measure accuracy in identifying presented words by comparing the hit rate with the rate of unrelated false alarms. An analysis of variance was then conducted. The results showed that correct recognition did not significantly differ among the three list types. Regarding false recognition, associative lists exhibited a significantly higher level of false recognition than both taxonomic and ad hoc lists. However, no significant differences were found between ad hoc categorical and taxonomic lists. These findings are discussed in relation to current theories of false memory and contribute to the understanding of the nature of semantic representation and the mechanisms underlying memory distortions.

Title

Evaluating assumptions on variance magnitude in repeated measures designs: an empirical evaluation

Author(s)

José A. López-López [1] , Julio Sánchez-Meca [1] , María Rubio-Aparicio [1] , Manuel J. Albaladejo-Sánchez [1] , Fulgencio Marín-Martínez [1] , Juan J. López-García [1]

[1] University of Murcia (Spain)

Abstract

In the literature on repeated measures designs, it is common to find assumptions about the magnitude of pre-test and post-test score variances. Some researchers argue for the homoskedasticity of both variances, justifying the use of the pooled standard deviation as the standardizer in effect size calculation. In contrast, others contend that post-test variance should be greater due to the influence of the treatment, advocating for standardization based solely on pre-test score variance. We provide an empirical evaluation of these assumptions using a database of primary studies from various meta-analyses in the field of clinical psychology. Specifically, we calculated the percentage of studies where post-test score variance exceeds pre-test score variance, and vice versa. Additionally, we assessed the proportion of studies in which these differences are statistically significant. Finally, to assess the practical relevance of this decision, we estimated the combined standardized mean change for each meta-analysis, comparing the results using the pre-test scores standard deviation versus the pooled standard deviation, and we compared the results. We discuss the implications of the results of the study.

Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

Title

Differential item functioning in online health-information seeking measurement

Author(s)
Patrícia Martinková [2] , Michaela Cichrová [3] , Petra Raudenská [1]

[1] Institute of Sociology of the Czech Academy of Sciences; [2] Institute of Computer Science of the Czech Academy of Sciences; Faculty of Education, Charles University; [3] Institute of Computer Science, Czech Academy of Sciences; Faculty of Mathematics and Physics, Charles University

Abstract

As digital health information becomes increasingly important in personal and public health decision-making, measuring internet usage for health-related purposes is crucial for understanding global digital health trends, which can contribute to designing policies and web-based tools. The International Social Survey Programme (ISSP) introduced a standardized questionnaire to assess online health information-seeking behaviour and collected data across diverse populations in 30 countries. Detailed item-level analysis, including detection of differential item functioning, may provide granular information on between-group differences across diverse populations.

In our work, we focus on testing differential item functioning (DIF) with respect to the country and age of the respondents to assess whether the questionnaire measures online health information-seeking behaviour consistently across different ages and nationalities and to better understand differences among respondents. If DIF is present, individuals with the same level of engagement but from various countries or age brackets may respond differently to certain items, which may remain unnoticed if the comparison is conducted based on total scores only.

Traditional DIF detection methods compare a fixed number of demographic groups, meaning that respondents would need to be divided into a fixed amount of age cohorts. However, such a grouping can lead to information loss and may fail to detect more complex DIF patterns. To address this, we introduce a new DIF detection method utilizing varying coefficient regression models, which allows for a more nuanced analysis of age-based DIF. These models enable us to test the statistical significance of DIF using submodel tests and to assess the practical significance of DIF by using area-based effect size measures. We apply our newly proposed method to the dataset and compare the results with those obtained by grouping age into cohorts. We also discuss differences with respect to countries.

Title

Effect of Class Imbalance and Multicollinearity on Parameter Estimation in Binary Logistic Regression

Author(s)
María Lucía Feo-Serrato [1] , Clara Cuevas Ureña , Ricardo Olmos Albacete [2]

[1] Complutense University of Madrid; [2] Autonomus University of Madrid

Abstract

Class imbalance—or, more broadly, rare events data—presents a common challenge in binary logistic regression (BLR), particularly in contexts where the phenomenon of interest has low prevalence (e.g., rare diseases or risk of abuse). The likelihood function employed in BLR tends to underestimate the probability of rare events (Oommen et al., 2011). Moreover, rare events exert a disproportionate influence on the variance-covariance matrix used to compute the standard errors of the estimated coefficients; consequently, any error in estimating the probabilities of these events can be amplified within this matrix, ultimately compromising the precision of the coefficient estimates (King and Zeng, 2001).

The primary objective of this study is to analyze the effects of class imbalance and multicollinearity on parameter estimation in binary logistic regression, and to assess which techniques might mitigate their impact, thereby yielding more consistent and efficient model estimates.

Monte Carlo simulations were employed to examine the influence of class imbalance (ranging from 50:50 to 95:5), multicollinearity (none, moderate, and high), and sample size (from 250 to 10,000) on parameter recovery. The estimation methods compared in the study included BLR, Ridge, LASSO, Elastic Net, and SMOTE, with performance evaluated in terms of average bias, standard error (SE), and root mean square error (RMSE).

The results indicate that BLR maintains moderate bias under conditions of low or moderate imbalance (≥70:30) and large sample sizes (>1,000). However, under extreme imbalance (95:5), both SE and RMSE increase significantly. Ridge and Elastic Net demonstrated greater stability in scenarios characterized by pronounced imbalances and small sample sizes, whereas SMOTE exhibited high variability and substantial bias in cases of severe imbalance.

From a theoretical perspective, this study contributes to a nuanced understanding of the impact of class imbalance in explanatory contexts, which are prevalent in psychology and other behavioral sciences. Moreover, it underscores the importance of selecting appropriate methods tailored to the specific conditions of the data.

Specifically, regularization methods—particularly Ridge and Elastic Net—emerge as promising tools for managing imbalances in practical applications. These methods are especially valuable in contexts where the stability of estimates is paramount, such as in public health studies or security risk analyses. Nevertheless, their application should be accompanied by rigorous evaluations using key metrics such as RMSE, SE, and probabilistic calibration indices.

In summary, while BLR proves effective under moderate imbalances and large samples, Ridge and Elastic Net are preferable in more complex scenarios. Validating these findings with real-world data and exploring advanced approaches, such as Bayesian methods, is recommended to further enhance the reliability of explanatory models.

Title

Trends in Reliability Induction Practices in Neuropsychology: A Systematic Review

Author(s)
Raimundo Aguayo-Estremera [1] , Ángeles Correas [1] , Julio Sánchez-Meca [2] , Eva Ruiz [1] , Emilio Verche [1]

[1] Complutense University (Spain); [2] University of Murcia (Spain)

Abstract

Reliability is an important property that all psychological measurement instruments must demonstrate. Researchers should report the reliability of tests based on the scores obtained from their own samples. However, in many cases, they engage in a malpractice known as reliability induction, either by failing to report it (by omission) or by not reporting it with their own data (by report). Induction by report can further be divided into induction by vague reporting, when it is simply stated that the test's reliability was appropriate in a previous study, and induction by precise reporting, when a specific value or range of values is provided.
In the present study, we aim to examine the trends in reliability reporting in the field of neuropsychology, taking into account the different types of reliability induction. A systematic review was conducted of studies that used standardized tests and were published over the last 25 years in neuropsychology journals. Specifically, studies published in the two journals with the highest impact factor within quartiles 1, 2, and 3 of the Journal Citation Report were selected. From the year 2000 to 2024, five articles per year were randomly selected from each journal among all issues published that year.
Since methodological studies with recommendations on the need to report reliability indices using researchers' own samples have been published over time, we expect the rate of reliability induction to be lower in recent years than in earlier years. Furthermore, based on findings from previous meta-analyses on reliability generalization, we anticipate a higher rate of reliability induction by omission than by report. The findings will contribute to the ongoing debate on results replicability in neuropsychology.

Title

How to use observable responses and hidden states of recurrent neural networks to reason about cognitive aspects of language?

Author(s)
Alejandro Martínez-Mingo [1] , José Ángel Martínez-Huertas [2] , Guillermo Jorge-Botana [3]

[1] IE; [2] UNED; [3] UCM

Abstract

The proposal of formal models to study psychological processes underlying language comprehension and production has long been a central concern in Cognitive Science. One of the main aims of this research line is to uncover the internal representations involved in language processing and the operations that transform these representations within contextual constraints because context and representation dynamically interact to create meaning. Neural network models have provided formal mechanisms for this purpose. These computational models allow for a precise characterization of the representations and operations hypothesized in cognitive theories. Among them, recurrent neural networks (RNNs) with long short-term memory (LSTM) mechanisms have emerged as a particularly useful tool to model the sequential contextual dependencies of language. Here, we illustrate the formalization of expectation shifts through an RNN model and align it with insights from experimental studies. To do so, we present a mental experiment analyzing both the external expressions (model outputs) and internal representations (hidden states) of RNNs.

Title

Evaluation of competence in dementia care among formal caregivers: Adaptation of the Sense of Competence in Dementia Care Staff scale (SCIDS) to Spanish

Author(s)
Arantxa Gorostiaga [1] , Igone Etxeberria [1] , Fátima García-Pena [1]

[1] University of the Basque Country UPV/EHU

Abstract

The evaluation of formal caregivers' ability to manage psychological and behavioral symptoms, as well as other problematic behaviors associated with dementia, is of great relevance for several reasons. On one hand, a lack of knowledge about dementia and the inability to handle these symptoms add additional stress to the burden these caregivers may experience. Moreover, assessing this competence is essential for identifying training needs and designing tailored interventions targeted at this population. Among the available instruments for this purpose, the Sense of Competence in Dementia Care Staff (SCIDS) scale developed by Schepers et al. (2012) demonstrates good psychometric properties and is well-suited for measuring this construct, as it assesses perceived competence in dementia care and symptom management. However, this instrument is not yet available in Spanish. This study presents the first findings regarding the psychometric properties of the Spanish version of the SCIDS. The scale's translation followed a back-translation design, involving four experts in dementia care and psychometric aspects related to item development. In the study's initial phase, cognitive interviews were conducted with 15 individuals from the target population to evaluate its validity evidence based on response processes. After that, a pilot phase was carried out with a sample of 50 participants from the target population to analyze the initial functioning of the items. Finally, the version obtained from this phase was administered, along with the necessary validation battery, to a sample of 247 formal dementia caregivers (93.1% women), aged between 18 and 66 years (M = 46.73; SD = 11.92). Regarding validity evidence based on internal structure, two models were tested: a unidimensional structure and a four-dimensional structure (Professionalism, Relationship Establishment, Care Challenges, and Adaptation to the Person), replicating the instrument's original structure. The best-fitting model was the four-dimensional structure (CFI = .949; TLI = .938). These dimensions demonstrated acceptable internal consistency (Cronbach's alpha = .77, .76, .83, and .73, respectively). The initial results from the adaptation process of the SCIDS into the Spanish language and culture indicate that it possesses adequate psychometric properties. In the coming months, the study aims to expand the sample and provide additional validity evidence for the scale.

Title

Latent Trait-State-Occasions models to analyze state and trait components of Loneliness in European adults and their associations with social contact

Author(s)

Laura Galiana [1] , José M. Tomás [1] , Patricia Sancho [1]

[1] Universitat de València

Abstract

Introduction: Loneliness may be considered a risk factor for overall health, especially in the old age, and it is defined as a subjective distressing experience that results from perceived isolation or inadequate meaningful connections (Prohaska et al., 2020). Loneliness is associated with a greater risk of health problems, similar in terms of mortality to that caused by smoking up to 15 cigarettes a day (Holt-Lunstad et al., 2017). Loneliness is monitored in virtually all international and national surveys of adults and older adults, such as ELSA, SHARE, HRS, MIDUS, JAGES, or CLSA, and this opens the possibility to analyze its variability and change through the aging process. Specifically, it could be of great interest to address how much of loneliness in the old age reflects stable trait variance as well as more labile (state) variance (Fleeson & Noftle, 2009). The aims of this work are: a) to estimate the amount of trait vs. occasion (state) variability in loneliness; and b) to accommodate covariates related to social contact to stablish their effects on both trait and occasion latent variables.
Method: The data belong to waves 6, 7 and 8 of the SHARE longitudinal study on European older adults. For the purposes of this study, only participants at least 60 years old at wave 6 were selected. Their mean age at the first wave was 71.72 years (SD= 8.27), and 55% were women. The variables of interest are the short version (three indicators) of loneliness UCLA-R Loneliness Scale and two indicators of social contact for receiving or giving help. The models estimated were Trait-State-Occasion (TSO) models that partitions variance of loneliness into trait, occasion-specific and autoregressive.
Results: The estimated TSO model had an extremely good fit to the data (chi-square= 83.019, df= 21, p< .001; RMSEA= .008, 90% CI [.006, .009]; CFI= .999; SRMR= .012). Trait variance was slightly lower than 60%, while occasion-specific variance was around 40%. Autoregressive effects explained a relatively low percentage of the variance in each wave of loneliness (2.3 to 2.6%). The TSO model with covariates also fitted the data extremely well. The associations of the covariates were significantly associated with the loneliness trait. However, the associations with the occasion factors were much lower, and in general non-significant.
Conclusions: Longitudinal data on loneliness in the old age showed that an important part of its variance is trait-like or stable, which should be accounted for when longitudinal explanatory models are performed. Results also showed that covariates may have a greater association with stable variance than with occasion specific variance.

Title

Purifying the ability from external variables.

Author(s)
Daniil Talov [1]

[1] HSE University

Abstract

In social sciences, measuring constructs and evaluating their relationships with other variables is often complicated by external factors. These external variables can bias estimates and provide alternative explanations for results. This creates a demand for quantitative methods that correct estimates of the relationships between variables. The aim of this study is to investigate methods for purifying the target ability, measured by the target test, from external variables. Three methods are used for purifying abilities: (1) linear regression, (2) orthogonal and (3) oblique bifactor models. To investigate the functioning of these methods, we conduct a simulation study. The simulations demonstrate that it is possible to purify the target ability in two ways: by completely removing (via regression and orthogonal bifactor models) or by partially retaining (via oblique bifactor models) the confounding variance. We also provide a real-data example: the assessment of mathematical literacy in the first grade of elementary school. At this age, children experience difficulties with reading, so the tasks have to be voiced. In this case, phonological literacy interferes with the success of solving items. The differences in interpretation of these methods are discussed in application to the real-data example.

Title

Coping Flexibility Among Young Adults from Six Countries: Cross-Cultural Validation of the Coping Flexibility Scale (CFS)

Author(s)

Manuel Sánchez García [1] , Fermín Fernández Calderón [1] , Adrian J. Bravo [2] ,
Bella M. González Ponce [3]

[1] Departamento de Psicología Clínica y Experimental, Universidad de Huelva; [2] University William & Mary (USA); [3] Universidad de Extremadura

Abstract

Youth is a period of significant change in which contextual demands can lead to maladaptation and impaired personal and social functioning. An adequate repertoire of coping strategies during this critical life stage is essential for achieving personal and social adjustment. In this regard, cognitive flexibility—operationalized as the ability to adapt coping strategies in response to changing environmental demands (Cheng et al., 2014; Kato et al., 2012)—is strongly associated with better psychological and social adjustment.

The Coping Flexibility Scale (CFS; Kato, 2012) is one of the most widely used instruments for assessing coping flexibility. It comprises 10 items divided into two dimensions: evaluative coping (five items) and adaptive coping (five items). While this instrument has demonstrated evidence of adequate psychometric properties, some studies have identified issues related to its structural validity and the performance of specific items (Janicka, 2015; Soltys et al., 2015).

The present study aims to cross-culturally test the psychometric properties of the CFS in its Spanish and English versions. To do this, we conducted two different studies. In Study 1, we accessed 3753 university students from six countries (USA, Canada, Spain, Argentina, England and South Africa). We conducted confirmatory factor analyses to test four models. Our results supported an 8-item unidimensional measure (CFI = .992; TLI = .989, RMSEA = .117, Cronbach alpha = .89; items 2 and 7 were removed), which was invariant across the six countries. Cronbach´s Alpha ranged between .79 and .90. Moreover, this model showed evidence of validity according to the relationships between the CFS scores and other variables such as depression, anxiety and emotion regulation. In study 2, we used targeted sampling procedure to recruit a community sample of 612 cannabis-using young adults in Spain. In terms of internal structure, the results replicated those of Study 1, this is, they supported an 8-item unidimensional measure in which items 2 and 7 were removed (CFI = .994; TLI = .992, RMSEA = .066, Cronbach alpha = .83). Evidence of validity according to the relationships between the CFS scores and other related measures were also provided. In particular, CFS scores were negatively associated with metal health measures (anxiety, depression, stress), and problematic cannabis use.

Our results support the use of the 8-item unidimensional measure of the Coping Flexibility Scales across Spanish and English-speaking countries. Also, its use with different young-adults subpopulations (students-non students, and people using and not using cannabis) is supported.

This work was funded by the Agencia Estatal de Investigación (Ministerio de Ciencia, Innovación y Universidades, Spain; MICIU/AEI//10.13039/501100011033/), grant number PID2020-118229RB-I00 (Principal Investigator: Fermín Fernández Calderón). The authors thank the Office of the Provost of William & Mary for faculty start-up research funds to Dr. Bravo in support of this work. Dr. Bravo was also awarded the Margaret S. Glauber Faculty Fellow Award to support this work.

2.11 STATE OF ART: The past, present and future of meta-analytic structural equation modeling

Title

STATE OF ART: The past, present and future of meta-analytic structural equation modeling

Author(s)
Suzanne Jak [1]

[1] University of Amsterdam

Abstract

Meta-analytic structural equation modeling (MASEM), originally referred to as model-based meta-analysis, involves testing structural equation models on meta-analytic data. The technique is being applied in a broad range of fields, including education, psychology, environmental research, information security, medicine, and ecology. In this talk I will outline various methods that can be used to apply MASEM. I will explain how different methods may lead to different (possibly incorrect) conclusions, and consider the pros and cons of the methods currently available. As MASEM is a relatively new technique, there are many opportunities to extend existing approaches, enabling researchers to make better use of available data. Examples of necessary developments include the analysis of dependent effect sizes, handling effect size heterogeneity, synthesizing raw data, analyzing mean structures and evaluating model fit. I will therefore conclude my talk by presenting a research agenda for MASEM.

2.12 STATE OF THE ART: Big data

Title

STATE OF THE ART: Big data

Author(s)

Enrique Alonso García [1]

[1] Cuncillor of State - Spain

Abstract

To do.

2.13 KEY NOTE: Careless Responding in Survey Research: Is There Hope?

Title

KEY NOTE: Careless Responding in Survey Research: Is There Hope?

Author(s)

Ana Hernández

Abstract

In survey research, especially under unsupervised online conditions, careless responding—also referred to as insufficient effort responding—remains a significant threat to data quality. When respondents fail to engage meaningfully with questionnaire content, the resulting bias can weaken psychometric properties, distort correlations, and lead to erroneous conclusions. Recent estimates place the prevalence of careless responding between 10% and 40%, depending on survey design, context, and detection methods (Kam & Meyer, 2015; Oppenheimer et al., 2009; Ward et al., 2017). This keynote will synthesize current evidence on best practices for preventing, detecting, and managing careless responding.

Prevention strategies should reflect the dual nature of careless responding. Empirical evidence, including recent longitudinal studies from our own team, indicates that response attentiveness can vary across time and context, and may be influenced by both individual traits and situational demands (e.g., Tomas et al., 2024; Hasselhorn et al., 2023).Some individuals are consistently attentive or inattentive (trait-like), while others shift depending on situational context—such as fatigue, time pressure, or lack of interest. To address both patterns, researchers should combine context-sensitive strategies (e.g., optimizing survey length or timing) with broader, person-focused approaches like motivational instructions or commitment pledges, which can reduce carelessness even among those predisposed to inattention.

Detection strategies should be multifaceted. While attention check items offer a direct, in-survey method to flag inattentiveness, their effectiveness may decline over time as participants become familiar with them (Kam & Chan, 2018). Post-hoc statistical indices such as longstring response patterns, psychometric synonyms/antonyms, and Mahalanobis distance can be useful (Yentes, 2023), although researchers are encouraged to adopt model-based techniques, such as constrained factor and IRT mixture models (e.g. Kam & Cheung, 2023; Ulitzsch et al., 2022) and multilevel latent class analyses (Hasselhorn et al., 2023), which allow for the classification of random, patterned, and attentive respondents—without the need for additional survey items.

Managing careless responding requires more than simply discarding data. Once CR has been detected, researchers must make thoughtful decisions about how to handle it. Model-based approaches, such as constrained factor mixture models, can help disentangle trait-relevant from trait-irrelevant response patterns at the group level. However, Edwards (2019) recommends alternative strategies, such as statistically controlling for CR indices or using them as moderator variables in substantive models. These approaches acknowledge that CR can systematically influence results and should be modeled—not merely eliminated—to preserve data quality and enhance replicability.

There is hope—but only if we treat careless responding as a central concern rather than a peripheral nuisance. By integrating prevention, detection, and thoughtful data management strategies, researchers can substantially improve data quality in health and social sciences. Thus, researchers should adopt rigorous and transparent practices in dealing with careless responding in survey research.

2.14 Session 1 : ”Development and validation of psychometric instruments”

Title

Choose your own PAS? Studying the validity of the Perceptual Awareness Scale

Author(s)
Alicia Franco-Martínez [1] , Carmen Peiro-Lanchares [1] , Miguel A. Vadillo [1] , Alicia Ferrer-Mendieta [1]

[1] Universidad Autónoma de Madrid

Abstract

Any psychometrician knows how complex it is to measure constructs such as intelligence or personality. Psychological tests have addressed this task including various items and dimensions (ideally) representative of the construct under measurement. But now imagine that you intend to measure, not a trait, but a particular feature of a brief experience: the perceptual awareness of a certain stimulus. For many decades, the measure was simple: Did you see the stimulus? Seen / Unseen. The reasonable extension to a gradual measure came across by Ramsøy and Overgaard in 2004, who presented the “Perceptual Awareness Scale”(PAS), a 4-point scale with which participants had to report from ‘No experience’, ‘Brief glimpse’, ‘Almost clear image’, to ‘Absolutely clear image’in each trial. Since then, some authors have explicitly discussed the validity of this scale and proposed crucial requirements (e.g., to include the specific stimulus’feature in the PAS labels). However, most of the hundreds of papers using the PAS modify its labels and categories lacking a proper justification of these modifications. This is particularly problematic when researchers implicitly assume that different versions of the PAS are invariant measures (i.e., that a PAS=1 means the same regardless of the study), without empirical support for this assumption.

While pretesting the procedure of an ongoing replication study conducted by our team, we observed substantial changes in pilot participants’ performance when varying the PAS labels and instructions. Following up on this unexpected finding, in the present work, we have run a typical unconscious processing experiment, where trials are always preceded by a PAS assessing the visual perception of the stimulus presented. In an experimental between-subjects design, we have randomly assigned one of the conditions to each participant, varying the instructions, the presence of labels, and the versions of the PASs employed. Our goal is to determine which combination is more appropriate to measure perceptual awareness. In doing so, we will propose different ways to obtain and study validity evidence in this experimental measuring context. Additionally, we will explore the stability of the PAS labels’meanings across the experiment.

Title

Exploring the stability and temporal variability of Consideration of Future Consequences (CCF): A Trait-State-Occasion (TSO) model.

Author(s)

Geraldy Sepúlveda Páez [2] , Alejandro Vásquez Echeverría [1] , Alfredo Rodríguez Muñoz [2] , Alejandro Díaz-Guerra [2] , Mirko Antino [2]

[1] Univerdad de la República de Uruguay; [2] Complutense University of Madrid

Abstract

Introduction: Consideration of Future Consequences (CFC) is one of the temporal constructs that has undergone a broad development in psychology in recent years. Although initially conceptualized as a stable individual trait, recent research suggests that it may fluctuate over short periods of time (days or weeks). However, this possible variability has been scarcely explored, and previous literature typically assesses stability using test-retest correlation coefficients, which depend on the interval between measurements (higher in short lapses and lower in long intervals), limiting the understanding of CFC stability. To address these limitations, this study employs a Trait-State-Occasion (TSO) latent variable model, which allows decomposing variability into time-varying and time-invariant sources while also accounting for measurement error. Specifically, we apply this model to analyze the stability of CFC in Spanish workers. Methods: Two independent samples ($n_1 = 120$; $n_2 = 272$) were evaluated at four-time points, with measurement intervals of one week in the first sample and two months in the second. We first assessed longitudinal measurement invariance and subsequently fitted TSO models for each data set using the Lavaan package in R Studio. Results: Preliminary results indicate that the TSO model provides a robust modeling framework for modeling intra-individual fluctuations in CFC, particularly in the bimonthly sample, where significant temporal patterns were detected. Conclusions: These results highlight the importance of integrating latent variable models into longitudinal research to improve measurement precision in psychology, minimize the impact of measurement error, and obtain more reliable estimates over time.

Title

Evidence of the validity of the Cognitive Reserve Scale (CRS, Escala de Reserva Cognitiva, ERC). Cognitive reserve as a protective factor against the risk of violence in youth.

Author(s)
Elena Ortega-Campos [1] , Leticia De la Fuente Sánchez [1] , Mery Estefanía Buestán-Játiva [1] , Mª Dolores Roldán-Tapia [1] , Juan García-García [1]

[1] University of Almería

Abstract

Cognitive reserve has been used to explain the lack of correspondence between brain deterioration and an individual’s cognitive, behavioural, and functional performance (Stern, 2013). Over time, the scope of cognitive reserve research has expanded beyond the field of ageing and dementia, finding applications in contexts such as cognitive impairment and the development of problematic behaviours (Conté et al., 2024; Mena et al., 2024). Furthermore, studies have consistently documented a moderate to strong association between antisocial behaviour and cognitive difficulties, particularly in the domains of executive functioning and social cognition. The purpose of this study is to analyse the validity evidence of the Cognitive Reserve Scale (CRS, León et al., 2011, 2016) as a predictive tool for risk and protective factors associated with antisocial behaviour in youth, assessed using the Structured Assessment of Violence Risk in Youth (SAVRY, Borum et al., 2002). The sample consisted of 67 youth people subjected to educational judicial measures in Andalusia. The mean age was 16.6 years, with 73.1% being boys and 26.9% girls and 80.6% of Spanish nationality. Based on the type of judicial measure imposed, 59.7% were in a CIMI (Centro de Internamiento de Menores Infractores, Youth Detention Centre), 20.9% in SIMA (Servicios Integrales de Medio Abierto, Open Environment Services) and 19.4% in a cohabitation group for both boys and girls. Additionally, 29.9% had a prior record in the youth justice system. Regression models between the CRS and SAVRY scores were estimated. The regression model indicates a moderate relationship between the subdimensions of the CRS and the SAVRY RTS. Specifically, hobbies and family relationships are the most important predictors of the model. These results should be interpreted with caution as a result of the small sample size. However, they reinforce the importance of assessing cognitive reserve in adolescents and highlight its potential role as a protective factor against the risk of violence. This offers new perspectives for research on factors related to antisocial behaviour and the development of preventive strategies and interventions in the field of youth justice.

Title

First steps in the construction of a new item pool to adaptively measure numerical reasoning in university students

Author(s)
Pilar Algaba-Cenizo [1] , Ana Sanz-Cortés [2] , María Dolores Nieto-Cañaveras [3] , Juan F. Luesia [1] , Milagrosa Sánchez-Martín [1]

[1] Universidad Loyola Andalucía; [2] UNIE Universidad; [3] Nieto-Cañaveras

Abstract

Numerical reasoning (NR) is a key competence associated with higher academic performance, especially in specific fields of study. Higher education has proven to be highly relevant in academic tasks involving critical thinking, analytical thinking, and problem-solving. Assessing NR in the university admission process would provide crucial insights into the candidate's profile and enable universities to implement targeted actions to foster and enhance these skills. The measurement of numerical reasoning traditionally involves the administration of fixed-length questionnaires, with the corresponding risks they may entail in high-stakes settings (e.g., cheating). In this scenario, computerized adaptive testing (CAT) offers an alternative to overcome some drawbacks of conventional testing methods while providing more efficient measurements.

This study aimed to construct and analyze the psychometric properties of a first set of items to adaptively measure NR that will form part of a larger bank currently under construction. Based on a previous pilot study conducted with university students, different logical reasoning schemes were established to generate items according to five difficulty levels. This initial pool was administered to a huge sample of undergraduate students, and preliminary analyses were conducted to select the items with better psychometric properties for subsequent pool calibration. After that, a post-hoc simulation study was carried out to assess the performance of the CAT. The CAT is expected to be efficient and highly accurate in measuring NR using a small tailored set of pool items. Finally, some suggestions are offered regarding the different phases involving the construction of the item pool.

Title

Using confirmatory factor analysis as tool for discriminating between attribute and method effects

Author(s)
Karl Schweizer [1]

[1] Goethe University Frankfurt

Abstract
Research addressing the suitability of confirmatory factor analysis (CFA) measurement models for discriminating between common systematic variation associated with the measured attribute and common systematic variation of method effects is reported. The CFA standard version that is specified according to the congeneric model of measurement includes one latent variable that is expected to account for all the common systematic variation of data and, therefore, may not be suitable for discriminating between different types of common systematic variation. Alternatives for achieving such discrimination are variants of the tau-equivalent measurement model which we refer to as tau-based models. Such models require assumptions characterizing the type of common systematic variation in order to account for what they are expected to account.

The empirical part of this study included the simulation of data including components reflecting the item-position effect, effect of speededness, high subset homogeneity (HSH) and wording effect besides the component reflecting the attribute. These data were investigated by one-factor congeneric and tau-equivalent and two-factor tau-based measurement models. The congeneric model led to good model fit but only discriminated between attribute and method effect in the case of HSH. In contrast, CFA with variants of the tau-based measurement models indicated good model fit whenever the specification as one-factor or two-factor model corresponded to the type of method effect. The two-factor tau-based model yielded good model fit and at the same time accomplished correct discrimination between types of common systematic variation.

Title

BERO: A New Perspective on the Psychometric Assessment of Socially Aversive Traits

Author(s)
Jaime García-Fernández [1] , Covadonga González-Nuevo Vázquez , Álvaro Postigo [1]

[1] University of Oviedo

Abstract

Introduction: Dark personality consists of an association of different traits traditionally understood as socially aversive: Psychopathy, Machiavellianism, Narcissism, among others. A substantial amount of research discusses the overlap between these traits, as many of them share similar facets. Thus, a theoretical model is proposed, consisting of eight components (Authoritarianism, Greed, Cruelty, Insensitivity, Irresponsibility, Manipulation, Arrogance, and Vengeance), whose definitions do not overlap.

Objective: The aim of this research is to create a questionnaire that provides empirical evidence to support the aforementioned theoretical model.

Participants: Three samples from the general adult Spanish population were selected. The first two were used for the pilot study: one with 206 individuals (Mage = 39.53, SDage = 17.95, 73.79% women) and another with 237 (Mage = 43.63, SDage = 16.37, 67.09% women). The third sample was used for the main study, with 1,064 participants (Mage = 42.42, SDage = 13.38, 63.06% women).

Methodology: A set of 160 items was developed to assess the eight components, which were reviewed by a team of psychometricians and a panel of experts composed of 22 professionals from various fields of psychology. The initial items underwent a qualitative and quantitative pilot test (item analysis and exploratory factor analysis). After removing poorly performing items, the revised questionnaire was administered to the main sample to analyze each dimension using Item Response Theory techniques. Additionally, measurement invariance across sex, evidence of convergent/discriminant validity of its factors, and the internal structure of the complete battery were examined.

Conclusions: A final battery consisting of nine traits is proposed, whose scores have shown strong evidence of content validity and internal structure for assessing dark personality in the general adult Spanish population, confirming the proposed theoretical model. Limitations inherent to the assessment of these traits will be discussed, as well as future research directions, such as adapting this instrument to the organizational setting.

2.15 Session 15 : ”Multilevel models and Individual differences”

Title

Comparing nested multilevel models with the Vuong test

Author(s)
José María Santa Olalla Tovar [2] , Belén Fernández-Castilla , José Ángel Martínez-Huertas [1]

[1] UNED; [2] Department of Methodology of Behavioral Sciences. National Distance Education University (Spain)

Abstract

Vuong test is an extension of Likelihood Ratio Test that allows to compare non nested models. Besides, there is no difference between Likelihood Ratio Test and Vuong test for nested models. Voung test has been only available for classical models for years. The dependence of Level 1 units blocks application of Vuong test for multilevel models. Moreover, the lack of independence blocks Likelihood Ratio Test implementation too. However, Likelihood Ratio Test is usually applied through anova function in R.

Recently Vuong test has been implemented for SEM and Multilevel models in the nonnest2 R package. The nonnest2 package version of Vuong test, in the case of multilevel models, overcome the problem of lack of independence of level 1 units. For this reason, Vuong test as implemented in nonnest2 might be more efficient than anova function when comparing nested multilevel models if the sample is clearly dependent.

The objective of this work is to contrast the efficiency of Voung test with that of Likely Ratio Test for nested model comparison. We will try to answer the following question: Under which conditions does Vuong test through nonnest2 perform better that Likelihood Ratio Test through anova R function.

To achieve that aim, a simulation study will be performed for nested multilevel models'comparison. Different sample sizes of level 1 and level 2, inter group and residual variance and effect size of the predictors will be simulated as conditions of the study.

The null model, a model with a predictor and a model with two independent variables, one of them non significative will be estimated.

The simulation study is being developed now, and though results are not available yet, we will dispose them in July 2025. We have developed a preliminary pilot study that suggest that Likelihood Ratio Test for dependent data is less efficient as the sample size increases. Once the simulation study will be done Voung test as implemented in nonnest2 might achieve high proficiency while Likelihood Ratio Test implemented through anova function fails.

Title

Predicting Intersectional Inequalities Using Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA)

Author(s)
GEORGE LECKIE [1]

[1] University of Bristol, UK

Abstract
Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) is a recently developed multilevel regression modeling approach for investigating social inequalities in individual outcomes. Grounded in intersectionality theory, MAIHDA quantifies social inequalities across intersections of multiple social identities (e.g., gender, ethnicity, social class) rather than examining identities in isolation.

Proponents of MAIHDA argue that its predicted intersectional means are statistically superior to simple means derived from descriptive statistics or conventional regression models. However, this claim has yet to be formally tested. In this study, we derive and analyze analytical expressions to compare the bias, variance, and mean squared error properties of two competing MAIHDA-based mean estimators against simple means.

Our findings show that MAIHDA means outperform simple means, with the best results achieved by the approach that decomposes intersectional means into additive and non-additive components. However, the relative advantages of the two MAIHDA estimators depend on the nature of intersectional inequalities and the sample data. All three prediction methods converge as the overall magnitude of inequalities increases, departures from additivity decrease, and intersection sizes grow.

Thus, the benefits of MAIHDA are most pronounced when inequalities are subtle—whether in magnitude or in hidden processes affecting only certain social identity combinations—and when data on some intersections, such as multiply marginalized groups, are sparse.

Title

semnova: An R Package for Investigating Interindividual Differences in Experimental Effects on Latent Variables

Author(s)
Axel Mayer , Marcel Koppka , Benedikt Langenberg [1]

[1] Maastricht University

Abstract

Introduction. Interindividual differences are a fundamental aspect of experimental research, yet many statistical methods primarily focus on estimating average effects, overlooking the variability in how individuals respond to experimental manipulations. Understanding these differences is essential for psychological and behavioral research, as it provides insights into the diverse ways individuals interact with experimental conditions. However, traditional statistical methods impose restrictive assumptions like sphericity, limiting their ability to account for individual variability and measurement error. To address these limitations, we introduce semnova, an R package designed to model interindividual differences in experimental effects using latent variables. By extending structural equation modeling (SEM), semnova provides a flexible framework that determines both mean effects and their variances, enabling a more comprehensive analysis of experimental data.

Methods and Results. semnova builds on the latent growth components approach, which can be used to model experimental effects as growth components using a customized contrast matrix. This approach allows researchers to estimate within- and between-subject interactions while simultaneously accounting for measurement error. A key feature of semnova is its ability to estimate not only mean effects but also their variability. This also facilitates the examination of how individual characteristics moderate effects of experimental manipulations. Furthermore, semnova supports a multi-group framework with stochastic group sizes and variance heterogeneity, making it applicable across diverse experimental designs. Full information maximum likelihood for handling missing data and robust estimators for dealing with non-normality are readily available.

To illustrate its capabilities, we apply semnova to a longitudinal study on children's reading efficiency, tracking their development from grade one to grade four. The dataset includes eye-tracking measures such as fixation duration, re-fixation time and total viewing duration. semnova is used to estimate latent growth trajectories and analyze the impact of experimental conditions, such as sentence type (regular vs. Landolt sentences) and dyslexia status, on reading efficiency. The model specification incorporates user-defined contrast matrices to capture complex interaction effects and includes latent variables to account for measurement error. Additionally, the method supports the examination of measurement invariance across different participant groups, ensuring that the observed effects are not confounded by structural differences in data measurement.

Discussion and Conclusion. Beyond longitudinal designs, semnova can be applied in various experimental paradigms, including ecological momentary assessment studies, where interventions are administered repeatedly over time. By bridging traditional experimental designs with modern SEM-based techniques, semnova empowers researchers with a robust and flexible tool to investigate individual differences in experimental effects. Future developments will further expand its functionality, enhancing its applicability in experimental psychology and related fields. The ability to assess sphericity violations, model latent growth trajectories and capture interindividual variability allows semnova to offer deeper insights into the mechanisms underlying experimental effects, making it a valuable resource for researchers seeking a more refined approach to statistical modeling.

Title

Pooling Correlation Matrices in Meta-Analysis: Addressing Hierarchical Effect Size Dependencies

Author(s)
Diego G. Campos [1] , Ronny Scherer [1]

[1] University of Oslo

Abstract

Meta-analyzing correlation matrices has become an essential tool for synthesizing relationships among constructs, measures, or concepts across studies. Traditional univariate and multivariate meta-analytic models allow researchers to create a single, pooled correlation matrix which can then be used in advanced analyses, such as meta-analytic structural equation or network modeling. However, the presence of multiple correlation matrices derived from diverse sources within studies—such as different samples, locations, or labs—introduces hierarchical dependencies. This effect size multiplicity complicates the pooling process and has been largely overlooked in existing meta-analytic approaches. To address this challenge, we propose a Multilevel, Multivariate, and Random-Effects Modeling (MLMV-REM) framework that pools correlation matrices and accounts for their hierarchical dependencies. This innovative framework enables meta-analysts to explore various assumptions about random-effect dependencies, facilitating the selection of an appropriate meta-analytic baseline model. By incorporating hierarchical structures into the analysis, our approach enhances the exploration of heterogeneity of pooled correlation matrices at various levels of analysis.

Title

A Multi-Method Approach to Investigating Between-Group Differences in Latent Variables

Author(s)
Marcos Romero-Suárez [2] , Jesús Mª Alvarado-Izquierdo [1] , Marta Evelia Aparicio-García [1]

[1] Complutense University of Madrid; [2] Universidad Autónoma de Madrid

Abstract

In psychology, much scientific research is based on the development of latent variable models designed to represent psychological constructs and capture the complexity of human reality. This work proposes a multi-method approach to the measurement and analysis of these constructs, combining advanced techniques that allow for more robust, standardized, and replicable results. This approach optimizes the integration of different methods, overcoming the inherent limitations of each technique through the joint application of various procedures.

The example presented in this study analyzes differences in the factors underlying the Conformity to Masculine Norms Inventory (CMNI) across two age groups.

First, Exploratory Graph Analysis (EGA) was used to check correct dimensionality and model specification and to avoid problems associated with incorrect item grouping. Then, Bootstrap Exploratory Factor Analysis (bootEFA) was used to assess the stability and robustness of the factor structure, minimizing the risk of capitalization on chance. Confirmatory Factor Analysis (CFA) was then performed to confirm the adequacy of the proposed model. The derived fit indices allowed the evaluation of the fit between the theoretical model and the empirical data. Next, a measurement invariance (MI) test between the two age groups was performed to test the equivalence of the model structure in both groups. Then, a Structural Equation Model (SEM) was applied, in which the grouping variable (age) acted as a predictor of the CMNI latent variables, providing information and quantifying the differences in the CMNI latent variables between the two age groups. Finally, recommendations for different scenarios were presented, highlighting the advantages of this multi-method approach to construct analysis.

2.16 Symposium : "Advances in Investigating Response Behavior"

Title

The Influence of Time of Day on the Occurrence of Careless and Insufficient Effort Responding

Author(s)
Tobias Deribo [1] , Ulf Kroehne

[1] DIPF | Leibniz Institute for Research and Information in Education

Abstract
Questionnaires are a cornerstone of scientific research when wanting to measure non-cognitive constructs. However, low motivation to complete them can lead to improper responses and compromise the validity of the drawn conclusions (i.e., Maniaci & Rogge, 2014; Podsakoff et al., 2012), especially when using unsupervised online formats (i.e., Kroehne et al., 2020). One specific factor related to response motivation may be the time of day the questionnaire was undertaken (i.e., Kouchaki & Smith, 2014; Olsen et al., 2017). Here, prior research specifically points to night-time as a risk factor, as it may be coupled with increased exhaustion, tiredness, and depleted cognitive resources (i.e., Dickinson and Whitehead, 2015). Furthermore, this effect may be enhanced when surveys are work-related but completed outside the professional context. Therefore, this study wants to investigate how the time of survey processing impacts unmotivated response behavior in the form of Careless and Insufficient Effort Responding (C/IER; Huang et al., 2015) and if there are other variables related to the time of survey processing. The analysis draws on data from N = 2,699 teachers and N = 711 pedagogical staff participating in an online questionnaire to measure school and teaching development conditions. Teachers provide an interesting sample here, as they are often susceptible to work outside of regular working hours (e.g., Forsa, 2022). A combination of straight-lining and rapid responding, was used as indicators of low response motivation (Curran, 2016). From these indicators, we derived the proportion of unmotivated responses per respondent. To address the research question, a Bayesian Zero-Inflated Beta Regression (i.e., Ospina & Ferrari, 2010) was applied to predict the appearance and the number of unmotivated responses. Predictors were gender, measures for work-related fatigue, as well as time of survey processing. The model was estimated with weakly informative priors and four chains with 5.000 iterations (half as burn-in). The applied model converged well with a Potential Scale Reduction Factor of < 1.01 for all modeled parameters and an Effective Sample Size > 1.000 (Bürkner, 2017).
The study examined response behavior in an school survey and found that survey timing and role type influence data quality. While teachers showed stable response patterns regardless of time or work-related fatigue, pedagogical staff showed a lower probability for C/IER outside of regular working-hours. However, no practically significant effects were found regarding the proportion of unmotivated responses.
The findings challenge simple assumptions about when and why C/IER occurs and underscore the importance of considering participant role and state when designing online surveys. Enabeling staff to complete surveys during regular working hours may remain advisable, but results also suggest that some individuals may be more engaged outside of typical work times.

Title

Real-time detection of unmotivated response behavior in questionnaires - Can immediate feedback influence future response behavior?

Author(s)
Frank Goldhammer , Ulf Kroehne , Lothar Persic-Beck , Leonard Tetzlaff , Carolin Hahnel

Abstract

Unmotivated responses, identified using response times (as rapid guessing in cognitive tests, Wise & Kong, 2005; or as rapid responding in questionnaires, as part of the careless and insufficient effort responding, C/IER), are a known threat to validity (e.g. Wise, 2017). It is known from the literature that unmotivated response behavior occurs more frequently in low-stakes assessments (Wise et al., 2009), for male test takers or respondents (e.g. DeMars, Bashkov, & Socha, 2013) and with increasing item numbers (Lindner et al., 2019). However, specific psychometric models for identified rapid responses at item level (e.g., Deribo et al., 2021) or incorporating response time effort (RTE) as a process indicator at person level can only indirectly improve data quality and the validity of measurements as a post hoc correction based on already contaminated data. This paper examines how real-time detection of unmotivated response behavior during the data collection is possible and whether immediate feedback on the observed unmotivated response behavior as micro-intervention influences future responding. Using an experimental design with between-subject variation, feedback on leaving a questionnaire page that indicates missing answers, monotonous (i.e., “straightlining”) or rapid (i.e., “rapid responding”) answers in a computerized questionnaire (experimental group) is compared with feedback that only indicates missing answers (control group). The questionnaire providing log event data necessary for the identification of item-level response times in questionnaires with several items per page was administered in a nation add-on study to PISA 2022 (N=705). The position of contiguous questionnaire screens, each containing one scale, was counterbalanced using a balanced design with 18 booklets. Real-time detection of rapid responding was implemented using algorithmic processing of log events (Kroehne & Goldhammer, 2018) to extract the average answering time (AAT). AAT is known from previous analyses (Kroehne et al, in press) to show a bimodal distribution in the presence of rapid response behavior, and a conservative time threshold of 1.0 seconds was chosen for the detection of rapid response behavior in the experimental condition. The results confirm the expected bimodal distribution of the AAT, supporting the two hypothesized response processes in the experimental condition and control group. For both male and female 15-year-old students, a significant effect of the feedback on the average response time and the probability of showing quick response behavior (standardized odds ratio of 0.867 for boys and 0.734 for girls) was found. In addition to the direct effects of the micro-intervention, which remains significant when controlled for position effects, we report further indirect effects on data quality (reliability, differential item functioning and latent correlations), and present descriptive results of an inserted in-situ question to test takers on how the identified responses should be used. While the real-time detection of unmotivated response behavior affects future response behavior, the overall effect sizes of the micro-intervention are low. In the concluding section, the practical significance of the results for future computerized surveys is discussed.

Title

Experimental Validation of Model-Based Identification of Careless and Insufficient Effort Responding

Author(s)
Esther Ulitzsch [1] , Gabriel Nagy , Irina Uglanova [2]

[1] University of Oslo; [2] Leibniz Institute for Science and Mathematics Education

Abstract

Self-report surveys often suffer from careless and insufficient effort responding (C/IER), which refers to responses provided without paying attention to the items'content. Mixture modeling approaches are promising tools to assess C/IER by means of latent class variables. However, evidence for the validity of interpreting the latent class variable as C/IER is still pending. To shed more light on this issue, this paper presents the results of a pre-registered survey experiment. Specifically, we examine the ability of a recently developed mixture item response theory model (Uglanova, Nagy, & Ulitzsch, in preparation) to detect C/IER in self-report measures of dark personality traits. We evaluated two validity arguments: the relationships of the latent class (1) with experimentally manipulated survey conditions and (2) with alternative indicators of C/IER. Experimental conditions were designed to evoke or prevent C/IER by manipulating the instructions, the presence of cognitively demanding tasks, and the number of preceding items. Alternative indicators of C/IER were attention check items, item content recognition tasks, and self-reported C/IER. Using the bias-adjusted three-step approach to relate latent class membership to external variables we found that (a) respondents in the evoking C/IER condition were more likely to be assigned to the C/IER class than respondents in the preventing condition, and (b) respondents assigned to the C/IER class exhibited lower performance on all alternative indicators of C/IER than those in the attentive class. Overall, these results were consistent with our pre-registered hypotheses, providing validity arguments to support interpreting the latent class variable as representing C/IER.

Title

A Multilevel Mixture Item Response Theory Model for Partial Engagement in Proficiency Tests

Author(s)
Gabriel Nagy , Esther Ulitzsch [1]

[1] University of Oslo

Abstract

Disengaged test-taking behavior is a problem in low-stakes assessments. To account for low engagement, popular approaches rely on item response times to classify responses as disengaged (rapid guessing) or inconspicuous (engaged). Although conceptually elegant, this binary classification has been found to miss a substantial proportion of disengaged responses. This paper introduces an extended classification of engagement that includes “partial (dis)engagement”. To this end, a Multilevel Mixture Item Response Theory (MMIRT) model is proposed that classifies engagement at the item level. Partially engaged responses are specified to be associated with response times that fall between the very short response times of rapid guesses and the response times of engaged responses. Responses are classified on the basis of within-individual response time distributions, meaning that the model accounts for individual differences in habitual time expenditure. Disengaged responses are modeled as the result of a guessing process, while partially and fully engaged responses are both related to the proficiency variable via a three-parameter response model. The MMIRT model can be estimated using maximum likelihood techniques via the expectation maximization algorithm. The MMIRT model is illustrated with data from the TIMSS 2019 science assessment. Multiple-choice items presented at the beginning and end of the test in a rotated test design were analyzed. In the U.S. sample of eighth graders, test performance was lower on items presented at the end of the test. The lower performance could not be explained by models based on the binary classification of response engagement. In contrast, the proposed MMIRT model suggested that the decline in performance was due to an increase in partially disengaged responses.

Title

Investigating the interplay of text rereads with IRT parameters: Rereads render hard items easier and easy items harder

Author(s)
Esther Ulitzsch [1] , Gabriel Nagy , Jana Welling

[1] University of Oslo

Abstract
In reading comprehension tests, test-takers can choose to reread the text of the task while working on an item. Up to now it is not well understood how rereading the text relates to test performance and its measurement. To close this gap, the aim of the present study was to investigate the relationship between text rereads on one hand and item parameters of item response models and test performance on the other hand. We specified three different item response mixture models that distinguish on the response level between the three latent classes rapid guessing, solution behavior with text rereads and solution behavior without text rereads. The different models assumed either (1) equal item parameters, (2) equal item discriminations but varying item difficulties, or (3) varying item parameters between the two different solution behavior classes. In a reading comprehension test of the German National Educational Panel Study (N = 1933 students, 14 multiple-choice items), the second model with equal item discriminations but varying item difficulties between the two latent classes fitted the data best. Descriptive analysis revealed that the reread class did not differ extensively from the no reread class in the average item difficulty, but rather exhibited less variation in the item difficulties, rendering hard items easier and easy items harder. Furthermore, the tendency to reread the text positively predicted test performance. The results highlight the importance of investigating process data beyond item response times, which can help to better understand the test-taking process as well as its interplay with the measurement of test performance.

Title

Benefits of Process Data for Evaluating the Differential Effectiveness of App-Based Treatments

Author(s)
Esther Ulitzsch [1] , Janne Torkildsen , Jarl K. Kristensen , Marie-Ann Sengewald [2]

[1] University of Oslo; [2] Leibniz Institut for Educational Trajectories

Abstract
Differential effect analysis can reveal the preconditions for effective interventions by highlighting variations in intervention outcomes. The growing use of digital tools, such as learning apps, provides rich process data on response times and response behavior, offering insights into how participants interact with these apps. We use this information source and bridge psychometric research on controlling for disengaged responding with differential effect analysis to evaluate how variations in the usage of learning apps contribute to heterogeneity in intervention effectiveness. Specifically, we consider different response-time-based indicators to identify disengaged behavior, including thresholds for overly short response times, Gaussian mixture modeling, and model-based approaches that integrate item responses and response times (e.g., Wise & Kong, 2005; Wise, 2017; Ulitzsch et al., 2020; 2023). We demonstrate how these indicators can be integrated into the EffectLiteR framework, which specifies a structural equation model for differential effect analyses with latent variables (Mayer et al., 2016; Sengewald & Mayer, 2024). Finally, we compare the different modeling strategies and investigate the benefits of using the disengagement indicators in an empirical application. For this, we rely on the work of Torkildsen et al. (2022), who constructed an app-based morphological training program and evaluated its effectiveness in a randomized controlled trial with 717 second-grade students. Using the empirical data, we examine the heterogeneity of the morphological training effects in relation to the pre-treatment characteristics of the students and the gains achieved by including the different disengagement indicators, focusing on their impact on explained outcome variance and effect size differences. Our findings identify baseline characteristics that predict greater benefits from the training and highlight how different modeling strategies for disengagement indicators influence the conclusions. Beyond the practical insights into the utility of process data, the results demonstrate the application of the advanced modeling strategies for differential effect analysis.

2.17 Symposium : ”Advancing Dynamic Methods for Modeling Change Over Time”

Title

Modeling overnight lags in daily emotion dynamics

Author(s)
Pablo Fernández Cáncer [2] , Eduardo Estrada [1]

[1] Universidad Autónoma de Madrid; [2] Universidad Pontificia Comillas

Abstract
Introduction. Experience sampling methods (ESM) are an increasingly popular strategy for studying affective processes (i.e., mood and emotions). In these studies, the emotional state of one or more individuals is measured several times a day during multiple days or weeks. A unique feature of these studies is the spacing

of observations: measurements are frequent during waking hours but separated by a much longer interval overnight while participants sleep. This uneven distribution poses challenges for dynamic models, where emotional states are represented as a function of previous states and dynamic noise. Importantly, the overnight gap may induce changes in emotional dynamics that cannot be explained solely by the length of the interval.

For example, emotional states at bedtime may exert an influence on morning affect that differs from daytime patterns. Despite its potential impact, the role of overnight lags has been largely overlooked in the literature. Typical approaches either ignore these effects or exclude nighttime intervals entirely, which simplifies the data structure but may overlook meaningful dynamics in the transitions between days. In this study, we evaluate the efficacy of various modeling strategies to address overnight effects within the framework of statespace models. Specifically, we investigate how overnight lags can be incorporated to account for changes in emotion dynamics that occur between consecutive days.

Method. To evaluate the performance of the strategies compared, we conducted a Monte Carlo study under a range of conditions that are frequent in experience sampling studies. We also applied the proposed approaches to existing datasets on affect dynamics to illustrate their implementation and practical utility.

Results and discussion: We discuss the implications of modeling overnight dynamics, highlighting the importance of accurately capturing these effects for a more nuanced understanding of daily emotional processes.

Strengths, limitations, and future directions for improving the handling of irregular time intervals in ESM research are also considered.

Title

Kalman scores for the estimation of planned and unplanned missing individual observations in accelerated longitudinal designs

Author(s)
Eduardo Estrada [1] , José Ángel Martínez-Huertas [2] , Ricardo Olmos [1]

[1] Universidad Autónoma de Madrid; [2] UNED

Abstract

A popular cost-effective way of collecting longitudinal data is the accelerated longitudinal design (ALD). In ALDs, participants from different cohorts are measured repeatedly but the measures provided by each participant cover only a fraction of the time range of the study. It is then assumed that the common trajectory can be studied by aggregating the information provided by the different converging cohorts. ALDs are, therefore, characterized by a very high rate of planned data missingness. Additionally, it is very common that most
longitudinal studies present unexpected participant attrition leading to unplanned missing data. A way for analyzing this data is the latent change score (LCS) model within a Continuous-Time State-Space Modeling framework (CT-SSM). This CT-SSM model allows computing Kalman scores, which can be used to estimate individual observed and unobserved scores. We simulated an accelerated longitudinal design where we manipulated different conditions such as the sample size, the unplanned missing data mechanism (MCAR, MAR, MNAR), and the severity of the unplanned missingness. Results showed that the Kalman scores were able to estimate both (1) data points that were expected but unobserved and (2) data points that were outside the age range observed for each case (i.e., to estimate the individual trajectories for the complete age range un-
der study). These results have important implications for practitioners in psychology and education because they make it possible to accurately forecast individual longitudinal trajectories and to make individual-level decisions considering the model predictions. This presentation summarizes part of the results of a recent publication: https://doi.org/10.1037/met0000664

Title

State-Space Models for Identifying Abrupt Changes in Cognitive Development

Author(s)

Eduardo Estrada [1] , Marcos Romero-Suárez [1] , Pablo Fernández Cáncer [2]

[1] Universidad Autónoma de Madrid; [2] Universidad Pontificia Comillas

Abstract

State-space models (SSMs) provide a powerful framework for modeling dynamic systems, capturing both intra-individual and inter-individual variability in longitudinal data. In the context of cognitive development research, one interesting feature of SSMs is their ability to model deviations, or “shocks,”in individual trajectories. Such shocks may signal atypical changes that could be considered outliers within developmental processes. In this study, we adapt a semi-exploratory procedure proposed by You et al. (2020) to the context of cognitive development, using the dynr package (Ou et al., 2019) in R.

Our main objectives are to: a) propose a novel SSM designed to detect outliers in developmental trajectories; and b) evaluate its performance in terms of accuracy of outlier detection and recovery of the population parameters.

To evaluate these objectives, we performed an extensive Monte Carlo study. First, we generated data based on empirical trajectories. We manipulated several simulation conditions, including sample size, number of time points per participant, timing of shocks, and proportion of participants affected by shocks. Next, we examined the impact of these factors on group-level parameter bias, and the balanced accuracy of the individual outliers identification. Based on our findings, we discuss the utility of SSMs to detect abrupt environmental changes affecting cognitive development.

Title

When Should I Measure? Finding the Best Sampling Schedule for Recovering Longitudinal Dynamics in Panel Data Studies with Continuous-Time Models

Author(s)
Eduardo Estrada [1] , Nuria Real-Brioso [1]

[1] Universidad Autónoma de Madrid

Abstract

One of the key questions in longitudinal research is when to take measurements of the variables of interest. Panel studies usually focus on the dynamics between two processes over time (e.g., depressive symptoms and self-esteem), and include few repeated measures (<10). This forces researchers to find the most efficient way to design their study and collect their data. Recently introduced in Psychology, continuous-time models are very convenient in this context, as they can accommodate irregularly spaced measurements, both between and within individuals.

Previous research on deciding the optimal time interval between measurements have proposed various criteria, such as using the time interval at which the overall cross-effects are largest, or the time interval leading to best estimation reliability. However, relatively less attention has been paid to the effect of stochastic innovations (i.e., dynamic error) on the sampling design, despite its key role in the stability of the system.

In a Monte Carlo simulation, we used state-space continuous-time models to characterize the dynamics of two variables measured longitudinally through panel designs. We compared various sampling schedules, including those suggested in recent research, to evaluate their effectiveness in recovering bivariate trajectories under various levels of stochasticity. We discuss the strengths and weaknesses of different sampling approaches and provide practical recommendations on using continuous-time modeling in panel data studies.

2.18 Symposium : ”Methodological Advances in Meta-analysis”

Title

Visualization of heterogeneity in forest plots

Author(s)
Wolfgang Viechtbauer [1]

[1] Maastricht University

Abstract

The findings of a collection of studies addressing a common research question can be visualized in terms of a forest plot, showing the effect sizes of the individual studies together with a corresponding confidence interval. A four-sided polygon (sometimes called a summary 'diamond') is often added to such a plot to depict the results from a meta-analysis pooling together the effect sizes, where the center of the polygon corresponds to the pooled estimate and the ends of the polygon represent the bounds of the confidence interval for the pooled estimate. However, this only communicates the size of the average effect and how precisely it is estimated. In addition, it is equally important to indicate the degree of heterogeneity among the findings, that is, the variability in the underlying true effects. Such information (e.g., the results from the Q-test, the I2 statistic, and the estimate of τ2 from a random-effects model) is often only added textually underneath the plot. In this talk, I will describe several alternative visualizations of the amount of heterogeneity in terms of the prediction interval and by showing the entire prediction distribution. This also raises interesting issues when applying a back-transformation of the results (such as exponentiation when meta-analyzing log-transformed estimates or the hyperbolic tangent function when meta-analyzing Fisher r-to-z transformed correlation coefficients), since this impacts not only the shape of the prediction distribution, but also what the back-transformed estimates represent. The various types of visualizations discussed are already implemented in the metafor package for R and can be readily used by practitioners.

Title

Relationship between repeated measures in clinical psychology studies: an empirical evaluation

Author(s)
Juan José López García , Manuel Jesús Albaladejo Sánchez , Fulgencio Marín-Martínez [1] , Jose Antonio Lopez Lopez , María Rubio Aparicio [2] , Julio Sánchez-Meca [1]

[1] University of Murcia (Spain); [2] University of Murcia

Abstract
Effect sizes are commonly used in meta-analysis, as they provide a tool to summarize the results from each primary study in a common metric. In psychology and related fields, meta-analyses often involve integrating continuous variables measured with different scales across studies, which leads to using standardized mean differences as the effect size index. One of these indices is the standardized mean change (SMC), which quantifies within-group treatment effects when the primary studies have examined the effectiveness of a treatment program using a repeated measures design, and the dependent variable has been measured on a quantitative scale. However, different procedures have been proposed to calculate this effect size, and some of them make assumptions which are hardly verifiable in practice, namely homoscedasticity and a specific value for the correlation between measurement points. This presentation will explore the potential impact on the results if some of these assumptions are violated, using a range of simulated scenarios.

Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

Title

Bias and Mean Squared Error of Six Estimators of the Standardized Mean Change in Pretest-Posttest Designs

Author(s)
Juan José López García , Manuel Jesús Albaladejo Sánchez , Fulgencio Marín-Martínez [1] , Jose Antonio Lopez Lopez , María Rubio Aparicio [2] , Julio Sánchez-Meca [1]

[1] University of Murcia (Spain); [2] University of Murcia

Abstract
The standardized mean change is widely recognized as a key effect size index in pretest-posttest one-group designs with quantitative dependent variables. Different parametric versions of this index are available, depending on the standardizer used to scale the mean difference into standardized units. In addition, various estimators can be applied to each parameter. This study used a Monte Carlo simulation to assess the bias and mean squared error of several estimators for the standardized mean change. Key factors, such as population effect size, population correlation, sample size, and heterogeneity of pretest and posttest population variances, were systematically manipulated. The results offer valuable insights for selecting the most efficient estimator, taking into account the chosen parameter, study characteristics, and the potential of integrating effect sizes into meta-analyses.
Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

Title

An improved homogeneity test for meta-analysis of standardized mean differences

Author(s)
Juan Botella [1] , Juan I. Durán [1] , Manuel Suero [1]

[1] Universidad Autónoma de Madrid

Abstract

In meta-analysis, the Q statistic is traditionally used for testing the hypothesis of homogeneity of the parametric effect sizes of the set of studies. Several critiques have been posed to that test, especially when applied to the standardized mean difference (g). Among them, that the weights are based on estimated, not true, variances, that the variances of the estimates correlate with the own g values, and that it is assumed a wrong distribution of g (normal) although it is actually a linear transformation of a Student's t. We present an improved test of homogeneity of g values based in the Mixture Model of Suero et al (in press) in which most of the problems highlighted are solved or greatly reduced. Specifically, the variances of g in the studies are independent of the own g values, and the true distribution of g is acknowledged and transformed to a normal distribution. Although the variances are still estimated, we show that their impact in the performance of the test is negligible under the Mixture Model. We present the results of an extensive Monte Carlo simulation to assess the performance of the classical Q test, an alternative that use weights based only on the samples sizes (effective sample size, Qñ) and two normalizing transformations (those of Johnson-Welch and Laubscher). The results show that whereas the classical Q test yields a too low rate of type I errors and Qñ an unacceptable large rate, the two normalizing transformations yield rates within a comfortable 4% –6 % range. Furthermore, the rate of correct rejections (estimated power) is always higher than that of the Q test. Taking all the results together, we conclude recommending the new test for meta-analysis of g values, using the estimates provided by the Mixture Model and the normalizing transformation of Johnson-Welch.

Title

Statistical power of random-effects meta-analyses of clinical psychological interventions

Author(s)
Rubén López Nicolás [1] , Alejandro Sandoval-Lentisco [2] , Jose Antonio Lopez Lopez , Robbie van Aert , Julio Sánchez-Meca [3]

[1] Universidad de Castilla la Mancha; [2] Universidad Autónoma de Madrid; [3] University of Murcia (Spain)

Abstract

Underpowered studies are ubiquitous in psychology and related disciplines. Meta-analysis can help alleviate this problem, increasing the statistical power by combining the results of a set of primary studies. However, this is not necessarily true when we use a random-effects model, which is currently the predominant approach when carrying out meta-analyses. In this study, we examined the statistical power of a sample of 141 meta-analyses on the effectiveness of clinical psychological interventions. Additionally, we compared the estimated statistical power of these meta-analyses with the power of the individual studies that comprised them and computed the minimum number of primary studies needed to achieve 80% statistical power. To do so, we used different analytical approaches and a Monte Carlo approach. The statistical power of random-effects meta-analyses was computed under different values of the true effect size and levels of heterogeneity. Our results show that under certain scenarios, the hypothesis test of the null-hypothesis of no average effect is underpowered. These scenarios were characterised by small true effect sizes, high heterogeneity, and a small number of included studies in the meta-analysis. Statistical power of the meta-analysis could also be lower than the median or maximum power of the included primary studies. These results are discussed in light of the statistical basis of random-effects meta-analysis, and recommendations are made for applied researchers. Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

Title

Reliability generalization of the Emotional Quotient Inventory Youth Version (EQ-i: YV): A meta-analytic structural equation modelling approach

Author(s)

Raimundo Aguayo-Estremera [1] , Alejandro Veas Iniesta [2] , Jose Antonio Lopez Lopez , Julio Sánchez-Meca [3] , José Antonio López Pina

[1] Universidad Complutense de Madrid; [2] Universidad de Murcia; [3] University of Murcia (Spain)

Abstract

Recent research has identified several limitations in traditional methods for conducting meta-analyses of reliability generalization, such as the lack of equivalence between total and subscale reliability indices and the violation of error independence assumptions. In response, multivariate statistical techniques have been developed to offer more accurate estimations of measurement instruments, one of which is meta-analysis of structural equation modelling (MASEM). MASEM offers significant advantages, including the ability to combine correlation matrices from primary studies and to estimate factor models more efficiently. This communication demonstrates the application of MASEM to the Emotional Quotient Inventory Youth Version (EQ-i:YV), a widely used tool for assessing emotional intelligence in children and adolescents worldwide. By employing a MASEM approach, we will derive more robust reliability estimates using Omega values, which are more suitable for multidimensional measures compared to traditional Cronbach's alpha. Ultimately, we aim to enhance our understanding of the psychometric properties of the EQ-i:YV and contribute to advancing theoretical development in the field of emotional intelligence assessment.

Funding: MICIU/AEI /10.13039/501100011033/ and FEDER funds, European Union, grant no. PID2022-137328NB-I00

2.19 STATE OF THE ART: Dealing with publication bias in a meta-analysis

Title

STATE OF THE ART: Dealing with publication bias in a meta-analysis

Author(s)
Robbie van Aert

Abstract
Meta-analysis is the statistical methodology to synthesize findings across multiple studies. However, publication bias is arguably one of the most important threats to the validity of a meta-analysis. One major consequence of publication bias is overestimation of the meta-analytic effect size. To address this, various methods have been developed to correct for publication bias in a meta-analysis and also to test for its presence.

This presentation will start with providing a short overview of evidence for the presence of publication bias in the literature. I will then introduce several methods to test and correct for publication bias in a meta-analysis. Both traditional methods (e.g., fail-safe N and the trim-and-fill method) and nowadays recommended methods (e.g., selection model approaches and regression based methods) will be discussed. Finally, I will highlight recent advances in the field and outline directions for future research.

2.20 STATE OF THE ART: Educational programs evaluation

Title

STATE OF THE ART: Educational programs evaluation

Author(s)

José Saturnino Martínez García [1]

[1] Agencia Canaria de Calidad Universitaria y Evaluación Educativa

Abstract

To do

2.21 Poster Session 4

Title

Global versus domain-specific scores in the measurement of cognition: A study of the differences in correlations with relevant criteria

Author(s)
Sara Martínez-Gregorio [1] , Patricia Sancho [1] , Mireia Abella [1] , Irene Fernández [1] ,
Aitana Sanz [1]

[1] Universitat de València

Abstract
Background. The study of cognition is tackled from very heterogeneous approaches in the literature, some of them regarding the way it is measured. Often, cognition is conceptualized as a global ability and operationalized as a single score. However, researchers warn against this practice given that it can obscure subtle changes due to lack of sensitivity. Instead, using cognitive domains and deriving domain-specific scores is advised. In this study, we employed a global score of cognition as well as domain-specific scores of memory, orientation, visuospatial ability and executive functioning to assess differences in the relationship of these scores with theoretically-relevant correlates of cognitive ability: satisfaction with life (SwL), social activity participation (SA), intellectual activity participation (IA), limitation with activities of daily living (ADL), limitations with instrumental activities of daily living (IADL), quality of life (QoL), depression, number of chronic disorders and educational level. Methodology. We employed data from the 9th wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). We selected participants aged at least 60 years who had responded to the cognitive battery included in the survey. The final sample included 55,569 individuals from 27 European countries and Israel. Individuals were mostly female (56.8%) and had an average age of 72.07 (SD = 7.97). We derived scores of memory, orientation, visuospatial ability and executive functioning, as well as a global score made up the sum of the four domain scores. Given the ordinal nature of some of the variables and the non-normal distribution of others, we employed Spearman's correlation coefficients to estimate the relationship between global and domain-specific cognitive scores and the criteria. Correlation contrasts were performed using Fisher's Z transformation. Results. Correlations between the domain-specific scores and the criteria differed significantly ($p < .05$) from the correlations between the global cognitive score and the criteria in almost all cases. The only exceptions were the correlation of memory and SA, the correlations of memory and orientation with ADL, and the correlation between memory and educational level. Except for the correlations of memory with SwL and QoL, the correlations of domain-specific scores and the criteria were smaller than those of the global score and the criteria. Discussion. Results from this study suggest that global cognitive scores are not equivalent to single domain scores and should not be used interchangeably. In some cases, the effect sizes of the correlations were moderate/small for the global score, while they dropped to negligible/small for domain scores. Therefore, research in the cognitive arena ought to reconsider the use of global or domain-specific scores to capture cognition, as they do not behave equivalently. Future research should examine potential differences in the behavior of these scores longitudinally, in order to assess their sensitivity to detect change over time. Moreover, contrasts between correlations of the different domain-specific scores and relevant criteria ought to also be studied.

Title

Psychometric Properties of the Online Version of the Edinburgh Postnatal Depression Scale (EPDS) for its Use in Spanish Pregnant and Postpartum Women.

Author(s)
Maria Fe Rodríguez-Muñoz [1] , Carmen Rodríguez-Domínguez [2] , Sara Domínguez-Salas [2] , Diego Gómez-Baya [3] , Emma Motrico [2] , Irene Gómez Gómez [4]

[1] Universidad Nacional de Educación a Distancia; [2] University of Seville; [3] Universidad de Huelva; [4] Universidad Loyola Andalucía

Abstract

Introduction: Perinatal depression is a public health issue and is considered one of the main complications during the perinatal period. The Edinburgh Postnatal Depression Scale (EPDS) is one of the most frequently used instruments to detect perinatal depression in women. The EPDS is a 10-item self-reported scale with a 4-Likert-type response scale. Despite numerous studies about the psychometrics properties of the EPDS have being published, there are heterogeneous results regarding its internal factorial structure. The most frequent factorial structures previously found in both international and Spanish studies were the one-factor, the two-factor (depression and anxiety) and the three-factor (depression, anxiety and anhedonia) structure. In addition, it must be noted that item factor loadings varied across factors in the studies. This study aimed to obtain different sources of validity evidence (internal structure and relationship with other variables) and to analyse the psychometric properties of the online version of the EPDS for its use in Spanish pregnant and postpartum women.
Method: This study followed the Standards guidelines. Data were collected online. Exploratory factor analysis (EFA) was conducted using the principal axis factoring extraction method and Promax rotation. Confirmatory factor analysis (CFA) was carried out using the Robust Unweighted Least Squares method. Both the EFA and the CFA were conducted through cross-validation procedures. The fit of the one-factor and the two-factor (depression and anxiety) structure proposed in previous studies was also tested. Factorial invariance between pregnant and postpartum women was explored. The correlation of depression (EPDS) with anxiety (GAD-7) and post-traumatic stress disorder (PTSD checklist) was analysed. Internal consistency of the EPDS was evaluated through Cronbach's alpha and McDonald's Omega coefficients. In addition, EPDS item analysis was performed.
Results: The EFA revealed a three-factor structure that showed a good fit of the model to the data through CFA for pregnant (CFI=.995; NNFI=.993; RMSEA [95% CI] =.047 [.032; .062]) and postpartum (CFI= .996; NNFI= .994; RMSEA [95% CI] =.039 [.027; .051]) women. The one-factor (CFI≥ .974; NNFI≥ .960; RMSEA≥ .080) and the two-factor (depression and anxiety; (CFI≥ .979; NNFI≥ .973; RMSEA≥ .070) structure presented poorer fit indexes compared with the three-factor structure. In addition, it must be noted that in the sample of postpartum women, item 3 presented high factor loadings (FL) in factors 1 (depression; FL= .48) and 2 (anxiety; FL= .34). Positive (r > .500) and significant (p-value < .001) correlations were found between the depression, anxiety and anhedonia dimensions with the GAD-7 and the PTSD checklist. The Cronbach's alpha and McDonald's Omega coefficients exceeded the optimal cut-off (0.70).
Conclusions: The three-factor structure presented the best-fit indexes for both pregnant and postpartum women. Item 3 "I have blamed myself unnecessarily when things went wrong"presented higher loading factors in depression and anxiety subdimensions for postpartum women and higher loading factors in anxiety for pregnant women. However, due to theoretical and statistical reasons, a homogenised factorial structure was proposed for both pregnant and postpartum women: items 1 & 2 (anhedonia), items 3 - 6 (anxiety) and items 7 - 10 (depression).

Title

PSYCHOMETRIC PROPERTIES OF THE ROSENBERG SELF-ESTEEM SCALE IN PATIENTS DIAGNOSED WITH BREAST CANCER

Author(s)
MARÍA VICTORIA CEREZO GUZMÁN [1] , Lorena M Soria-Reyes [1] , Rafael Alarcón [1] , María J. Blanca [1]

[1] University of Malaga

Abstract

Background. Breast cancer patients need to maintain self-esteem during the disease process, as it is crucial for their mental health. It is therefore vital to have an adequate tool for measuring self-esteem in these women. Although the Rosenberg Self-Esteem Scale (SES) has been widely used in both research and clinical settings, its psychometric properties have not, to our knowledge, been studied in this specific population. The present study aimed to address this gap. Method. Participants were 170 women residing in Spain (mean age 51.22 years, SD = 8.85) who completed the SES and scales assessing other psychological variables, namely emotional distress, wellbeing and optimism. Validity evidence based on the internal structure of the scale was obtained through confirmatory factor analysis (CFA). The reliability of SES scores was assessed by calculating Cronbach's alpha and McDonald's omega coefficients. Validity evidence based on relationships with other variables was obtained by examining associations between SES scores and those on the measures of emotional distress, wellbeing and optimism. Results. The CFA indicated a one-factor structure, with acceptable fit indices. Reliability coefficient for SES scores was above .80. Correlations with other variables indicated a significant negative relationship with scores on emotional distress, and significant positive correlations with scores on the measures of wellbeing and optimism. Conclusion. The SES is a valuable tool for measuring self-esteem in the context of breast cancer, providing useful information for psychological assessment.

Title

The Workplace Ostracism Scale: A Reliability Generalization Meta-Analysis

Author(s)

Palmira Faraci [1] , Giusy Danila Valenti [1]

[1] University of Enna ”Kore”

Abstract

Although the Workplace Ostracism Scale (WOS) is a popular instrument for assessing workplace ostracism, research examining its psychometric properties is limited. Reliability is a fundamental quality of a psychometric tool. It is not a property of the test itself but of its test scores, as it depends on the sample's characteristics, the number of items, and the response format. A good practice to foresee the reliability of a scale's scores is to quantitatively integrate multiple reliability estimates derived from different administrations. Reliability Generalization (RG) meta-analyses synthesize these estimates, identifying typical reliability levels and factors influencing variability in reliability outcomes.
We performed an RG meta-analysis following the recommendations for conducting and reporting Reliability Generalization Meta-Analyses (REGEMA Checklist). Data were gathered from databases including Psychology and Behavioral Science Collection, MEDLINE, Scopus, Web of Science, Wiley Online Library, SAGE Journals, ScienceDirect, and Google Scholar.
Studies were included if they: (a) used the WOS with an adult sample, (b) reported reliability coefficients for the WOS scores, (c) were published in English in peer-reviewed journals, and (d) used the 10-item, seven-point Likert version of the scale. Moderator variables coded to assess reliability influences included: (a) participants'gender distribution and mean age, (c) WOS language, (d) country of administration, (e) focus (psychometric vs. applied), and (f) mean and standard deviation of the WOS scores. Our RG meta-analysis focused on Cronbach's alpha, the most commonly reported reliability coefficient.
Reliability estimates were standardized using Bonnett's (2002) transformation to normalize their distribution. Between-study heterogeneity was assessed with Cochran's Q test ($p < 0.05$) and Higgins'I^2, with thresholds of 25%, 50%, and 75% indicating low, moderate, and high heterogeneity, respectively (Higgins et al., 2003). Meta-regressions were then conducted to evaluate the moderation effects of categorical and continuous variables.
From an initial 1,672 records, 44 articles (50 independent samples) were retained, comprising 15,868 participants. The sample was 40.43% male, with a mean age of 32.77 years (SD = 6.90). The WOS was primarily used in English (35.8%) and Chinese (38.3%) in applied research (84.9%). The mean WOS total score was 1.98 (SD = 0.08). The mean Cronbach's alpha was .93 [CI: .92–.95], ranging from .71 to .99, with significant heterogeneity (Q = 2,634.29, $p < .001$; $I^2 = 98.15$). Moderation analyses revealed that the standard deviation of WOS scores was the only significant moderator ($p < .001$), explaining approximately 37% of the total variance.
A limitation of this RG meta-analysis is its reliance on a single reliability index, Cronbach's alpha, which has known drawbacks. Including additional reliability coefficients could have offered a more comprehensive view of WOS score reliability. Furthermore, while variability in WOS scores explained much of the heterogeneity in alpha values, a significant portion remains unexplained. Future RG meta-analyses should consider additional moderators, such as work status, job category, or contract type. However, despite these issues, our findings suggest that the WOS is a reliable measure of workplace ostracism, with adequate internal consistency and no significant discrepancies across demographics, country, or language of administration.

Title

Network Analysis of the Illness Management and Recovery Scale (IMR) in Individuals with Mental Disorders

Author(s)
María Dolores Hidalgo [1] , Albert Sesé [2] , Nuria Martín-Ordiales [1] ,
María Pilar Martín Chaparro [1] , Maite Barrios [3]

[1] University of Murcia; [2] Universitat de les Illes Balears; [3] University of Barcelona

Abstract

Understanding the complex relationships between key factors in mental health recovery is essential for improving clinical interventions. The Illness Management and Recovery Scale (IMR) is widely used to assess recovery-related processes, yet little is known about the underlying structure of its items when analyzed through network models. This study explores the interconnections between IMR items using network analysis, assesses the stability of the network structure, and examines measurement invariance between users of mental health services and healthcare professionals.

The IMR was completed by a sample of 172 users of mental health services diagnosed with a severe mental disorder and 167 healthcare professionals treating them. A network model was estimated using the Triangulated Maximally Filtered Graph (TMFG) model and the walktrap algorithm. To ensure robustness, bootstrap resampling was applied to assess the stability of edge weights and centrality indices. Additionally, an Exploratory Graph Analysis (EGA) was conducted to identify the most stable latent structure of the scale, followed by a configural and metric invariance analysis to compare factor structures between users and professionals.

Network analysis revealed 39 non-zero edges among the 15 IMR items, with coping strategies (IMR11) emerging as the most central node, closely linked to distress from symptoms (IMR6), functional difficulties (IMR7), and symptom recurrence (IMR9). Items related to social support (IMR3, IMR4, IMR5) and substance use (IMR14, IMR15) formed distinct but interconnected subgroups. The EGA model identified a predominant three-factor structure, with high stability in items related to symptom management, social and health behaviors, and personal recovery strategies.

Measurement invariance analysis indicated that 13 out of 15 items demonstrated configural and metric invariance, suggesting a largely consistent factor structure across user and professional responses. However, IMR9 (Symptom Recovery) exhibited higher factor loadings in users, implying a stronger association with their recovery perceptions compared to professionals.

These findings highlight the central role of coping strategies in mental health recovery and provide evidence for a stable three-factor structure of the IMR in a clinical population. The results support the use of network analysis to refine assessment tools and tailor interventions. Future research should further explore non-invariant items to enhance the interpretability of the IMR across different stakeholder groups.

Title

Adaptation and Validation of the the Philosophy of Social Science Inventory (PSSI) in Polish Cultural Context

Author(s)
Martyna Jarota [1] , Sławomir Pasikowski [1]

[1] University of Lodz

Abstract

Objective:
The aim of this study was to adapt and assess the psychometric properties of the Philosophy of Social Science Inventory (PSSI), originally developed in English (Johnson & Onwuegbuzie, 2004; Sheehan & Johnson, 2012), for the Polish-speaking population.
Method:
The study involved 449 participants (54% women, 46% men, M = 38.7 years, SD = 10.4), who were instructors in research design and measurement in psychology, education, and sociology. They completed the Polish version of PSSI along with additional tools measuring related constructs, including the Profile of Individual Preferences of an Investigator (PIPB-80; Nosal, 1986) and the Survey of Attitudes Toward Statistics (SATS-36; Schau, 2003).
The adaptation process included back-translation, content validity assessment, and pilot testing. The psychometric properties were examined using factor analysis (EFA, CFA), reliability analysis (Cronbach's α, McDonald's ρ), and validity analysis (convergent and discriminant validity).
Results:
Confirmatory factor analysis (CFA) showed an unsatisfactory model fit for the original structure (CFI = 0.94, TLI = 0.89, RMSEA = 0.12). Exploratory factor analysis (EFA, Oblimin rotation) suggested a two-factor structure (KMO = 0.81; Keiser criterion: eigenvalue >1 and scree plot analysis: 2, explained variance = 0.61). The average factor loadings were F1: M = 0.61, F2: M = 0.58, differing from the original model. Cronbach's α values for the subscales were 0.87 and 0.78, respectively. Convergent validity was confirmed through significant correlations with related theoretical constructs.
Conclusions:
The findings confirm that the Polish two-factor version of PSSI demonstrates satisfactory reliability and validity, suggesting its potential application in research.

Title

Development and Validation of the Sport Causality Orientations Scale (SCOS)

Author(s)
María J. Raimundi [1] , María F. Molina [2] , Isabel Castillo [3] , Mauro G. Perez-Gaido [4] , Octavio Álvarez [3] , Vanina I. Schmidt [5] , Inés Tomás [3]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Psicología Básica, Aplicada y Tecnología (Argentina); [2] Universidad Nacional de Tres de Febrero (Argentina); [3] Universitat de València (Spain); [4] Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional de Tres de Febrero (Argentina); [5] Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad de Buenos Aires (Argentina)

Abstract

Athletes often perceive contextual factors in sports differently, which influences their motivation and their tendencies to respond in specific ways, affecting their participation and sport experience. According to the Self-Determination Theory, these individual differences are known as causality orientations and they represent separate tendencies to: (a) focus on those features of the context that allow the individual to initiate and regulate their behavior in a self-determined way (autonomy orientation); (b) pay attention to environmental cues that provide the individual with rewards or punishments, regulating their behavior in a more controlled way (control orientation); (c) orient oneself towards obstacles that diminish one's potential to achieve one's goals, thereby considering oneself to be incompetent and without control over one's environment (impersonal orientation). As no instrument has been developed to measure this concept in the sports context, this presentation aims to introduce the development and validation of the Scale of Causality Orientations in Sport (SCOS).

The present study employs a vignette format, as previously utilized in related causality orientation scales, whereby situations are delineated and respondents are invited to indicate their likelihood of responding in alignment with each of the three causality orientations using a 7-point Likert scale. The generation of a pool of 20 situations resulted in a total of 60 items, with the situations being adapted from previous scales and new ones being created to represent relevant collective sporting events. The situations were designed to encompass events preceding, during, or following a competition or training session, during extended periods of physical exertion, and social interactions with teammates or coaches. The scale was then administered to five experts in the domain of SDT, specializing in the fields of sport sciences and sport psychology, who have experience working with adolescent athletes. These experts evaluated the scale's general instructions, situations, and items, offering comments and suggestions for improvement. The evaluation focused on the precision and breadth with which the relevant constructs were represented. Following these evaluations, the wording of some situations and items was modified, resulting in a scale with 18 situations (54 items) that was applied to 276 adolescent athletes (M = 16.14 years). During the administration of the scale with the participants, comments and questions were recorded. In general, the adolescents reported that the scale was easy to understand.

Using confirmatory factor analysis, a stepwise reduction of situations was conducted to ensure a good fit of the scale. The final version consists of seven situations (21 items) evaluating the three causality orientations: one before competition, one during, and two after; two situations related to practice and one addressing general motivation.

The psychometric properties of the SCOS were deemed acceptable, including factorial validity (R-CFI = .929; R-TLI = .907; R-RMSEA = .043; SRMR = .058) and internal consistency (omega-autonomy = .67; omega-control = .69; omega-impersonal = .70). A subsequent study provided further support for the validity of the SCOS, thereby confirming these initial findings. Consequently, the SCOS can be considered a valid and useful instrument for measuring causality orientations in the sport context.

Title

Psychometric validation of the Reported and Intended Behaviour Scale (RIBS) in the Spanish-speaking population

Author(s)
Hernán Sampietro [1] , Georgina Guilera [1] , Jorgina Taulé [1] , Maite Barrios

[1] University of Barcelona

Abstract

Background: Stigma is a significant obstacle to mental health recovery, is a barrier to seeking help and is linked to social exclusion. The Reported and Intended Behaviour Scale (RIBS) is one of the most widely used scales to evaluate stigma against people diagnosed with a psychiatric disorder. The RIBS consists of two parts. The first part has four items that are answered as "Yes", "No", and "Don't know", while second part has four items answered with a 5-option Likert scale. Although the RIBS has been used in some studies with Spanish-speaking populations, no prior research has systematically adapted and validated the scale for Spanish-speaking adults. This study aimed to fill this gap by gathering evidence of the validity and reliability of the RIBS to assess stigma in a Spanish cultural context.
Methods: A total of 361 adults (M = 33.5 years, SD = 15.0), the majority of whom were women (80.6%) and had completed secondary education (46.3%), participated in the study. The Spanish version of the RIBS was administered alongside the Attributional Questionnaire Abbreviated Version (AQ-14) and the Community Attitudes Toward the Mentally Ill Scale (CAMI). The scale' s dimensionality was examined using a confirmatory factor analysis. Internal consistency was assessed using McDonald's omega and Cronbach's alpha, and relationships with other variables through Spearman correlation coefficients.
Results: A 49.3% currently live or have lived with someone with a mental health problem, 50.1% currently work or have worked with such individuals, 44.6% currently have or have had a mental health problem themselves, 79.5% currently have or have had a close friend with a mental health condition. Factor analysis supported the one-factor structure of the RIBS, with factor loadings ranging from .80 to .91, and a good model fit. Reliability was excellent (ω = .89, α = .88). Additionally, the RIBS showed moderate to high correlations with the AQ-14 (r = .51) and the CAMI (r = .58).
Conclusion: The findings support the RIBS scores as valid and reliable for assessing stigma in the Spanish-speaking population. Its application can contribute to evaluating and improving interventions aimed at reducing stigma.

Title

Application of Exploratory Graph Analysis (EGA) to a scale of depressive symptomatology

Author(s)
Zaira Torres Romero [1] , José M. Tomás [1] , Mireia Abella [1] , Irene Fernández [1] , Aitana Sanz [1]

[1] Universidad de Valencia

Abstract

Background: Depressive disorders are one of the two most common mental disorders worldwide and highly prevalent among older adults. Therefore, its early detection is of outmost interest for older adults'healthy aging. Among the instruments employed to measure depressive symptomatology in the adult population, the EURO-D scale stands out as a short instrument that harmonizes pre-existing scales: the Geriatric Mental State-Automated Geriatric Examination For Computer Assisted Taxonomy (GMS-AGECAT), SHORT-CARE, the Center for Epidemiologic Studies Depression Scale (CES-D), the Zung Self-Rating Depression Scale (ZSDS), and the Comprehensive Psychopathological Rating Scale (CPRS). The EURO-D scale has been widely used and has multiple validation studies. However, its factor structure is still under debate and Exploratory Graph Analysis (EGA), a novel technique derived from network psychometrics, constitutes a promising new alternative to analyze its dimensionality. For that reason, the aim of this study is to explore the dimensionality of EURO-D using EGA.
Methodology: The sample comprised 46317 adults from the 8th wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). Participants were from 26 European countries and Israel, had an average age of 71.33 years (SD = 9.34) and 42.6% were male. The sample was randomly split into two subsamples: a derivation sample (n = 23,282) and a cross-validation sample (n = 23,035). A two-step strategy was followed. First, EGA was applied to the derivation sample to determine the underlying factor structure of the EURO-D. For this, two estimation methods were used: the Graphical Least Absolute Shrinkage and Selection Operator (glasso) and the Triangulated Maximally Filtered Graph (TMFG). Next, the resulting factor structure using each of the estimation methods was tested by means of Confirmatory Factor Analysis (CFA) in the cross-validation sample to assess model fit.
Results: The results of the EGA indicated a two-factor structure for the EURO-D, composed by "affective suffering"and "lack of motivation". Both glasso and TMFG estimation methods consistently identified this two-factor structure, with slight variations in the allocation of the suicidality and fatigue items. CFA results confirmed that both structures provided an adequate fit to the data. Results of the CFA based on the glasso EGA were: ⊠2(53) = 2472.71, p< .001, CFI= .941, RMSEA= .045, 90%CI [.044, .047], SRMR= .061. Results of the CFA based on the TMFG EGA were: ⊠2(53) = 2356.17, p< .001, CFI= .944, RMSEA= .044, 90%CI [.043, .046], SRMR= .061.
Conclusions: This work presents a combination of an exploratory technique, EGA, and a confirmatory technique, CFA, that has allowed providing additional evidence of the factor structure of a commonly used scale. The results support a two-factor structure on the EURO-D scale with alternative allocation of the fatigue and suicidality items. Present results are discussed against previous studies reporting two and three-factor solutions with different allocation of these items.

Title

The 20-item and 10-item Spanish versions of the Positive and Negative Affect Schedule (PANAS): Psychometric properties in a sample of nursing students

Author(s)

Javier Sánchez-Ruiz [1] , Michael A. West [2] , Gabriel Vidal-Blanco [1] , Noemí Sansó [3] , Juan Gómez-Salgado [4] , Philip Larkin [5] , Laura Galiana [1]

[1] Universitat de València; [2] Lancaster University; [3] Universitat de les Illes Balears; [4] Universidad de Huelva; [5] University of Lausanne

Abstract

Background: Hedonic well-being has been repeatedly related to health outcomes. Specifically, in samples of nurses, positive affect has contributed to explain stress, burnout and compassion satisfaction, whereas negative affect has been related to secondary traumatic stress. For hedonic well-being measurement, he Positive Affect and Negative Affect Schedule, in its long and short versions, is one of the most used instruments. Aim: The aim of this study is to present evidence on the internal structure of both the 20-item and 10-item Spanish versions of the PANAS in a sample of nursing students. Methods: Research took place at the University of Valencia and the University of the Balearic Islands (Spain). Participants were 925 nursing students, in the first year of the Nursing Degree. A sequence of models, including four confirmatory factor analyses, was hypothesized, estimated and tested, in the two versions: model 1, a one-factor model; model 2, a two-correlated factors model; model 3, a general factor with a method factor model; and model 4, a bifactor model. Evidence of reliability estimates for the best-fitting model, together with relations with other variables related to nurses'stress and burnout (which included hope, general-self-efficacy, resilient coping, optimism, and mindfulness), were also gathered. Results: Evidence pointed to an adequate fit of models 2, 3 and 4 in both the 20- and the 10-item versions. As models 3 and 4 showed negligible CFI differences when compared to the most parsimonious model 2, this latest was retained as the best representation of the data. Reliability estimates were adequate for both versions, and similar pattern of relations with other variables was found. Conclusions: The 20- and 10-item Spanish versions of the PANAS assesses two correlated factors of positive and negative affect, when applied in a sample of nursing students. Positive and negative affect, measured with this instrument, are related to hope, general self-efficacy, resilience, optimism, and mindfulness, key abilities for future prevention of nursing students and professional nurses'stress and burnout.

Title

Validation of the Spanish Gaming Transfer Phenomena Scale

Author(s)
Laura Maldonado-Murciano [1] , Angelica Ortiz de Gortari [2] , Georgina Guilera [3] , Juana Gómez-Benito [3] , Maite Barrios [3]

[1] International University of Catalonia; [2] University of Bergen; [3] University of Barcelona

Abstract

Game Transfer Phenomena (GTP) refers to the transfer of video game experiences into real life (i.e., altered sensory perceptions, mental processes and behaviours) (Ortiz de Gortari, 2019). In order to assess these experiences, the Game Transfer Phenomena Scale (GTPS) was developed (Ortiz de Gortari, Pontes, & Griffiths, 2015). The objective of this study is to validate the scale in a European Spanish-speaking population in terms of dimensional structure and internal consistency.
A sample of 412 gamers (54.61% women, mean age 25.48 years, SD = 11) participated. The GTPS was adapted from English to Spanish using three parallel translations and a reconciliation process. Item descriptive statistics were obtained. Confirmatory factor analysis was conducted to assess the scale's dimensionality, while Cronbach's alpha and McDonalds'omega were used to evaluate reliability. Correlations with problematic gaming and game session length were also examined. Analyses were carried out with the R packages lavaan and psych.
Item response distributions appeared to be right-skewed. The five-factor structure was confirmed (CFI = .996, SRMR = .016), with factor loadings ranging from .64 to .95. Cronbach's alpha values ranged from .75 and .84, and McDonalds'omegas between .77 and .84. The GTPS scores strongly correlated with measures of problematic gaming and session length.
The preliminary assessment of the psychometric properties of the European Spanish version of the GTPS shows that it is a suitable tool for measuring GTP in the Spanish-speaking population.

Title

Using machine learning on multiple true-false item texts to predict the difficulty of best single-answer items: Identifying domain-specific text features beyond readability

Author(s)
Lubomír Štěpánek [2] , Čestmír Štuka [1] , Martin Vejražka [1] , Patrícia Martinková [3]

[1] First Faculty of Medicine, Charles University; [2] Institute of Computer Science of the Czech Academy of Sciences; First Faculty of Medicine, Charles University; [3] Institute of Computer Science of the Czech Academy of Sciences; Faculty of Education, Charles University

Abstract

Accurately estimating item difficulty is crucial for designing fair and effective assessments, particularly in high-stakes exams such as medical faculty admissions. This study investigates subject-specific textual elements that significantly influence item difficulty beyond traditional readability features and explores the predictive potential of machine learning algorithms in estimating the difficulty of best single-answer items derived from multiple true-false items. Using historical admission test data from the First Faculty of Medicine, Charles University in Prague, we employ pre-calibrated difficulty estimates of multiple true-false items to predict the difficulty of their reformulated best single-answer counterparts.

Our approach goes beyond traditional textual features related to readability, such as word counts, vocabulary frequency, lexical similarity, and readability indices (Štěpánek, Dlouhá, & Martinková, 2023). Instead, we aim to leverage domain-specific contextual elements within item wording – particularly in subjects like physics, chemistry, and biology – that influence difficulty. These contextual and semantic elements include conceptual and knowledge representation features (such as domain-specific taxonomy or terminology abstractness), semantic embedding and contextual features (such as algorithm-estimated text complexity using large language models), syntactic and structural complexity (including text mode, sentiment density, and diction analyzed using language models), cognitive and conceptual load features (e.g., missing or aberrant information in the item wording), and domain-specific features (such as chemical or mathematical notation, formulas, or figures), among others. By adopting this approach, we seek to uncover key linguistic or conceptual patterns in item wording that strongly impact difficulty levels. Machine learning techniques are applied to identify these domain-specific difficulty-related textual and contextual features. The dataset includes multiple years of admission test responses, allowing us to match item wordings with test-takers'performance and apply the Rasch model for difficulty estimation. By comparing pre-calibrated multiple true-false item difficulties with the predicted and observed difficulties of their best single-answer versions, we evaluate the effectiveness of our approach in predicting difficulty shifts caused by item reformulation.

The findings of this study may contribute to the field of educational assessment by demonstrating how machine learning can enhance difficulty estimation, particularly when transitioning between item formats. The extracted textual features provide insights into the linguistic and cognitive factors influencing item difficulty, which can inform test construction and item design in high-stakes assessments.

References:
L. Štěpánek, J. Dlouhá, and P. Martinková, “Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms”, Mathematics, vol. 11, no. 19, p. 4104, Sep. 2023, issn: 2227-7390. doi: 10.3390/math11194104. [Online]. Available: http://dx.doi.org/10.3390/math11194104.

Title

Modeling difficulty of listening comprehension items with machine learning

Author(s)
Filip Martinek [2] , Patrícia Martinková [1] , Jan Netík [1]

[1] Institute of Computer Science of the Czech Academy of Sciences; Faculty of Education, Charles University; [2] Institute of Computer Science of the Czech Academy of Sciences

Abstract

Understanding the difficulty of items in educational assessment is crucial for test development, teaching, and learning. While prior research used text features and machine learning to model the difficulty of reading comprehension items (Štěpánek et al., 2023), listening comprehension (LC) presents additional complexity. Beyond textual factors, LC item difficulty can be influenced by several additional factors, including variations in speech rate, voice characteristics, and others.

In this study, we analyze data from English, German, and French listening sections of the Czech Matura exams. We extract text features commonly used in reading comprehension difficulty modeling, including readability indices and similarity measures. Additionally, we incorporate audio-specific features, such as speech rate, confidence scores from automatic speech recognition (ASR) models, and voice distinguishability derived from a speaker diarization model. This feature set allows us to capture the linguistic and auditory effects on LC item difficulty.

Using these features, we train a regression model to predict the final LC difficulty as described by student response data. The optimal model is selected based on accuracy measures of different models and model parameters, and compared to predictions of content experts to assess alignment between computational modeling and human evaluation.

The model is implemented into an interactive application to aid exam development. This tool aims to provide educators and test developers with insights into LC item difficulty, allowing for more informed test development.

Title

Measurement invariance and latent mean differences across Basque and Spanish versions of the Multidimensional Frailty Scale (MFS)

Author(s)

Arantxa Gorostiaga [1] , Joanes Lameirinhas [1] , Jone Aliri [1] , Igone Etxeberria [1] ,
Nekane Balluerka [1]

[1] University of the Basque Country UPV/EHU

Abstract

The concept of frailty was first introduced in medical literature to explain differences in health status among individuals of the same chronological age. Since then, it has evolved into a key concept in geriatrics and gerontology, fostering a growing body of research on its characteristics and implications. Frailty is generally defined as a state of increased vulnerability to external and internal stressors, significantly heightening the risk of adverse health outcomes such as falls, cognitive impairment, physical disability, hospitalization, and mortality. Although several instruments have been developed to assess frailty, most fail to encompass all its dimensions and lack satisfactory psychometric properties. Additionally, as stated in the guidelines of the International Test Commission (2017), comparative statements about respondent performance levels in two language groups should not be made unless measurement invariance has been established for the test scores being compared. Therefore, the aim of this study was to analyze the measurement invariance across languages of a new instrument called Multidimensional Frailty Scale, originally developed in Basque (Hauskortasun Multidimentsionalaren Eskala, HME) and Spanish (Escala de Fragilidad Multidimensional, EFM), to assess frailty in older adults, and to examine latent mean differences between both groups. The sample consisted of 484 individuals aged 65 and older (58.5% for the Spanish version and 41.5% for the Basque version). The MFS is a 29-item scale designed to assess multidimensional frailty, encompassing five dimensions: physical, cognitive, affective, social, and environmental frailty. Multi-group confirmatory factor analysis was used to evaluate the measurement invariance of the five EFM dimensions across languages. The results showed that the constrained model, with equivalent thresholds and factor loadings for Spanish and Basque respondents, demonstrated an adequate fit (CFI=0.984; TLI=0.984; RMSEA=0.039 [90% CI: 0.034-0.045]). Thus, measurement invariance across languages was successfully demonstrated. Additionally, latent mean differences indicated that participants who completed the Basque version of the instrument scored significantly lower in physical and affective frailty compared to those who completed the Spanish version. The magnitude of the difference between these means, calculated using Cohen's d, ranged from small to moderate. This study provides evidence of measurement invariance across the Spanish and Basque versions of the MFS, ensuring that frailty is assessed equivalently in both languages. These findings contribute to the validation of a robust assessment tool for frailty in aging populations, supporting its use in both clinical and research settings.

Title

Psychometric properties of the Brief Resilient Coping Scale (BRCS): A first study of its longitudinal measurement invariance in the Spanish context

Author(s)

Michael A. West [1] , Juan Gómez-Salgado [2] , Gabriel Vidal-Blanco [3] , Noemí Sansó [4] , Philip Larkin [5] , Javier Sánchez-Ruiz [3] , Laura Galiana [3]

[1] Lancaster University; [2] Universidad de Huelva; [3] Universitat de València; [4] Universitat de les Illes Balears; [5] University of Lausanne

Abstract

Background: Resilience has proved to be essential for nursing students due to their higher levels of academic stress. Among the developed tools for its measurement, the Brief Resilient Coping Scale (BRCS) is commonly used in the nursing context. However, no evidence on its longitudinal measurement invariance has been provided. Aim: The aim of this study is to present evidence on the psychometric properties, including its longitudinal invariance, of the BRCS in a sample of Spanish nursing students. Methods: A longitudinal design was used. Research took place at the University of Valencia and the University of the Balearic Islands (Spain). Participants were 257 nursing students, in the first (Wave 1) and second year (Wave 2) of the Nursing Degree. Analyses included descriptive statistics, reliability estimates, confirmatory factor analysis, and a longitudinal measurement invariance routine. Results: Evidence of reliability showed by the scale was adequate, and a one-factor solution for the structure was found in the two time occasions. Additionally, the BRCS showed evidence of invariance over time. Conclusions: The Spanish version of the BRCS is a brief instrument that can contribute to the assessment of resilience in nursing students, a key ability for both nursing students and professional nurses. Additionally, evidence on its longitudinal measurement invariance has been gathered. Therefore, the results of current research point to the adequacy of the Spanish version of the BRCS for assessing changes in nursing students'resilient coping, whether they are caused by educators'behaviors and methodology interventions.

Title

Response Options in Multiple-Choice Items: Increased Difficulty or Improved Reasoning?

Author(s)

Joaquín Juliani Aguado [1] , Manuel López Pavón [1] , Clara Patricia De Lacy Pérez de los Cobos [1] , María del Valle Jiménez Jaraba [1] , Ángel Manuel García-Carmona [1] , Samuel Ranz Castañeda [1] , Dulcinea Raboso Paniagua [1] , Fermín González Pereiro [1] , Pablo Ares-Gastesi [1] , Susana Sanz Velasco [1]

[1] Universidad San Pablo CEU

Abstract

Introduction: When writing multiple-choice items, the "None of the Above" (NOTA) option has been widely defended and criticized. Some authors argue that it artificially increases item difficulty and may even be detrimental because it negatively affects item discrimination, making it harder to distinguish between students with more knowledge and those who have not acquired it. However, other authors suggest that, apart from being an easy-to-create option, forcing students to confront this option, especially in items with non-numerical solutions, positively impacts their reasoning, particularly when NOTA is the correct option. This could lead to improved performance on similar or future tests. Therefore, the objective of this study is to assess the impact of this response option. Method: A sample of 443 students from various degree programs, such as Economics, Business Administration and Management, and Marketing, completed an inferential statistics test consisting of 12 items with three response options. Different versions of the same test were created to ensure that all items were answered in three formats: (1) a version with three distinct response options, (2) a version where NOTA was the correct answer, and (3) a version where NOTA was one of the distractors. Additional data on student performance, such as results from other tests in the same subject, were also collected. Comprehensive analyses will be conducted in R, first considering the Classical Test Theory (CTT) indicators for the scores of the test and the items, and later, performing some crossed random-effects multilevel models, to consider the items and students as two random factors or the model, to account the variability of both factors. Results: As preliminary results, the reliability of the obtained scores was calculated using Cronbach's alpha coefficient for the three models: Model A had a value of 0.638, Model B had a value of 0.406, and Model C had a value of 0.685. Given that the test consisted of only 12 items, the reliability coefficient was also calculated, and the Spearman-Brown formula was applied to determine the optimal test length for each version: 38 items for Model A, 68 for Model B, and 23 for Model C. Additionally, difficulty and discrimination indicators were calculated for each item. Regarding difficulty, the findings confirm that the inclusion of NOTA increases item difficulty, particularly when NOTA is the correct answer. In terms of discrimination, a general decrease was observed when NOTA was presented as a response option, with the lowest discrimination occurring when NOTA was the correct answer. However, this indicator depends on the specific model used, requiring further analysis to draw precise conclusions. Conclusion: The inclusion of NOTA appears to negatively impact item properties by reducing item discrimination. Regarding difficulty, it would be interesting to explore whether the increase in difficulty is artificial and therefore negative, or whether it enhances students' reasoning skills. This relationship will be further investigated with additional assessment tests on the subject.

Title

Location-Scale Models in Meta-Analysis: A Comparison of Analytic Frameworks

Author(s)
Manuel J Albaladejo-Sánchez [1] , Jose Antonio Lopez Lopez , Wolfgang Viechtbauer [2]

[1] University of Murcia, Murcia (Spain); [2] Maastricht University (Maastricht, The Netherlands)

Abstract

Location-scale models have recently been proposed for their use in meta-analysis, allowing the simultaneous testing of moderators of the mean (location) and variance (scale) of the distribution of true effects. They provide a powerful tool for evidence synthesis to address questions that have not yet been explored, and their implementation in the metafor R package makes them more accessible to applied meta-analysts. A simulation study conducted by our team found that, out of the different frequentist and likelihood-based methods currently available to researchers in metafor, profile-likelihood intervals and permutation tests generally perform better in terms of interval estimation and Type I error rates, respectively. However, a number of scenarios were identified where all methods under comparison showed a poor performance, sometimes even running into estimation problems. These issues were more frequent in conditions with a small to moderate number of studies, which are ubiquitous in practice, raising concern on the applicability of location-scale models in many applied scenarios.

Bayesian inference provides an interesting alternative to frequentist and likelihood-based methods, and the use of priors has the potential to solve some of the estimation problems and limitations which may arise when analysing small datasets. Location-scale models may also be fitted within a Bayesian framework using the brms R package. In this study, we used several example datasets to compare the results of the different analytic frameworks available nowadays to fit location-scale models in meta-analysis.

Title

Psychometric properties of the Scientific Reasoning Scale

Author(s)
Margherita Lanz [1] , Rossella Caliciuri [1]

[1] Università Cattolica del Sacro Cuore

Abstract
Scientific reasoning (SR) plays a crucial role in daily life, influencing individuals' abilities to formulate scientific questions, collect data, critically evaluate information, and draw informed conclusions. Currently, there are limited tools available for measuring SR that are tailored to specific age groups or countries, which restricts the comparability of results due to the unique characteristics of these tools. The Scientific Reasoning Scale (SRS), validated in both the United States and Turkey, assesses an individual's capacity to evaluate scientific evidence. To enhance generalizability across diverse populations, it is essential to validate this scale within different cultural contexts.

This study aims to validate the SRS in the Italian context using a representative sample of 897 Italian adults aged 18 and over. A survey was conducted through Qualtrics. This validation process, employing a unified view of validity, seeks to establish the scale's validity through various evidence sources: factorial structure, generalizability, convergent validity, criterion-related validity, known-group evidence, and reliability. To examine the factorial structure of the scale, we intend to use Structural Equation Modeling (SEM) within the framework of Classical Test Theory (CTT) and conduct Confirmatory Factor Analysis (CFA). Furthermore, to discuss the scale's generalizability, we plan to perform measurement invariance analyses to ensure that the scale holds the same meaning across different genders and age groups.

The results will facilitate an assessment of the adequacy of this instrument's adaptation and potentially address the existing gap in the evaluation of SR skills within the Italian context.

Title

Validation of the Spanish Nijmegen Gender Awareness in Medicine Scale (SN-GAMS) in Clinical Psychology Students

Author(s)
Francisco González Espejito [1] , Concepción Serrador-Díez [2] , Rebeca Pardo-Cebrián [2] , Irati Garrido-González

[1] Departamento de Psicología Biológica y de la Salud, Facultad de Psicología, Universidad Autónoma de Madrid; [2] Facultad de Ciencias Biomédicas y de la Salud, Universidad Europea de Madrid

Abstract

Background: Gender awareness in healthcare is essential for ethical and effective professional training. The Spanish Nijmegen Gender Awareness in Medicine Scale (SN-GAMS), previously validated in nursing, assesses gender sensitivity and gender role ideology. This study examines its psychometric properties in a sample of Spanish clinical psychology students. Method: The SN-GAMS was administered to 333 psychology students (80% women; M = 27,52; DT=7,470) in Spain using a 5-point Likert scale. It comprises three factors: Gender Sensitivity (GS; 12 items), Gender Role Ideology towards Patients (GRI-P; 11 items), and Gender Role Ideology towards Professionals (GRI-Pro; 6 items). Three items were removed following expert focus group evaluations. Due to non-normality, polychoric correlations were used. Dimensionality was assessed using parallel analysis with resampling, principal component extraction, the mean eigenvalue criterion, and the minimum average partial (MAP) index. Additionally, parametric bootstrap exploratory graph analysis (EGA) with 500 replications, graphical LASSO regularization, and the Louvain algorithm were applied. Confirmatory factor analysis (CFA) was conducted using robust weighted least squares estimation (WLSMV) following the original structure. Compared to previous studies, the statistical techniques applied in this research not only provide a more comprehensive analysis but are also better suited to the characteristics of the data. Results: The Kaiser-Meyer-Olkin index (KMO = 0.84) confirmed the suitability of the correlation matrix. Parallel analysis with polychoric correlations, more appropriate for asymmetric data, and EGA, suitable for moderately correlated factors, supported the three-factor structure. CFA showed good fit indices (RMSEA = 0.060; CFI = 0.981; TLI = 0.979), improving on previous validations. All items loaded above 0.40, explaining 47.3% of variance. Measurement invariance across gender was demonstrated at configural, metric, and scalar levels. Convergent validity was supported by significant correlations with the Ambivalent Sexism Inventory (ASI), where GS correlated negatively and GRI factors positively. Criterion validity analysis showed that students reporting gender-related training exhibited higher GS and lower GRI scores, aligning with theoretical expectations. Conclusion: This validation confirms the SN-GAMS as a reliable instrument for assessing gender awareness in psychology students. Findings highlight its relevance for ethical and inclusive training, particularly in clinical settings.

Title

The Madrid Loneliness Questionnaire (MLQ): Development and Validation in the Contemporary Sociocultural Context

Author(s)
Laura Esteban-Rodríguez [4] , Francisco González Espejito [5] , Gema Blasco-Novalbos [1] , Agustín Haro-León , Eduardo José Pedrero-Pérez [2] , Elisa Lillo-López [3] , Josselyn Victoria Sevilla-Martínez , Elena Díaz-Zubiaur

[1] Departamento de Evaluación, Calidad y Sostenibilidad, Unidad Técnica de Calidad e Investigación, Madrid Salud; [2] Departamento de Psicobiología, Universidad Nacional de Educación a Distancia; [3] Subdirección General de Prevención y Promoción de la Salud, Madrid Salud; [4] Instituto de Investigación Biomédica, Hospital 12 de Octubre; [5] Departamento de Psicología Biológica y de la Salud, Facultad de Psicología, Universidad Autónoma de Madrid

Abstract

Background: Loneliness has become a pressing issue in contemporary societies, particularly in urban settings where social transformations have redefined interpersonal relationships. Beyond its psychological distress, loneliness has been associated with significant health risks, including increased morbidity and mortality. Given rapid technological advancements, economic fluctuations, pandemics, and shifts in relationship dynamics, existing loneliness measures may not fully capture the nuances of the experience in modern contexts. This study presents the development and validation of the Madrid Loneliness Questionnaire (MLQ), designed to assess loneliness across adulthood and in alignment with contemporary sociocultural realities. Methods: The study utilized a general population sample comprising 1,526 participants aged 18 to 87 years (77.06% women). To assess structural validity, the sample was divided into two subsamples: one for exploratory factor analysis (EFA, n=623) and another for confirmatory factor analysis (CFA, n=903). Measurement invariance across gender was tested using multigroup CFA. Convergent validity was examined through correlational analyses with scores on the Perceived Stress Scale (PSS) and self-esteem measures. Internal consistency was evaluated via Cronbach's α and McDonald's ω coefficients. Results: The EFA and CFA supported a three-factor structure with excellent model fit indices (GFI=0.986; SRMR=0.05). The identified factors were Social Skills, Partner Relationships, and Emotional Isolation. Multigroup CFA confirmed configural, metric, and scalar invariance across gender, underscoring the scale's applicability in both male and female populations. The PSS showed significant positive correlations with the MLQ and its subscales, particularly moderate associations with the Withdrawal, Isolation, and Total Loneliness scores, while the Partner Relationships subscale showed weaker associations. Furthermore, negative correlations with self-esteem (Rho= -0.36 to -0.63) provided additional evidence of construct validity. The scale demonstrated excellent internal consistency (Cronbach's α=0.93, McDonald's ω=0.92). Conclusions: The MLQ emerges as a psychometrically robust instrument for assessing loneliness in the general population. Its demonstrated measurement invariance supports its use across gender groups, reinforcing its generalizability. The scale holds potential for further applications in diverse populations, including clinical cohorts and older adults. Future research should investigate the MLQ's incremental validity and predictive capacity for mental health outcomes, strengthening its role in prevention and intervention strategies targeting loneliness.

Title

An Empirical Network Analysis of the Carnism Structure based on the 4N Scale: Spanish and English Versions

Author(s)

Antonio J. Rojas Tejada [1] , Claudia Suárez-Yera [1] , María Sánchez-Castelló [1] , Orlando Dumitru Costin [1] , Javier Mayoral López [1]

[1] University of Almería

Abstract

Carnism is an ideology that justifies meat consumption using arguments known as the three Ns: eating meat is Natural, Necessary, and Normal. Subsequently, the "Nice"argument was added to these Ns since it was found in the literature that "eating meat is pleasant"was one of the main motives people feel motivated to give reason for and maintain their meat consumption. To measure this ideology, the 4N Scale was developed to measure Carnism. Our research team has conducted recent work based on factor analysis to identify the best theoretical framework for measuring Carnism's 4Ns. In this context, the 4N Scale showed improvement when using an alternative 3N (natural, necessary, and nice) framework, even compared to the 3N's original approach (natural, necessary, and normal). The rationale behind the alternative 3Ns approach arose from the concern about the methodological problems encountered in different studies in the Normal dimension, and the high factual content in the wording of the items belonging to this subscale. The present study aimed to analyze the structure of the Ns in the Spanish and English versions of the 4N Scale using empirical networks. The analysis focused on testing the network structure when the four initial dimensions (natural, necessary, normal, and nice) or the three dimensions alternatively proposed by the authors are introduced (natural, necessary, and nice), for the Spanish and English scale versions. The sample consisted of 265 Scottish and 272 Spanish participants. Data collection was carried out in parallel. Participants, who completed an online questionnaire, were obtained by convenience sampling. The analyses were performed by EBICglasso estimation and exploratory graphical analysis, to detect the number of substructures within the network. Centrality measures were considered to estimate the relative position of the items within the network. The resulting networks with the 4Ns in the Spanish and English versions were similar, but neither formed a unique structure for the Normal dimension. However, when the alternative 3N framework was used, the resulting networks formed distinct structures for each N in both versions. These results corroborate the findings of our team using factor analysis and imply further evidence that the best dimensional configuration for measuring carnism is composed of natural, necessary, and nice dimensions, since the Normal dimension does not contribute to the measurement of carnism as presented so far.

Title

Assessing the Performance of the Healthcare Access and Quality Index: A Methodological Challenge in Global Health Metrics

Author(s)
Tomislav Meštrović [1]

[1] University North, Croatia / Institute for Healthcare Metrics and Evaluation, US / University of Washington, School of Medicine, US

Abstract

The quantification of healthcare access and quality (HAQ) has long been a methodological challenge due to disparities in data availability, inconsistencies in measurement approaches, and also the complexity of disentangling health system performance from socioeconomic determinants. Here, the aim is to present an innovative and refined methodology for measuring the HAQ Index across 204 countries and territories in a 30-year time period. This approach marks a significant advancement in global health metrics by integrating mortality-to-incidence ratios (MIRs) and risk-standardized death rates (RSDRs) to isolate healthcare performance from other confounding variables.

This approach builds upon the Nolte and McKee concept of amenable mortality –deaths that should not occur given timely and effective medical intervention –by refining its operationalization. Unlike previous versions of the HAQ Index, which relied on principal component analysis (PCA) weighting schemes, here an arithmetic mean of scaled MIRs and RSDRs across 32 conditions is employed. Such methodological shift enhances interpretability while maintaining the robustness of prior iterations.

A novel contribution of this approach the age-specific stratification of the HAQ Index into three groups: young (0–14 years), working (15–64 years) and post-working (65–74 years). This allows for a more nuanced assessment of healthcare access and quality over the life course, addressing a very significant gap in previous global health assessments. Each group's HAQ Index was computed separately, ensuring that observed changes reflect real healthcare performance improvements rather than demographic shifts.

To further refine the metric, absolute convergence analysis was conducted to assess whether countries with initially lower HAQ scores exhibited faster improvements over time. This convergence analysis was performed using the Socio-Demographic Index (SDI), allowing for differentiation between improvements driven by healthcare access versus broader socioeconomic progress. In this approach, integrating MIRs and RSDRs minimizes biases from disease incidence, improving in turn accuracy. Age-stratified measurement addresses a major limitation of previous indices, revealing persistent disparities in healthcare access among working-age and older adults despite improvements for younger populations. The transition from PCA to an arithmetic mean scoring system enhances interpretability while maintaining robustness. Furthermore, absolute convergence analysis enables a longitudinal assessment of healthcare improvements in lower-performing regions. All these refinements create a more reliable and actionable framework for evaluating healthcare access and quality globally.

This methodological advancement can be used for benchmarking progress and informing health policy. More specifically, the refined HAQ Index holds significant potential for integration into universal health coverage assessments, cross-country performance comparisons, as well as targeted health system interventions –which is currently the need in public and global health. The demonstrated divergence in HAQ Index scores among working-age and older adults underscores the urgency of policy action to ensure equitable healthcare access across all age groups. Future iterations may also further refine risk-adjustment methodologies and expand to incorporate additional indicators reflecting primary care effectiveness and health system resilience.

Note: This work is presented on behalf of Global Burden of Disease Healthcare Access and Quality Collaborators. The analysis was funded by Bill & Melinda Gates Foundation.

Title

Addressing Overfactoring in Mixed-Worded Scales: An Exploratory Application of the Random Intercept Item Factor Analysis (RIIFA)

Author(s)
Giuliana Nasonte [1] , Palmira Faraci [1]

[1] Psychometrics Laboratory, Department of Human and Social Sciences, University Kore of Enna (UKE), Italy

Abstract

Method variance due to wording effects represents a critical challenge in psychometric research and measurement validity, often distorting factor structures and inflating dimensionality estimates. The wording effect stems from the inclusion of negatively oriented items and occurs when individuals provide inconsistent responses to positively worded (PW) and negatively worded (RW) items that are intended to reflect the same substantive dimension. Traditional dimensionality reduction techniques, such as Exploratory Graph Analysis (EGA) and Parallel Analysis (PA), tend to overestimate the number of factors when applied to mixed-worded scales, potentially compromising the interpretability of psychological measures. This study evaluates the effectiveness of the random intercept item factor analysis (RIIFA) in addressing this issue by isolating method variance at the exploratory stage. RIIFA introduces an additional latent variable—the random intercept factor—which can be conceptualized as a method factor capturing individual differences in the use of the response scale. By doing so, it helps separate substantive variance from systematic variance introduced by item wording, ultimately leading to a more accurate operationalization of the intended construct. Using a large UK sample (N = 977) who responded to the Short Grit Scale (Grit-S), we first performed a redundancy analysis to detect locally dependent items and then compared standard EGA and PA solutions with their RIIFA-based counterparts (i.e., riEGA and riPA), where a random intercept factor was specified with unit loadings for both PW and RW items and its variance freely estimated. To compare the robustness of the factor solutions, stability was further assessed through bootstrap analyses. Based on the redundancy analysis, one item was removed, resulting in the refined 7-item version of the scale (Grit-S7). The results confirmed that EGA and PA overestimated the number of factors, identifying one factor loaded by PW items and another by RW items, whereas both riEGA and riPA provided a unidimensional solution. Moreover, RIIFA-based techniques demonstrated greater stability across 5000 bootstrap resamples compared to their traditional counterparts. These findings highlight the potential of RIIFA as a valuable framework for mitigating method variance in exploratory analyses, offering a more accurate estimation of the substantive dimensionality. By effectively controlling for an approximate portion of variance caused by item wording, RIIFA reduces artificial factor proliferation and enhances structural validity, making it a promising approach for researchers dealing with mixed-worded scales.

3 Friday, 25 July 2025

3.1 Session 18 : ”Latent factors and errors in psychometric measurement”

Title

Factor Analysis under multimodal latent distributions: A simulation study

Author(s)
Guido Corradi [1] , Jesús Alvarado [2] , Víctor Ciudad [3] , Oscar Lecuona [2]

[1] Universitat de Illes Balears; [2] Complutense University of Madrid; [3] Universitat de València

Abstract
Conceptual framework: Factor analysis is one of the most popular techniques to estimate latent variables in social sciences, both in their exploratory and confirmatory branches. Among others, the factor model assumes that latent variables are unimodal and normally distributed. This assumption can be challenging when working with several populations and cultural groups due to potential multimodal distributions, which impedes fairness and generalizability of research findings.
Objectives: To investigate the impact of multimodal latent variables on factor analyses, specifically examining the potential bias in parameter estimation.
Methodology: This simulation study approaches this question by applying exploratory and confirmatory factor analyses with multimodal latent distributions. More concretely, we created several latent variables sorted in the severity of their multimodality, both in the number of modes to their distance between them.
Results: As multimodality increases, factor weights exhibit a heightened bias towards positive values, which may be an aberrant cue for researchers. Fit indices and parallel analyses showed also biases with specific patterns as multimodality gets more pronounced. Only factor scores showed up as reliable descriptors of the latent multimodality.
Implications: Latent multimodal distributions can be a relevant challenge in latent variable models, which underscore the importance of careful consideration and interpretation by researchers. We suggest more research on factor scores as potential reliable indicators of multimodal latent variables, while also the limitations and applications of these findings.

Title

Initial Attempts to Clean Up the Messy Middle Problem

Author(s)

Jimmy de la Torre [1] , Zechu Feng [1] , Lin Luo [1] , Xiaopeng Wu [1]

[1] University of Hong Kong

Abstract

Accurately measuring learning growth is a crucial exercise because it provides various stakeholders with a comprehensive picture of students'progress and school quality over time. Recently, learning progression (LR) has been introduced as a framework to understand what students know at multiple levels of learning across different time points. Although promising, as LP allows for the integration of learning sciences, modern psychometrics, and rigorous assessment design, many issues therein remain unsolved. Chief among them is the issue of the messy middle (MM) because it directly affects the validity of inferences from LR-based assessments. MM refers to the muddled item difficulties, as well as student abilities at the intermediate levels of the progression. Previous studies have shown that MM, which can be construed as departures from expected item difficulty ordering, and in some instances, item difficulty categories, can result from fitting item response theory (IRT) models that may be too simple for the data. Consequently, in this study, a more complex IRT model, specifically, the four-parameter logistic model (4PL), is explored as a means to address the MM problem. A simulation study involving 15 items, where items 1-5, 6-10, and 11-15 were categorized as easy, medium, and difficult, respectively, was conducted. Items 5, 6, 10, and 11 were designated as boundary items (i.e., items at the boundary of a difficulty category), and the rest as nonboundary items. Moreover, items 5 and 11 (6 and 10) are designated as outer (inner) boundary items. Nonboundary item responses were generated using the 1PL, whereas boundary item responses using the 4PL. The item difficulty parameters were uniformly spaced from -1.4 to 1.4; furthermore, the discrimination parameters (a) of all items, as well as the guessing parameters (g) of items 6 and 11, and slip parameters (s) of items 5 and 10, were manipulated. To examine in an ideal condition how fitting the 1PL to 4PL data produces switches in the boundary item difficulty categorizations, which presupposes switches in item difficulty parameter estimates, the sample size was fixed at 100,000. It was found that the guessing and slip parameters interact with the discrimination parameters to produce category switches. In particular, if the discrimination is low (i.e., a = 0.5), category switches occurred when the inner guessing or slip parameter was high (i.e., g6, s10 ≥ .18); if the discrimination is high (i.e., a = 2.0), category switches occurred when the outer slip or guessing parameter was moderately high (i.e., s5, g11 ≥ .10); however, when the discrimination is average (i.e., a = 1.0), category switches occurred only when s5, g11 ≥ .04 and g6, s10 ≥ .08. In general, category switches can be represented by four regions of a coordinate system, where the x-axis represents the discrimination parameter, and the y-axis the guessing or slip parameter. Depending on the region, outer or inner items or both can either switch or not switch categories. No switches in the item difficulty estimates, hence, no item difficulty category switches were observed when the 4PL was fitted to the data.

Title

Unpacking the Impact of Measurement Error and Outliers on Three-Way Interactions: Evidence from Monte Carlo Simulations with Best-Practice Recommendations

Author(s)

Jeremy Dawson [1] , Torsten Biemann [2] , Christina Andres [2]

[1] University of Sheffield; [2] University of Mannheim

Abstract

Three-way interaction models are powerful analytical tools for investigating how combinations of multiple factors influence outcomes across fields like management, psychology, and the social sciences. Unlike two-way interactions, which focus on how a single moderator affects the relationship between two variables, three-way interactions enable researchers to analyse the joint effects of three predictors. Statistically, three-way interactions are analysed using regression models that include the main effects of all three predictors, the pairwise interactions between each combination of predictors, and a three-way interaction product term created by multiplying the values of all three predictors. While effective, this approach is inherently sensitive to deviations in the data, as the multiplicative nature of product terms amplifies even small deviations in the predictors. This sensitivity makes three-way interaction models particularly vulnerable to issues that compromise data integrity. Consequently, our study focuses on measurement error and outliers, as these are among the most critical factors affecting the validity and reliability of interaction analyses.

Measurement error in the predictor variables becomes exacerbated in the product term, attenuating true effects and decreasing statistical power. Outliers distort interaction estimates disproportionately, as even moderate values in one or multiple predictors can result in extreme product term values due to the multiplicative logic of the product term. While two-way interactions may still yield meaningful results under such conditions, that might not be the case for three-way interactions. Our Monte Carlo simulations demonstrated that measurement error in the predictor variables significantly reduces the reliability of the product term, leading to diminished power, downward-biased regression coefficients, and smaller effect sizes. Outliers were found to exert disproportionate influence by creating extreme product term values, resulting in spurious interactions or obscured genuine effects. These challenges were exacerbated in smaller samples and smaller effect sizes.

To address these issues, we recommend tailored approaches: Power analyses that explicitly incorporate predictor reliabilities to ensure adequate sample sizes for detecting interaction effects, even with measurement error present in the data. Scale shortening, often used to address practical constraints like survey space or participant fatigue, should be approached with caution. Our findings indicate that reducing scales to the commonly used minimum of three items significantly undermines statistical power for three-way interactions, even for high sample sizes. For addressing outliers, diagnostic techniques focusing on the dynamics of product term are essential. Initial evaluations of product term distributions via histograms can identify extreme values, while targeted methods like DFBETAS effectively pinpoint data points exerting disproportionate influence on interaction regression coefficients. Unlike more general detection methods like Cook's Distance, DFBETAS directly assesses the impact of individual observations on parameters affected by outliers. Non-linear transformations and robust MM-regression were less effective in mitigating distortions caused by outliers in our simulations, and we therefore do not recommend their use for outlier diagnostics.

Overall, this study contributes to the methodological toolkit for analysing three-way interaction models, addressing critical gaps in the literature and equipping researchers with strategies to handle the challenges posed by measurement error and outliers.

Title

Added value of subscores: Can we accurately evaluate it?

Author(s)
Wilco Emons [1] , Maria Bolsinova [1] , Angelina Kuchina [1]

[1] Tilburg University

Abstract

In this presentation, we will examine the concept of subscore added value, a critical topic in educational and psychological testing, focusing on its evaluation. Subscores are often reported to provide more detailed feedback to test-takers and educators, but their usefulness depends on whether they add value beyond the total score.
Haberman (2008) introduced a criterion, stating that a subscore has added value if the squared correlation between the subscore and the true subscore exceeds that of the total score and the true subscore. Since this method requires parameter estimates from a sample, one crucial issue that needs to be considered is sampling variability. Even if the subscores have added value, the sample estimates of the squared correlations with the true subscore may suggest they do not, and vice versa.
Sinharay (2019) proposed using hypothesis testing to address this issue. However, he restated the hypotheses in terms of correlations between the observed subscore or total score, and a parallel-form subscore. Sinharay suggested using established statistical methods for testing dependent correlations, such as William's (1959) t and Olkin's (1967) Z statistics, to determine the significance of the difference between these correlations.
Nevertheless, the properties of the traditional statistics may not fully apply to the context of the added value of subscores. Both tests assume a trivariate normal distribution, but this assumption may not hold for discrete test scores. Moreover, these tests assume all variables are observed, while only correlations between observed (subscore or total score) and unobserved (parallel-form subscore) variables are available in this context. Finally, these correlations are derived from the assumption that the correlation between the subscores on parallel test forms equals the squared correlation between an observed and a true subscore. However, their sampling distributions differ, which may bias statistical conclusions.
Sinharay's (2019) results obtained from resampling an existing empirical dataset have some limitations. To accurately evaluate the performance of the proposed statistics, it is crucial to control for the true population parameters and the sampling mechanism, which is impossible in real-data simulations.
To address these gaps, we present findings from a comprehensive simulation study evaluating the accuracy of Olkin's Z and William's t statistics within Sinharay's (2019) parallel-form approach and original Haberman's (2008) method. Furthermore, a non-parametric bootstrap procedure is employed as an alternative for testing the significance.
The results reveal that the performance of Olkin's Z and William's t statistics within Sinharay's parallel-form approach is overly conservative, with low statistical power across all conditions. In contrast, applying these tests to Haberman's framework shows varied performance: inflation occurs at low subscore reliability, while high reliability results in conservative performance. These statistics, however, are the most powerful under all conditions. The non-parametric bootstrap procedure appears slightly conservative but shows promise for determining subscore added value. Nevertheless, it may face challenges in detecting small effects, particularly at low reliability levels.
These findings provide valuable insights into the practical application of subscore evaluation methods. Future work will investigate the underlying factors influencing the performance of these statistical methods.

Title

Addressing convergence problems in latent variable models

Author(s)
Marcos Jiménez [1] , Mauricio Garnier-Villarreal [1] , Vithor R. Franco [2]

[1] Vrije Universiteit Amsterdam; [2] São Francisco University

Abstract

Convergence problems are difficult to circumvent in psychometric analyses. Sometimes, these problems arise because models are wrongly specified but when this is not the case, convergence issues are due to slow and rigid optimization algorithms that are unable to set proper constraints over the parameter space. In the latent R package, we implemented a new optimization framework, termed optimization on matrix manifolds, where these problems are correctly addressed. latent is a user-friendly and flexible package capable of fitting a wide variety of latent variable models with guarantees of high-speed performance and convergence. This new package offers a unified approach towards different statistical frameworks such as Structural Equation Modeling (SEM) and Latent Class Analysis (LCA). In all of these frameworks, model fitting problems happen. For example, Ultra-Heywood cases (negative variances) and, more generally, non-positive-definite latent covariances are problematic in SEM. In these cases, the researcher either needs to be cautious in the interpretation of the parameter estimates and disregard standard errors or needs to set up an ad-hoc model in order to attain proper convergence. Meanwhile, local minima are frequent in LCA and optimization algorithms are usually slow. We present solutions to these problems. First, we implemented in latent an algorithm (the partially oblique manifold) for estimating covariance matrices that are strictly positive-semi-definite. This algorithm can even estimate matrices that are sparse (with many zeroes in specific positions). This way, researchers can estimate properly any model and extract standard errors as long as the model is well-specified. Second, we show how LCA models can be estimated very fast with multiple starting values in the latent package due to high-performing optimization algorithms written in C++. This manner, latent offers an alternative to fit models that otherwise cannot be fitted and saves time in the estimation of complex models.

Title

From Random Effects to Common Factors: Latent Dimensionality Assessment in Experimental Psychology

Author(s)
Ricardo Rey Sáez [1] , Javier Revuelta [2] , Miguel A. Vadillo [1]

[1] Universidad Autónoma de Madrid; [2] Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, 28049 Madrid

Abstract

One of the main objectives of psychometrics is to determine how many common factors underlie observed responses. When there is a theoretical basis for specifying the latent structure, assessing model fit and making model comparisons allow researchers to evaluate latent dimensionality. In the absence of such a theoretical framework, dimensionality can be examined using data-driven approaches such as Horn's parallel analysis or Exploratory Graph Analysis (EGA). These strategies, among others, are often used as a preliminary step before fitting an (unconstrained) exploratory factor model.

Although psychometricians have spent many years developing and refining methods for evaluating latent dimensionality, applying these methods in other fields is not always straightforward. A clear example is experimental psychology, where interest in the psychometric properties of experimental tasks has grown in recent years. In this context, researchers design and administer multiple tasks (e.g., Stroop, Flanker, Simon) that aim to measure the same underlying cognitive process (e.g., behavioral inhibition). To analyze these data, several authors have suggested fitting a linear mixed model in which the random effects (i.e., the model's latent variables) represent each participant's experimental effect (e.g., Stroop effect, Flanker effect, Simon effect). This approach allows researchers to quantify individual differences while preserving the experimental structure; however, it's limited to capturing correlations between experimental tasks and does not leverage the full potential of psychometric latent-variable decomposition.

Recently, Mehrvarz and Rouder (2024) introduced a hierarchical psychometric model designed to estimate the common factors underlying experimental effects from multiple tasks. While this represents a significant advancement, determining the number of latent common factors can be challenging, particularly when the latent structure is unclear. Fortunately, because experimental effects are latent variables, these common factors can be viewed as "second-order factors"in the psychometric field, and established methods do exist for determining their number. Building directly on the approach proposed by Jiménez et al. (2023), the present study extends their procedures to the experimental domain. Specifically, our workflow for assessing latent dimensionality involves: (1) fitting a linear mixed model, (2) extracting each participant' s experimental effects for each task, (3) estimating the correlations among these effects, and (4) applying dimensionality detection techniques to the resulting correlation matrix to identify the number of common factors.

We conducted a simulation study to assess the performance of this strategy, comparing parallel analysis (with different factor extraction methods) and EGA (with different unidimensionality detection methods and clustering algorithms). The simulations varied the number of latent dimensions, sample size, factor loadings, and the reliability of the experimental effects. Results indicate that, as reliability and factor loadings increase, both EGA with the Louvain algorithm and parallel analysis using principal components extraction show a close to perfect accuracy in detecting the correct number of latent dimensions. This method therefore provides a straightforward and effective alternative for determining how many latent dimensions underlie experimental tasks.

3.2 Session 5 : ”Bayesian methods in Psychology and Statistics”

Title

Where Psychometrics Meets Experimental Psychology: Bayesian Hierarchical Factor Models for Response Times

Author(s)
Ricardo Rey Sáez [1] , Miguel A. Vadillo [1] , Javier Revuelta [2]

[1] Universidad Autónoma de Madrid; [2] Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, 28049 Madrid

Abstract
Traditionally, experimental psychology has focused on examining differences between groups or experimental conditions. In recent decades, however, interest in individual differences has intensified. Yet, moving from an experimental approach to a psychometric framework is not straightforward because well-replicated experimental effects (e.g., the Stroop effect) often exhibit low reliability. As reliability decreases, the rank order of participants on a measured attribute becomes less consistent upon repeated measurement. This directly affects the correlation between two measures, as correlation reflects whether they similarly rank participants. Consequently, the probability of incorrectly concluding that two experimental measures are independent increases. This spurious independence can encourage substantive interpretations, even when the tasks were originally designed to measure the same underlying cognitive process. Thus, low reliability makes it difficult to determine (1) whether multiple processes are related and (2) whether multiple tasks measure the same cognitive process.
To address this issue, several experimental researchers have developed hierarchical models tailored to the unique features of experimental designs. Among them, the proposal by Mehrvarz and Rouder (2024) stands out as particularly innovative, as it directly aligns with classical psychometric models. These authors introduced a hierarchical model—implemented in JAGS—that recovers common factors across multiple response-time experimental tasks. Conceptually akin to exploratory factor analysis, their model assumes a Gaussian distribution for the dependent variable. Their simulation study demonstrates that this model provides a more accurate estimate of the true correlation among experimental measures than previous approaches.
Although the original model assumes a Gaussian distribution for response times, in practice these often follow markedly skewed distributions. The consequences of this specification error remain unclear, highlighting the need for alternative models that better reflect the empirical shape of response-time distributions. Building on Mehrvarz and Rouder's (2024) work, this study develops exploratory, confirmatory, and 'semi-exploratory' hierarchical psychometric models incorporating skewed response time distributions, such as the ex-Gaussian and shifted-lognormal. These models are efficiently implemented in Stan, minimizing estimation time. A simulation study was conducted to examine the impact of fitting Gaussian models to data generated from skewed distributions—more closely reflecting the empirical nature of response times—and to assess the performance of the appropriate skewed-distribution models. Results indicate that (1) using Gaussian models on skewed data systematically underestimates correlations and increases their uncertainty, and (2) skewed models yield unbiased and efficient parameter estimates. Critically, the combined effects of underestimation and uncertainty may lead researchers to conclude that a correlation is not significantly different from zero even when a true correlation exists in the population. In sum, skewed-distribution models provide a robust alternative for studying individual differences in response-time-based experimental measures, offering significant advantages over Gaussian models in accurately capturing correlations between experimental measures.

Title

Exploratory Bayesian Nonparametric Methods in Psychological Sciences

Author(s)
Giuseppe Mignemi [1] , Andrea Spoto [2] , Giovanni Bruno [2] , Anna Panzeri [2]

[1] Bocconi University; [2] University of Padova

Abstract

Discrete Bayesian nonparametric (BNP) priors have received increasing attention in recent decades. They characterize a broad class of flexible models in which inference is drawn under minimal distributional assumption and latent clusters are naturally accommodated by the structure of these priors. This makes the BNP models an appealing statistical solution for modeling individuals'heterogeneity in psychological sciences in which overdispersion of participants' variability in real data poses serious concerns. Moreover, the verification of distributional assumptions under the parametric models might not be straightforward and the estimates might be severely affected by model misspecification. The BNP models overcome these issues as they estimate the parameters'density distribution from the data under less strict assumptions. Despite the flexibility of these methods, their applications in psychological sciences have been poorly investigated. Their theoretical complexity and the difficult implementation of the algorithms for the posterior computation have made these priors less considered than other Bayesian solutions within the most common models for social sciences, such as multilevel models and Item Response Theory (IRT) models. Significant theoretical and methodological progress has recently been made, and the application of these methods is now more straightforward. We propose some real case examples in which discrete BNP priors are a convenient solution to explore latent individual similarities and prevent inferences from overdispersion and model misspecification issues. Limitations and future directions are discussed, and user-friendly tutorial codes are provided to enhance a more direct and practical use of these methods in applied social and psychological sciences.

Title

Estimating Context Effects in Small Samples while Controlling for Covariates: An Optimally Regularized Bayesian Estimator for Multilevel Latent Variable Models

Author(s)

Martin Hecht [1] , Valerii Dashuk [2] , Oliver Lüdtke [3] , Steffen Zitzmann [2] , Alexander Robitzsch [3]

[1] Helmut Schmidt University; [2] MSH Medical School Hamburg; [3] Leibniz Institute for Science and Mathematics Education

Abstract

We introduce a novel approach for estimating between-group effects in two-level latent variable models, specifically designed to address challenges associated with small sample sizes and low Intraclass Correlation Coefficients (ICCs). At the core of this method is a regularized Bayesian estimator, developed to minimize the Mean Squared Error (MSE) in estimating between-group effects by optimally balancing bias and variance. This approach is further extended to incorporate covariates, enabling a more generalized and robust estimator.

To facilitate the adoption of the regularized Bayesian estimator, we developed the MultiLevelOptimalBayes R package, tailored for researchers in the social sciences. The package offers extensive tools for implementing the proposed approach, including flexible model specifications. Key features include precise estimation of between-group effects, evaluation of covariate effects, and a novel balancing approach to create optimally balanced datasets from unbalanced data. Additionally, the package supports the use of a delete-d jackknife technique for obtaining standard errors.

This estimation approach not only advances statistical methodology but also equips researchers with practical tools for achieving more accurate results, especially in scenarios with limited data availability.

Title

Extending Bayesian Regularization Methods to Multiple-Group Mediation Analysis

Author(s)
Emma Somer [1] , Milica Miočević [1] , Carl F. Falk [1]

[1] McGill University

Abstract

Partial measurement invariance testing is a crucial prerequisite for comparing structural relationships across multiple groups. Recently, frequentist regularization approaches, such as the lasso and elastic net, have been extended to the measurement invariance framework and involve applying a penalty function to differences across item intercepts and loadings to improve the detection of non-invariant items. Despite their promising performance, frequentist regularization approaches may produce biased estimates and pose challenges for inference due to the unavailability of standard errors in some conditions. To address these limitations, Bayesian regularization methods, such as spike-and-slab priors (SSP), have recently been extended to the differential item functioning framework. Bayesian regularization methods rely on shrinkage priors that peak at zero and contain heavy tails, pulling small parameter differences toward zero while leaving large effects unchanged. This study builds on previous work by evaluating the performance of Bayesian regularization approaches in terms of the bias, efficiency, and coverage of the indirect effect in a multiple-group mediation analysis model. We compare Bayesian regularization approaches –small-variance normal priors, Laplace priors, Bayesian adaptive lasso, SSP, and horseshoe priors –against multiple-group CFA and alignment. We vary the sample size (N = 200 and 500), proportion (1/3 and 2/3) and magnitude (small or large) of non-invariance, and the value of the indirect effect (ab = 0 or 0.144) for a single mediator latent variable model with six indicators per latent variable. Preliminary findings suggest that Laplace priors and the adaptive lasso provide low bias and adequate coverage of the indirect effect under large proportions and magnitudes of non-invariance, whereas small-variance priors experience more difficulty. Additionally, adaptive priors produce biased latent means and intercepts under some conditions. The study provides recommendations for researchers estimating indirect effects in the presence of measurement non-invariance.

Title

What Do We Not Know About Small Sample Performance of AR(1) Models?

Author(s)

Eva Ceulemans [1] , Ginette Lafit [1] , Sigert Ariens [1] , Zhiwei Dou

[1] KU Leuven

Abstract

Single case experimental designs and experience sampling methods are state-of-the-art designs to study psychological processes. These designs yield time series of data, which are commonly analyzed using first-order autoregressive [AR(1)] modeling. In the AR(1) model, a variable is a function of its own value at the previous time point (i.e., autoregressive effect). It is crucial that researchers estimate the model parameters accurately. Nevertheless, estimating and performing inference on AR effects is not straightforward in small-sample settings. This is important for psychological scientists, who are faced with a number of constraints which make small-sample issues particularly relevant for the field. The Ordinary Least Square (OLS) estimator is the conventional and one of the most popular methods for estimating the autoregressive effect. In this talk, we will start with showing that OLS estimator is asymptotically consistent but downward biased due to the violation of the strict exogeneity assumption (Hamilton, 1994). We argue that alternative estimation methods do not resolve the exogeneity issue and thus fail to address small-sample bias. In addition, we demonstrate that in the AR(1) model, estimators of the intercept \hat{\alpha} and the AR effect \hat{\rho} are correlated, leading to bias in the intercept estimator when \alpha\neq0.

Consequently, small sample bias in the AR effect estimator can inflate power and increase Type I error in intervention studies, particularly when the study design is imbalanced. To help with a better power/Type I error rate, we propose two correction methods for improving the accuracy of OLS estimators in such settings.

Title

Multiple imputation of incomplete non-linear terms, when sample size is extremely small

Author(s)
Kristian Kleinke [1]

[1] Universität Siegen

Abstract

Handling incomplete non-linear terms in regression models like interaction terms (to test for moderating effects) is not trivial (Kleinke, 2021). One state-of-the-art approach to do so is factorized regression (e.g. Keller, 2022), also known under the name “substantive model compatible multiple imputation”. Imputations are generated under a model, which is fully compatible to the analyst’s model and to the assumed data generating process. The approach is implemented in various software packages including the Blimp software (Keller & Enders, 2023). The focus of the present study is to evaluate the factorized regression approach implemented in Blimp in a variety of scenarios where sample size was extremely small, ranging from 20 to 200 participants. I evaluated bias in point estimates and in measures of uncertainty of the incomplete non-linear effect in a 10 (sample size) by 11 (strength of the moderating effect) by 2 (missing data percentage) factorial experiment. Results show that point estimates were usually unbiased unless sample size was very small. Confidence interval (CI) widths increased with decreasing sample size. CI widths also increased with increasing missing data percentage, reflecting the increasing estimation uncertainty due to missing data. Especially for smaller samples and for smaller effects, this could mean power problems. Confidence interval coverage was widely accurate and centered around the nominal level. For very small samples, some over-coverage was observed, indicating that the obtained standard errors were too conservative (i.e. too large), which could result in type-II errors. If possible, applied researchers should try to obtain sample sizes of at least around 50, when the interaction effect of interest involves two continuous variables.

Keller, B. T. (2022). An introduction to factored regression models with Blimp. Psych, 4(1), 10-37. https://doi.org/10.3390/psych4010002

Keller, B. T., & Enders, C. K. (2023). Blimp user’s guide (Version 3). Retrieved from www.appliedmissingdata.com/blim

Kleinke, K. (2021). Estimation of partially observed non-linear terms in a multilevel model: An evaluation of the robustness of ad hoc and state-of-the-art missing data methods. Psychological Test and Assessment Modeling, 63(3), 432–455.

3.3 Symposium : ”Bridging Psychometrics and Artificial Intelligence”

Title

Optimizing LLM Embeddings for Automatic Item Development and Validation

Author(s)
Hudson Golino [1]

[1] University of Virginia

Abstract

Large Language Models (LLMs) have shown promise in text clustering and dimensionality analysis through embeddings, yet their potential for optimization remains largely unexplored. We conducted a comprehensive simulation study to enhance the accuracy of LLM embeddings in trait mapping using Dynamic Exploratory Graph Analysis (Dynamic EGA). The simulation generated 200 items across 4 traits of Narcissistic Personality, randomly selecting 3-40 items per dimension. We analyzed 1,040,000 combinations across 260 embedding values (3-1300) in a 1536-dimensional space. Performance was evaluated using Total Entropy Fit Index (TEFI) and Normalized Mutual Information (NMI). Vector field analysis revealed complex dynamics between TEFI and NMI, with optimal performance occurring in regions of moderate TEFI values and NMI above 0.5. The number of items per dimension showed peak performance between 10-20 items, while embedding dimensions exhibited non-linear relationships with both metrics. A weighted scoring system prioritizing NMI (70%) over TEFI (30%) significantly outperformed traditional cross-sectional embedding approaches. The optimization demonstrated improved accuracy in concept mapping while maintaining structural stability, suggesting a promising direction for enhancing LLM-based text analysis methods.

Title

Does the result of an in-silica structural validity matches the structural validity computed in human-gathered data?

Author(s)
Lara Russell-Lasalandra [1]

[1] University of Virginia

Abstract

The rapid advancement of large language models (LLMs) has enabled automated psychological scale development, yet questions remain about the correspondence between in-silica and human-gathered validation. This study examines whether structural validity metrics computed during automated item development match empirical validation results. Using AI-GENIE (Automatic Item Generation and Validation via Network-Integrated Evaluation), we generated Big Five personality items using five LLMs (Mixtral, Gemma 2, Llama 3, GPT-3.5, GPT-4). AI-GENIE performed in-silica structural validation during item generation and selection. These items were then administered to independent U.S. samples (N = 1000 per model). Comparing the in-silica and empirical structural validity metrics revealed strong correspondence (average correlation r = .89, RMSE = 0.08) across all models. Network invariance tests between in-silica and human-gathered data showed configural (NCT = 0.12, p > .05) and metric invariance (NCT = 0.15, p > .05). These findings suggest that AI-GENIE’s insilica structural validation effectively predicts empirical structural validity, supporting its use in automated scale development.

Title

How Block Types and Social Desirability Shape Forced-Choice Questionnaire Automatic Assembly

Author(s)

Francisco José Abad García [1] , Miguel A. Sorrel [1] , Rodrigo Schames Kreitchmann [2] ,
Scarlett Escudero [1]

[1] Universidad Autónoma de Madrid; [2] National University of Distance Education

Abstract

The construction of forced-choice questionnaires often relies on item banks with single-stimulus or Likert-type items. In its simplest form, items must be paired to create a desired number of blocks. A key challenge in this process is pairing items while accounting for factors such as item polarity and social desirability, which can impact the quality of the measures. Recent combinatorial approaches, like genetic algorithms, leverage parametric optimization based on Likert data estimates. Alternatively, blueprint-based methods enable block assembly without such estimates, integrating expert judgments on social desirability. However, these approaches have yet to be systematically compared, which is the primary goal of this study. A Monte Carlo simulation and empirical analysis were conducted to compare block assembly using the genetic algorithm and blueprint-based methods, with and without considering social desirability. The main outcome of interest was trait score recovery. Four key factors were manipulated to assess their influence on this outcome: the number of heteropolar blocks, questionnaire length, the inclusion of social desirability ratings, and the correlation between social desirability and single-stimulus item parameters. Results indicate that parametric methods generally lead to superior trait score recovery, especially when only homopolar blocks are used or when social desirability is factored in—conditions commonly found in applied settings. These findings highlight the importance of optimizing assembly procedures. We also discuss how expert judgments can serve as proxies for item parameters, enabling efficient block assembly in the absence of empirical data on single-stimulus items.

Title

Predicting Item Response Theory Parameters from the Semantic Space of Computational Language Models

Author(s)

Diego Iglesias [1] , Francisco José Abad García [1] , Miguel A. Sorrel [1] , Ricardo Olmos [1]

[1] Universidad Autónoma de Madrid

Abstract

Parallel to the development of new technologies, computational language models have emerged as automated tools for analyzing semantic relationships between linguistic units. Due to their success in performing human-like tasks, such as vocabulary tests and sentiment analysis, interest in the practical applications of these models has grown exponentially, resulting in the development of larger models with enhanced predictive capabilities.

In this study, we examine whether the high-dimensional semantic space underlying computational language models, such as ChatGPT, can be used to predict item parameters. In ChatGPT, linguistic units are represented as n-dimensional embedding vectors, which can be manipulated through mathematical operations.

We extracted embeddings for an item pool of 220 items from an English vocabulary test. The loadings of each item in ChatGPT's 1536-dimensional space were used as independent variables to predict their corresponding item response theory item parameters. The predictive accuracy of various machine learning models was evaluated using cross-validation procedures and compared with human-expert ratings. Despite the relatively small size of the training set, preliminary results are promising (^2_=0.40). We discuss the potential of using larger datasets for training the predictive model and the promising role of generative artificial intelligence in creating large item pools with desirable psychometric properties at minimal cost.

3.4 Symposium : ”Understanding, detecting and managing careless responding in survey research.”

Symposium Overview

Understanding, detecting and managing careless responding in survey research.

Author(s)
Clara Cuevas Ureña , Ana Hernández [1] , Inés Tomás [1] , Anna Brown , Esther Ulitzsch [2] , Vicente González-Romá

[1] University of Valencia; [2] University of Oslo

Abstract
Careless and insufficient effort responding (C/IER) occurs when individuals do not pay sufficient attention to item content. This threatens the validity of measurement and research conclusions. This symposium presents state-of-the-art approaches to understanding, detecting, and managing C/IER in self-report data. Specifically, it examines both simulated and empirical data and focuses on different item formats, including Likert and ipsative (e.g. forced-choice) formats. The first presentation investigates the stability of C/IER over time, addressing whether it should be considered a stable trait or a transient state. Using longitudinal data from university students, the study examines C/IER patterns identified through Instructed Response Items (IRIs) and explores whether distinct subpopulations display stable or changing response behaviors. The second presentation compares different C/IER detection methods, contrasting attention check items (i.e. IRIs) with a model-based mixture IRT approach that does not require additional items. The effectiveness of these methods and their implications for data quality are discussed. The third presentation uses simulated data to show how different strategies for handling C/IER affect the psychometric properties of scales. It compares doing nothing regarding C/IER, removing careless respondents, treating C/IER as a control variable, and using it as a moderator variable. Finally, the fourth presentation examines two strategies for addressing C/IER in ipsative data. The first strategy identifies and removes careless respondents using “person fit”statistics, while the second controls for C/IER using method factors designed for Thurstonian IRT and factor models. Together, these four studies contribute to advancing best practices in survey data quality.

This symposium has been partially supported within the research project PID2022-141339NB-I00, funded by MCIU /AEI /10.13039/501100011033 / and by FEDER A way to make Europe, EU

Title

Testing the stability of careless responding over time

Author(s)

Inés Tomás, Ana Hernández, Clara Cuevas, Vicente González-Romá

Abstract

Background. Careless responding (CR) occurs when individuals do not pay adequate attention to item content. Research has shown that CR introduces bias and compromises data quality (Podsakoff et al., 2012), highlighting the need for effective prevention and management strategies (e.g., Arthur et al., 2021; Edwards, 2019; Ward & Meade, 2022). Different methods have been proposed to detect CR, one of them being Instructed Response Items (IRIs), which direct participants to provide specific answers. Failing these items serves as an indicator of CR. The use of IRIs stands out for its simplicity, transparency, and metric properties (Kam & Chan, 2018). Despite its significance, the nature of CR remains unclear. While some researchers consider CR as a stable trait (Meade & Craig, 2012), others argue it is transient state (Maniaci & Rogge, 2014). However, little empirical evidence has clarified this distinction. A recent study by Tomás et al. (2024), conducted with a sample of adult workers who were paid for their participation in the study, identified subpopulations with distinct CR patterns, some displaying stable CR behaviors, while others exhibited changes over time.

Objectives. This study aims to deepen the understanding of CR's nature and dynamics by analyzing its patterns over time in a sample with different sociodemographic characteristics (university students) and with different contextual factors (individuals were not financially compensated for their participation). Additionally, we examine whether CR operates as a trait or state for the entire population or if distinct subpopulations exist, some for whom CR is a trait and others for whom it is a state. To detect CR, we utilize IRIs.

Methods. A total of 360 Spanish university students (71.7% women; mean age = 25.6 years, SD = 6.3) participated in the study after being offered a free face-to-face training course. We used a within-subject longitudinal design with three data collection points, spaced at 3-month intervals. Participants were first contacted during their final semester (T1), approximately one month before graduation, followed by assessments nine months post-graduation (T2), and four months after T2 (T3). The trajectory of CR over time was modeled using latent growth modeling (LGM), and latent class growth analysis (LCGA) in Mplus.

Results and Conclusions. The results aligned with previous research (e.g., Tomás et al., 2024): while CR exhibited a stable response pattern over time at the population level, distinct subpopulations emerged, each displaying different CR trajectories. Notably, the subgroups identified in this study differed from those found by Tomás et al. (2024). In this study, three distinct subpopulations emerged: a relatively stable group (careful individuals) and two groups whose inattentiveness increased over time (one initially careful but becoming less attentive and another already careless that became even more inattentive). These findings contribute to the understanding of CR's nature and dynamics, highlighting the role of personal factors (e.g., age) and contextual factors (e.g., participation compensation) in shaping CR patterns over time.

This study has been developed within the research project PID2022-141339NB-I00, funded by MCIU /AEI /10.13039/501100011033 / and by FEDER A way to make Europe, EU

Symposium title

Understanding, detecting and managing careless responding in survey research

Title

Detecting careless and insufficient effort responding: A comparison of attention check and model-based approaches

Author(s)

Esther Ulitzsch, Ana Hernández, Inés Tomás, Clara Cuevas

University of Oslo and University of Valencia

Abstract

Background. Careless and insufficient effort responding (C/IER) on self-report measures produces responses that fail to accurately reflect the trait being measured, posing a major threat to the quality and validity of survey data. While detecting C/IER is vital to ensure validity of conclusions drawn from self-report data, it is a non-trivial endeavor, with each detection method involving distinct assumptions and limitations.

Objectives. This study compares two prominent approaches for C/IER identification and adjustment based on respondent behavior: (1) attention check items, which offer clear interpretability but require careful and parsimonious administration, limiting their ability to monitor C/IER comprehensively, and (2) a model-based mixture IRT approach, which avoids the need for additional items but relies on strong assumptions about respondent behavior.

Methods. Using data from five scales of a job quality survey completed by 707 respondents, we fitted an explanatory mixture IRT approach by means of R.

Results and Conclusions. We observed strong alignment between the two approaches: respondents identified as less attentive by one method were similarly flagged by the other. Overall, both approaches suggested that C/IER remained relatively stable throughout the course of the questionnaire. However, single attention check items consistently indicated substantially lower levels of C/IER at multiple points throughout the questionnaire compared to the scale-level C/IER rates implied by the model-based approach. Both methods had comparable impacts on adjusted psychometric properties. While correlations between latent constructs did not differ markedly from their unadjusted counterparts, adjusted trait estimates were less reliable, especially when obtained using the model-based approach, reflecting greater uncertainty in respondents' trait levels. Implications for C/IER identification and adjustment are discussed, arguing for a triangulation of different approaches.

This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 33160 and by the Spanish project PID2022-141339NB-I00, funded by MCIU /AEI /10.13039/501100011033 / and FEDER A way to make Europe, EU

Symposium title

Understanding, detecting and managing careless responding in survey research

Title

Detecting and managing careless and insufficient effort responding: A simulation approach.

Author(s)

Clara Cuevas, Inés Tomás, Ana Hernández

University of Valencia

Abstract

Background. Careless and insufficient effort responding (C/IER) occurs when respondents fail to give sufficient attention to item content, which leads to poor-quality data (Podsakoff et al., 2012). There are several methods to detect this phenomenon, one being Instructed Response Items (IRI), valued for its simplicity, robust metric properties, and ability to identify different C/IER patterns (Kam & Chan, 2018). While detecting C/IER is a crucial first step, deciding how to address this phenomenon once identified is equally important, as this choice can determine the extent of its impact on data quality.
Objectives. This study compares four strategies for managing C/IER and their impact on the psychometric properties of questionnaires, specifically reliability and validity evidence based on the internal structure: (1) using the total sample without adjustments, (2) excluding careless respondents to create a “clean”sample, (3) retaining the total sample while treating C/IER as a control variable, and (4) retaining the total sample while treating C/IER as a moderating variable.
Methods. We use simulated data based on the Big Five Questionnaire (Caprara et al., 1993) and the Maslach Burnout Inventory (Maslach & Jackson, 1981). A total of 180 conditions are manipulated, with variations in variables such as Severity of C/IER (25%, 50%, 75%, 100%), Percentage of C/IER (0%, 8%, 24%), or Sample Size (150, 300, 700). For each condition, 100 replications are run.
Expected results and Conclusions. Based on previous studies with empirical data (Tomás et al., 2023), we anticipate that using C/IER as a moderating variable (4) will be the most effective strategy. In contrast, using the total sample without adjustments (1) will likely be the least effective, given that C/IER is ignored. Regarding the exclusion of careless respondents (2), we anticipate a reduction in statistical power and its subsequent impact on the psychometric properties. As for (3) using C/IER as a control variable, based on previous empirical research examining its impact on questionnaire psychometric properties (Tomás et al., 2023) and substantive research models results (Tomás et al., 2025), we expect this strategy to be a less effective approach for addressing CR. We will provide recommendations for managing C/IER, helping to mitigate its impact on data quality in applied research.

This study has been developed within the research project PID2022-141339NB-I00, funded by MCIU /AEI /10.13039/501100011033 / and by FEDER A way to make Europe, EU

Symposium title

Understanding, detecting and managing careless responding in survey research

Title
Detecting careless responding in ipsative data

Author(s)
Anna Brown

University of Kent

Abstract

Background. To prevent response styles associated with the use of rating scales, test items may be presented in so-called ipsative (or relative to self) formats including popular 'forced choice', and also 'graded preferences'or 'proportions-of-total'. Like any other questionnaires, ipsative questionnaires can be subject to careless responding when respondents are not sufficiently motivated to give their full attention to the questions. However, detecting such responding can be more challenging than when using Likert scales because ipsative response formats usually involve comparisons between items measuring different traits and their modelling is inherently multidimensional. Moreover, the comparative nature of ipsative responses makes challenging the use of a method factor (latent variable) to control careless responding.
Objectives. This presentation will describe and evaluate two alternative strategies for dealing with careless responses in ipsative data: (1) identifying (and ultimately removing from the sample) careless responders using "person fit"indices designed for ipsative formats; and (2) controlling for careless responding using method factors specifically designed for Thurstonian IRT and factor models (Brown & Maydeu-Olivares, 2012).
Methods. The two approaches are illustrated on a sample of N=504 paid Prolific participants in a trial of the Leadership Styles Questionnaire (LSQ) measuring 24 personal styles with 88 multidimensional graded triplets. Under Approach 1, two "person fit"indices were computed for each respondent. The first index summarized the discrepancies between a person's observed responses and responses expected under the fitted Thurstonian measurement model, thus resembling the lco index (Ferrando, 2010). The second index summarized the concordance between a person's observed and expected responses by computing a correlation coefficient between them. Under Approach 2, a random intercept was added to the Thurstonian measurement model to control carelessness expressed as overusing one rating scale category.
Results and Conclusions. The concordance index had a median of 0.572 and a long left tail, identifying at least 8% of aberrant responders. The discrepancy index had a median of 0.820 and a long right tail, again identifying at least 8% of aberrant responders. The Thurstonian model with the random intercept factor fitted better than the baseline model (SRMR were .058 and .075, respectively), and the random intercept explained between 1% and 2% in the variances of observed responses. However, at the individual level the discrepancy, concordance and random intercept agreed only for careful responders. For careless responders, each index provided unique information about the nature of carelessness. We conclude with recommendations for the use of such indices in practice.

Symposium title
Understanding, detecting and managing careless responding in survey research

Title

Testing the stability of careless responding over time

Author(s)
Ana Hernández , Clara Cuevas Ureña , Inés Tomás Marco [1] , Vicente González Romá

[1] University of Valencia

Abstract

Background. Careless responding (CR) occurs when individuals do not pay adequate attention to item content. Research has shown that CR introduces bias and compromises data quality (Podsakoff et al., 2012), highlighting the need for effective prevention and management strategies (e.g., Arthur et al., 2021; Edwards, 2019; Ward & Meade, 2022). Different methods have been proposed to detect CR, one of them being Instructed Response Items (IRIs), which direct participants to provide specific answers. Failing these items serves as an indicator of CR. The use of IRIs stands out for its simplicity, transparency, and metric properties (Kam & Chan, 2018). Despite its significance, the nature of CR remains unclear. While some researchers consider CR as a stable trait (Meade & Craig, 2012), others argue it is transient state (Maniaci & Rogge, 2014). However, little empirical evidence has clarified this distinction. A recent study by Tomás et al. (2024), conducted with a sample of adult workers who were paid for their participation in the study, identified subpopulations with distinct CR patterns, some displaying stable CR behaviors, while others exhibited changes over time.
Objectives. This study aims to deepen the understanding of CR's nature and dynamics by analyzing its patterns over time in a sample with different sociodemographic characteristics (university students) and with different contextual factors (individuals were not financially compensated for their participation). Additionally, we examine whether CR operates as a trait or state for the entire population or if distinct subpopulations exist, some for whom CR is a trait and others for whom it is a state. To detect CR, we utilize IRIs.
Methods. A total of 360 Spanish university students (71.7% women; mean age = 25.6 years, SD = 6.3) participated in the study after being offered a free face-to-face training course. We used a within-subject longitudinal design with three data collection points, spaced at 3-month intervals. Participants were first contacted during their final semester (T1), approximately one month before graduation, followed by assessments nine months post-graduation (T2), and four months after T2 (T3). The trajectory of CR over time was modeled using latent growth modeling (LGM), and latent class growth analysis (LCGA) in Mplus.
Results and Conclusions. The results aligned with previous research (e.g., Tomás et al., 2024): while CR exhibited a stable response pattern over time at the population level, distinct subpopulations emerged, each displaying different CR trajectories. Notably, the subgroups identified in this study differed from those found by Tomás et al. (2024). In this study, three distinct subpopulations emerged: a relatively stable group (careful individuals) and two groups whose inattentiveness increased over time (one initially careful but becoming less attentive and another already careless that became even more inattentive). These findings contribute to the understanding of CR's nature and dynamics, highlighting the role of personal factors (e.g., age) and contextual factors (e.g., participation compensation) in shaping CR patterns over time.
This study has been developed within the research project PID2022-141339NB-I00, funded by MCIU /AEI /10.13039/501100011033 / and by FEDER A way to make Europe, EU

Title

Detecting and managing careless and insufficient effort responding: A simulation approach

Author(s)
Ana Hernández , Clara Cuevas Ureña , Inés Tomás Marco [1]

[1] University of Valencia

Abstract

Background. Careless and insufficient effort responding (C/IER) occurs when respondents fail to give sufficient attention to item content, which leads to poor-quality data (Podsakoff et al., 2012). There are several methods to detect this phenomenon, one being Instructed Response Items (IRI), valued for its simplicity, robust metric properties, and ability to identify different C/IER patterns (Kam & Chan, 2018). While detecting C/IER is a crucial first step, deciding how to address this phenomenon once identified is equally important, as this choice can determine the extent of its impact on data quality.
Objectives. This study compares four strategies for managing C/IER and their impact on the psychometric properties of questionnaires, specifically reliability and validity evidence based on the internal structure: (1) using the total sample without adjustments, (2) excluding careless respondents to create a “clean”sample, (3) retaining the total sample while treating C/IER as a control variable, and (4) retaining the total sample while treating C/IER as a moderating variable.
Methods. We use simulated data based on the Big Five Questionnaire (Caprara et al., 1993) and the Maslach Burnout Inventory (Maslach & Jackson, 1981). A total of 180 conditions are manipulated, with variations in variables such as Severity of C/IER (25%, 50%, 75%, 100%), Percentage of C/IER (0%, 8%, 24%), or Sample Size (150, 300, 700). For each condition, 100 replications are run.
Expected results and Conclusions. Based on previous studies with empirical data (Tomás et al., 2023), we anticipate that using C/IER as a moderating variable (4) will be the most effective strategy. In contrast, using the total sample without adjustments (1) will likely be the least effective, given that C/IER is ignored. Regarding the exclusion of careless respondents (2), we anticipate a reduction in statistical power and its subsequent impact on the psychometric properties. As for (3) using C/IER as a control variable, based on previous empirical research examining its impact on questionnaire psychometric properties (Tomás et al., 2023) and substantive research models results (Tomás et al., 2025), we expect this strategy to be a less effective approach for addressing CR. We will provide recommendations for managing C/IER, helping to mitigate its impact on data quality
in applied research.

This study has been developed within the research project PID2022-141339NB-I00, funded by MCIU /AEI
/10.13039/501100011033 / and by FEDER A way to make Europe, EU

Title

Detecting careless responding in ipsative data

Author(s)
Anna Brown

Abstract

Background. To prevent response styles associated with the use of rating scales, test items may be presented in so-called ipsative (or relative to self) formats including popular 'forced choice', and also 'graded preferences'or 'proportions-of-total'. Like any other questionnaires, ipsative questionnaires can be subject to careless responding when respondents are not sufficiently motivated to give their full attention to the questions. However, detecting such responding can be more challenging than when using Likert scales because ipsative response formats usually involve comparisons between items measuring different traits and their modelling is inherently multidimensional. Moreover, the comparative nature of ipsative responses makes challenging the use of a method factor (latent variable) to control careless responding.
Objectives. This presentation will describe and evaluate two alternative strategies for dealing with careless responses in ipsative data: (1) identifying (and ultimately removing from the sample) careless responders using "person fit"indices designed for ipsative formats; and (2) controlling for careless responding using method factors specifically designed for Thurstonian IRT and factor models (Brown & Maydeu-Olivares, 2012).
Methods. The two approaches are illustrated on a sample of N=1,338 volunteers who participated in a trial of an assessment measuring 24 non-cognitive skills with 276 multidimensional graded response pairs. Under Approach 1, two "person fit"indices were computed for each respondent. The first index summarized the discrepancies between a person's observed responses and responses expected under the fitted Thurstonian measurement model, thus resembling the lco index (Ferrando, 2010). The second index summarized the concordance between a person's observed and expected responses by computing a correlation coefficient between them. Under Approach 2, a random intercept was added to the Thurstonian measurement model to control carelessness expressed as overusing one rating scale category.
Results and Conclusions. The concordance index had a median of 0.532 and a long left tail, identifying at least 10% of aberrant responders. The discrepancy index had a median of 0.770 and a long right tail, again identifying at least 10% of aberrant responders. The Thurstonian model with the random intercept factor fitted better than the baseline model (SRMR were .055 and .068, respectively), and the random intercept explained between 1% and 2% in the variances of observed responses. However, at the individual level the discrepancy, concordance and random intercept agreed only for careful responders. For careless responders, each index provided unique information about the nature of carelessness. We conclude with recommendations for the use of such indices in practice.

Title

Detecting careless and insufficient effort responding: A comparison of attention check and model-based approaches

Author(s)
Ana Hernández , Clara Cuevas Ureña , Esther Ulitzsch [1] , Inés Tomás Marco [2]

[1] University of Oslo; [2] University of Valencia

Abstract

Background. Careless and insufficient effort responding (C/IER) on self-report measures produces responses that fail to accurately reflect the trait being measured, posing a major threat to the quality and validity of survey data. While detecting C/IER is vital to ensure validity of conclusions drawn from self-report data, it is a non-trivial endeavor, with each detection method involving distinct assumptions and limitations.
Objectives. This study compares two prominent approaches for C/IER identification and adjustment based on respondent behavior: (1) attention check items, which offer clear interpretability but require careful and parsimonious administration, limiting their ability to monitor C/IER comprehensively, and (2) a model-based mixture IRT approach, which avoids the need for additional items but relies on strong assumptions about respondent behavior.
Methods. Using data from five scales of a job quality survey completed by 707 respondents, we fitted an explanatory mixture IRT approach by means of R.
Results and Conclusions. We observed strong alignment between the two approaches: respondents identified as less attentive by one method were similarly flagged by the other. Overall, both approaches suggested that C/IER remained relatively stable throughout the course of the questionnaire. However, single attention check items consistently indicated substantially lower levels of C/IER at multiple points throughout the questionnaire compared to the scale-level C/IER rates implied by the model-based approach. Both methods had comparable impacts on adjusted psychometric properties. While correlations between latent constructs did not differ markedly from their unadjusted counterparts, adjusted trait estimates were less reliable, especially when obtained using the model-based approach, reflecting greater uncertainty in respondents'trait levels. Implications for C/IER identification and adjustment are discussed, arguing for a triangulation of different approaches.

This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 33160 and by the Spanish project PID2022-141339NB-I00, funded by MCIU /AEI
/10.13039/501100011033 / and FEDER A way to make Europe, EU

3.5 Symposium : ”Observational methodology”

Title

Preliminary Convergent-Discriminant Validity Evidence of the Methodological Quality Scale for Observational Methodology (MQSOM). A Confirmatory Factor Analysis

Author(s)
Salvador Chacón-Moscoso [1] , José Mena-Raposo [1] , Daniel López-Arenas ,
Susana Sanduvete-Chaves [1]

[1] University of Seville

Abstract
Introduction: designs based on observational methodology enable the systematic recording and subsequent quantification of the spontaneous behavior displayed by participants in natural contexts. These methods offer advantages including a low level of intervention, independence with respect to standardized measurement instruments and flexibility when applied in non-standardized intervention contexts. Consequently, observational methodology is frequently used in psychology, education or health, as well as in other social fields. A Methodological Quality Scale for Studies Based on Observational Methodology (MQSOM), a tool with adequate psychometric properties to measure the methodological quality of these studies, has recently been validated. Previous research has demonstrated convergent-discriminant validity evidence of MQSOM through bivariate correlations with other methodological quality tools. Objective: the aim of this communication is to further substantiate the validity of MQSOM by presenting the preliminary evidence of its convergent and discriminant validity obtained through confirmatory factor analysis (CFA). Methods: nine-hundred and twenty articles based on observational methodology were coded with MQSOM, Rigorous Mixed-Methods (RMM), Guidelines for Publishing Evaluations Based on Observational Methodology (GREOM) and Mixed Methods Appraisal Tool (MMAT), circumscribed to the field of Mixed-Methods studies. Then, a confirmatory factor analysis was conducted. Results: MQSOM dimensions exhibited moderate-to-strong loadings in latent factors together with those contrast instruments dimensions that addressed similar constructs, as well as low factor loadings considered theoretically incongruent. Conclusions: this work deepens in the strengthening of the MQSOM and provides additional evidence of its suitability to assess the quality of intervention programs based on observational methodology.

Title

Evaluating Methodological Quality in Football Studies: An Application of the MQSOM

Author(s)
Mª Teresa Anguera [1] , Salvador Chacón-Moscoso [2] , Daniel López-Arenas ,
Susana Sanduvete-Chaves [2]

[1] University of Barcelona; [2] University of Seville

Abstract

Introduction: observational methodology has been widely used in football research during the last decades, due to its low level of intervention and independence from standardized measurement tools. Observational methodology allows the systematic recording, quantification and analysis of player behaviors on the field. As scientific production based on observational methodology increases, there is a growing need to evaluate its methodological quality. A Methodological Quality Scale for Studies Based on Observational Methodology (MQSOM), a tool to measure the methodological quality of these studies, has recently been validated with adequate psychometric properties (RMSEA = 0.000, NNFI = 1, GFI = .98, AGFI = .97). The MQSOM comprises a second-order factor of Methodological quality (MQ; ω = .87; D = .55) containing two first-order factors: Quality of design (D1; 6 items; ω = .90; D = .46; ICC = .933 - .967) and Quality of measurement and analysis (D2; 5 items; ω = .68; D = .67; ICC = .797 - .988). Objective: this study presents the results of a systematic mixed-method review that applies the MQSOM to primary studies of observational methodology on football. Methods: descriptive statistics by country of affiliation, journal, object of study and event observed are presented. In addition, an analysis of proportions was performed to identify differences in terms of procedural characteristics. Finally, a two-stage cluster analysis was performed. Results: the analysis of proportions showed significant differences in terms of type of observation, coding manual specification, type of data, observation instrument, recording, control and analysis software, type of parameter, data quality control and analysis used. In addition, two-step cluster analysis produced five methodological quality profiles ranked in decreasing order by MQSOM score. Profile 1 exhibited high levels of methodological quality (GD = 0.81; D1 = 0.78; D2 = 0.89), Profile 2 exhibited moderate-to-high levels of methodological quality (GD = 0.76; D1 = 0.70; D2 = 0.83), Profile 3 exhibited low-to-high levels of methodological quality (GD = 0.56; D1 = 0.37; D2 = 0.79), Profile 4 exhibited low-to-moderate levels of methodological quality (GD = 0.46; D1 = 0.26; D2 = 0.70), and Profile 5 exhibited low levels of methodological quality (GD = 0.24; D1 = 0.06; D2 = 0.45). Conclusions: this application of MQSOM allows to obtain an updated snapshot of the observational methodology applied in football. This work highlights future improvements in terms of methodological quality and presents MQSOM as a valuable tool to assess the quality of sports intervention programs based on observational methodology.

Title

Methodological Quality Profiles in Basketball: A Systematic Review of Studies based on Observational Methodology

Author(s)
Mª Teresa Anguera [1] , Salvador Chacón-Moscoso [2] , Daniel López-Arenas ,
Susana Sanduvete-Chaves [2]

[1] University of Barcelona; [2] University of Seville

Abstract

Introduction: the application of observational methodology to basketball allows the organized record, quantification and analysis of behaviors displayed by players in the pitch, being commonly used in research due to its low intervention level and independence from standardized measurement tools. As scientific production based on observational methodology increases, so does the aim to assess its methodological quality. Recently, a methodological quality scale for studies based on observational methodology has been validated (MQSOM), being the first tool with adequate psychometric properties to measure the methodological quality of this studies. Objective: This work presents a systematic mixed study review that applies MQSOM in 220 primary studies based on observational methodology in basketball. Methods: Descriptive statistics of the main procedural characteristics were obtained, and proportion analysis were carried out. Finally, a two-step cluster analysis was carried out. Results: Proportion analysis provided significant differences in terms of observation type, observational design, observational instrument, codification manual specification, data type, recording, control and analysis software, type of parameter, data quality control, and analysis employed. Two-step cluster analysis showed methodological quality profiles in decreasing order based on MQSOM score. Conclusions: This work offers an updated image of the observational methodology applied in basketball, detailing future methodological quality improvements and presenting MQSOM as a valuable tool to assess the quality of intervention programs in sport based on observational methodology.

3.6 KEY NOTE: The Integrated Mixed Methods Transformation Approach

Title

KEY NOTE: The Integrated Mixed Methods Transformation Approach

Author(s)
Anthony Onwuegbuzie [1]

[1] University of Cambridge

Abstract

The field of mixed methods research continues to evolve, pushing the boundaries of methodological innovation to address complex and multifaceted research problems. This keynote address introduces the Integrated Mixed Methods Transformation Approach (IMMTA) as a meta-framework that systematically transforms monomethod research designs into fully integrated mixed methods research approaches. IMMTA fosters seamless integration of qualitative and quantitative elements across all research stages, resulting in richer, more comprehensive findings and maximizing methodological rigor. By embedding integration at all phases—design, data collection, analysis, and interpretation—IMMTA enhances the depth and applicability of research, particularly in interdisciplinary settings such as those at RAND.

In the second part of my keynote address, I will explore Critical Dialectical Pluralism (CDP) 2.0, an evolution of its predecessor, CDP 1.0, now positioned as a transformative multidimensional metaparadigm and metaphilosophy for mixed methods research. Grounded in the five pillars of social justice, inclusion, diversity, equity, and social responsibility (SIDES), CDP 2.0 represents a shift toward socially responsive and ethically engaged research practices. This meta-framework promotes participant empowerment by redefining their role as co-researchers and challenges traditional research hierarchies to foster an egalitarian and impactful research paradigm.

By bridging IMMTA and CDP 2.0, this keynote address offers a transformative perspective on mixed methods research, one that is methodologically rigorous and ethically profound. Attendees will leave with an enriched understanding of how to apply these paradigms to advance research, making an impact on policy and practice. This session promises to be a forward-looking discussion that reimagines the future of integrated research methodologies.