Contribution ID: **127** Type: **Symposium**

# Innovations in test development and validation

## Coordinator

Giulia Balboni*; Pasquale Anselmi°

## Affiliation

*University of Bologna, Italy; °University of Padova, Italy

## Communication 1

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

## Communication 2

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

## Abstract

The symposium introduces innovative methods and procedures that are useful in the development and validation of tests, as well as in the analysis of collected data, to facilitate a comprehensive understanding of the response behavior of individual respondents. The symposium is comprised of five talks. Anselmi et al. present Competence-based Test Development, a method for constructing, improving, and shortening tests that derive the skills an individual has available from the observed item responses. Faraci presents Exploratory Structural Equation Modeling as a valuable tool for clearly representing the complexity of psychological constructs and illustrates how it provides a compromise between Exploratory and Confirmatory Factor Analyses. Castellani et al. address the issue of the appropriate handling of Likert scale items as ordinal variables and illustrate the advantages of integrating Classical Test Theory and Item Response Theory. Maurer and Heller move within the framework of Knowledge Structure Theory and explore different options for capturing the discrepancy between the items an individual is capable of solving and the probability of this individual of capturing the discrepancy between the items an individual is capable of solving and the probability of this individual in failing items by inattention or solving them by guessing. Orsoni et al. propose the Probabilistic Information and Network Evaluation System, which integrates Bayesian networks and information theory and illustrates its utility in detecting specific cognitive patterns, difficulties, and potential inconsistencies in the responses provided by individual respondents. The five talks provide a comprehensive illustration of the methods, with examples drawn from both real and simulated data.

## Communication 3

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

## 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.

## Keywords

ESEM; data-driven approach; theory-driven approach

## Symposium title

Innovations in Test Development and Validation

## Authors

Pasquale Anselmi, *Jürgen Heller°*, *Luca Stefanutti*, Egidio Robusto*

## 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.

## Keywords

Bayesian-networks; entropy-based metrics; reliability diagnostics

## Affiliation

*University of Padova, Italy; °University of Tübingen, Germany

## Communication 4

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

## Authors

Alice Maurer, Jürgen Heller

## 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.

## Keywords

competence-based test development; competence-based-knowledge-structure theory

## Keywords

test-development; IRT; ESEM; KST; Bayesian-networks

## Affiliation

University of Tübingen, Germany

## 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.

## Keywords

probabilistic knowledge structure theory; response-errors

## Authors

Matteo Orsoni, Matilde Spinoso, Sara Giovagnoli, Sara Garofalo, Noemi Mazzoni, Giulia Balboni, Mariagrazia Benassi

## Authors

Aurora Castellani*; Anselmi Pasquale°; Roberto Cubelli^; Giulia Balboni**

## Affiliation

*University of Perugia, Italy; °University of Padova, Italy; ^University of Trento, Italy;*University of Bologna, Italy

## Authors

Palmira Faraci

## Affiliation

University of Bologna, Italy

## 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.

## Keywords

ordinal-variables; validation; CTT; IRT; invariance

## Affiliation

University Kore Enna, Italy

## Number of communicatios

5

## Communication 5

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

**Primary author:** BALBONI, Giulia (University of Bologna)

**Co-author:** Prof. ANSELMI, Pasquale (University of Padova)

**Presenter:** BALBONI, Giulia (University of Bologna)

**Session Classification:** Symposium : "Innovations in test development and validation"

**Track Classification:** Measurement: Measurement