EAM2025
XI Conference

23RD - 25TH
JULY
2025

Spain Tenerife
Canary Islands

European
Association of
Methodology

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

Farshad Effatpanah
TU Dortmund University, Germany

Universidad de La Laguna

Cabildo de Tenerife

Tagoro

cajasiete

Gobierno de Canarias
Consejería de Universidades, Ciencia e Innovación y Cultura
Agencia Canaria de Investigación, Innovación y Sociedad de la Información

ICI Instituto Canario de Igualdad

tea

hogrefe

Introduction

- Multiple-choice (MC) items are commonly used in high-stake tests.

  - **Advantage:**

  1. Ease of administration
  2. Objective scoring
  3. High scoring reliability
  4. Wide content coverage
  5. Cost-effective

  - **Disadvantage:**

    Susceptible to construct-irrelevant factors

  1. Testwiseness
  2. Number of response options
  3. Cheating
  4. Guessing
  5. Pattern Guessing

Test Booklet Variations

- To minimize the risk of cheating and ensure test security, multiple versions of the test booklets are constructed with variations in item order, option order, or both.

1) Position of Items Varies, Options Stay the Same

Suitable for maintaining item comparability while reducing position bias.

2) Position of Items is the Same, Options Vary

Good for detecting or minimizing response strategy effects.

3) Both Item and Option Positions Vary

Most secure and requires detailed linking/equating plans to ensure comparability.

Does changing the position of items and options affect the performance of test takers?

- **Common Assumptions:**

1) If test content is the same, then

   Changes in item order or option position do not affect item difficulty and other psychometric characteristics.

2) When alternate test forms are used,

   A common measurement metric is assumed.

   Equating is not needed unless (partially or entirely) different sets of items are used.

- **Problem:**

  Even when identical items are used across forms,

  Item difficulty parameters may change due to context effects

Violates the assumption of invariance and score comparability

Studies on Item and Option Position Effects

Authors & Year	Item/Option Position Effect Analysis	Model & Design	Final Results
Cizek (1994)	Option order effects on MC items (reordering answer options)	CINEG equating design with experimental manipulation of option order	**Significant but unpredictable** option order effects on item performance
Hahne (2008)	Item position effects using varying item orders	LLTM for incomplete data; virtual item concept; Andersen’s LR test	**No significant** position effect found
Hohensinn et al. (2011)	Item position effects in a 4th-grade math test	LLTM; booklet design; simulation study to test model power	**No significant** global item position effect found
Weirich et al. (2016)	Item position effects & interaction with test-taking effort over time	Microlongitudinal design; effort components modeled; position effects examined across test duration	**Significant linear** position effects found; moderated by declining test-taking effort
Zeller et al. (2017)	Item position effects vs. increasing item difficulty (Raven’s APM)	Confirmatory Factor Analysis; comparison of original vs. randomized item orders	**Both effects found;** item position effect distinct from difficulty ordering
Hohensinn & Baghaei (2017)	Option position effects in MC items (positions a, b, c, d)	LLTM	**Slight option position** effects found; correct answers at later positions were slightly more difficult
Liu et al. (2024)	Item block position effects and subject session order effects in digital problem-solving tasks	Structural Equation Modeling; analysis of ability and response time using TIMSS 2019 Grade 4 data (N = 27,682)	**Small but significant block and booklet effects**; earlier blocks received more time and yielded better performance

Does changing the order of response options across test booklets affect item difficulty or lead to differential item functioning (DIF) among test takers, despite identical item content?

Purpose and Research Questions

- This study aims to use the Rasch tree to investigate DIF in the reading comprehension section of a high-stakes MC test. The covariates used for DIF analysis are gender and booklet. The following research questions were addressed:

RQ1: Do the reading comprehension test items exhibit DIF with respect to test takers' gender?

RQ2: Do the reading comprehension test items exhibit DIF with respect to booklets?

Rasch Tree

- Strobl et al. (2015) proposed Rasch tree for detecting DIF in the Rasch model framework (Rasch, 1960/1980).
- It does not require pre-specified groups and can simultaneously incorporate several covariates.

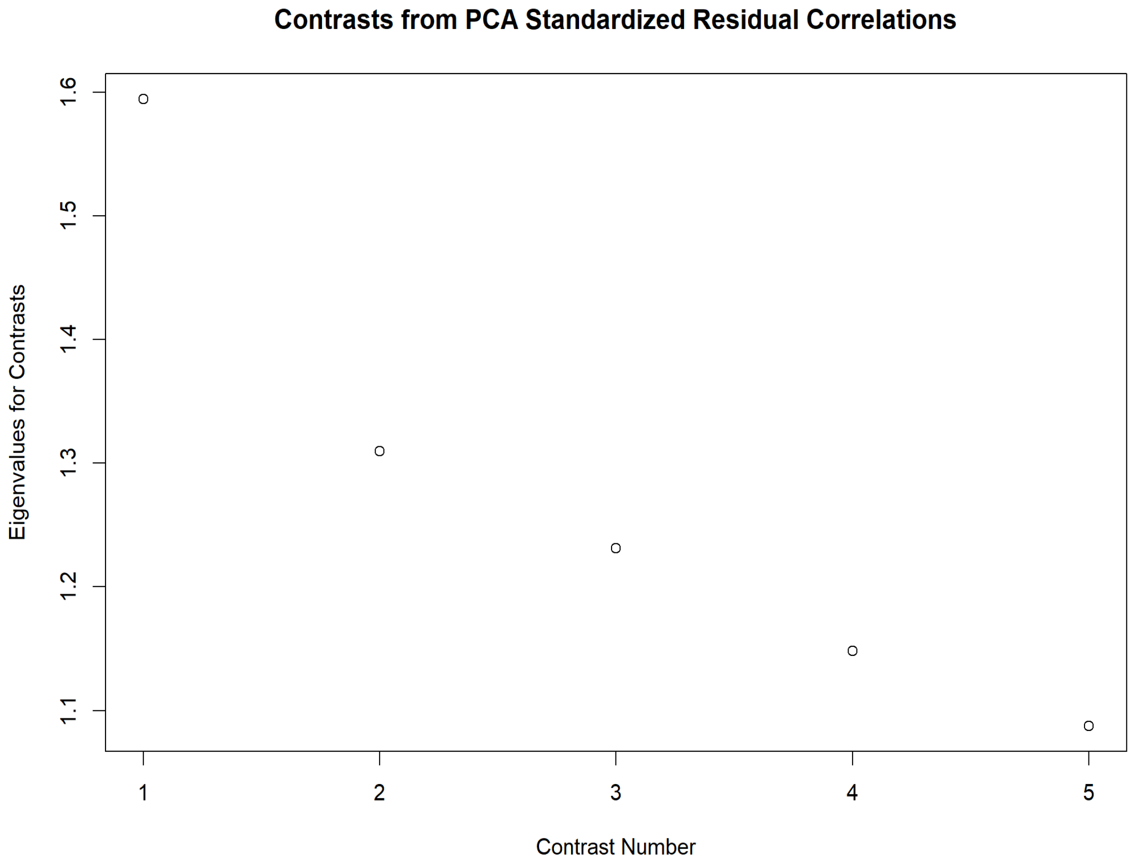
- The following steps are used to infer the structure of Rasch tree from the data (Strobl et al., 2015):

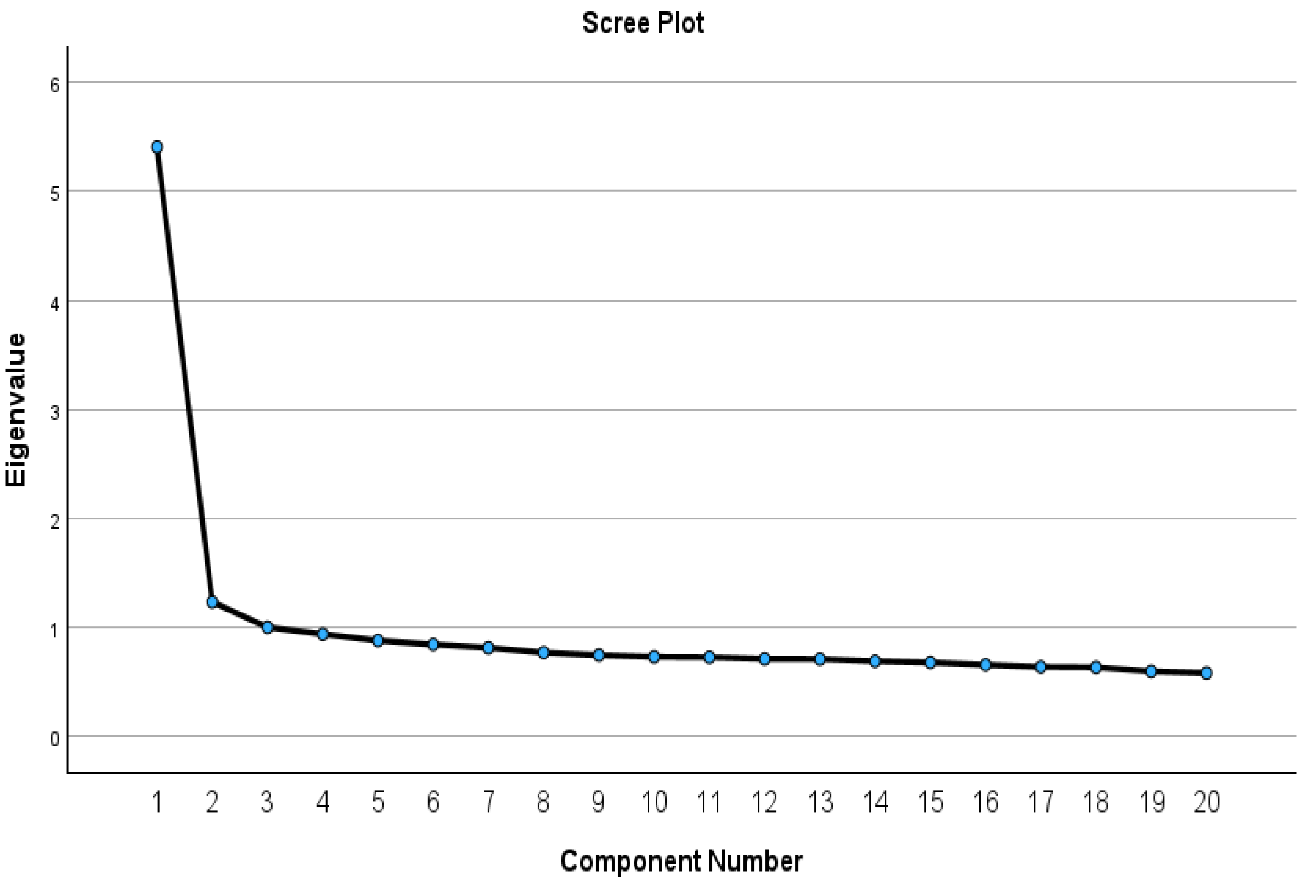
1) Estimate the item parameters jointly for all examinees in the full sample,
2) Assess the stability of the item parameters regarding each covariate.
3) If there is significant instability, split the sample along the covariate with the strongest instability and at the cut-point leading to the highest improvement of model fit.
4) Repeat Steps 1 to 3 recursively in the resulting subsamples until there are no more significant instabilities
5) Henninger et al. (2023) extended this method by introducing a new stopping criterion based on the Mantel-Haenszel odds ratio, aligning with the Educational Testing Service (ETS) DIF classification scheme to quantify effect sizes and flag DIF items.

Method

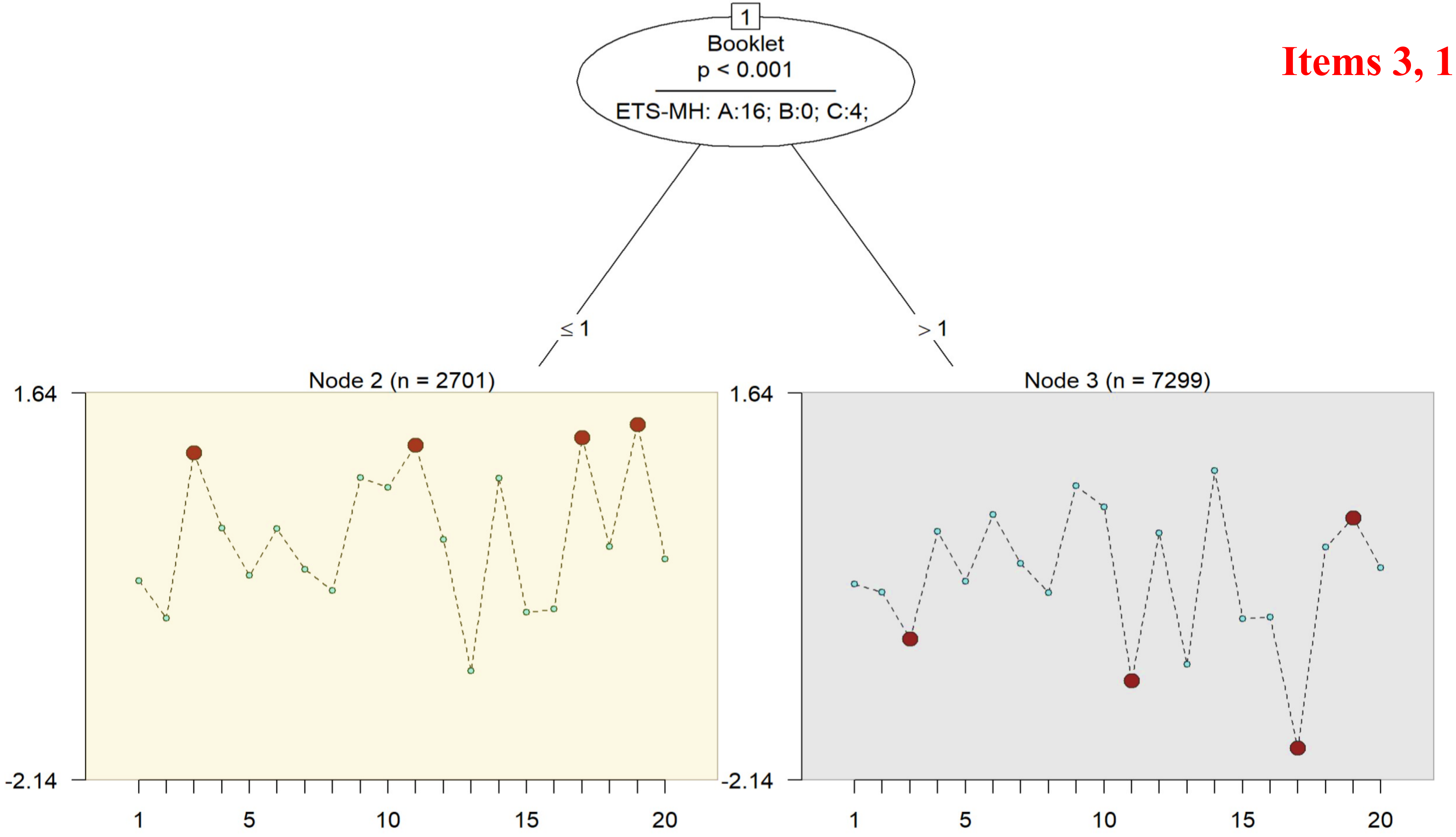
- Item responses (with no missing values) of 10,000 test takers to 20 reading comprehension test items presented across four booklets were analyzed using the ‘*psychotree*’ R-package (Zeileis et al., 2024).
- There were 6,821 (68.2%) females and 3,179 (31.8%) males. Their mean age was 18.79 (SD = 2.86).
- The Cronbach reliability of the reading test was 0.851, with confidence intervals of 0.847–0.855.
- The test content and item positioning were consistent across booklets, with only the positions of the answer options varying.

Results: Checking Dimensionality





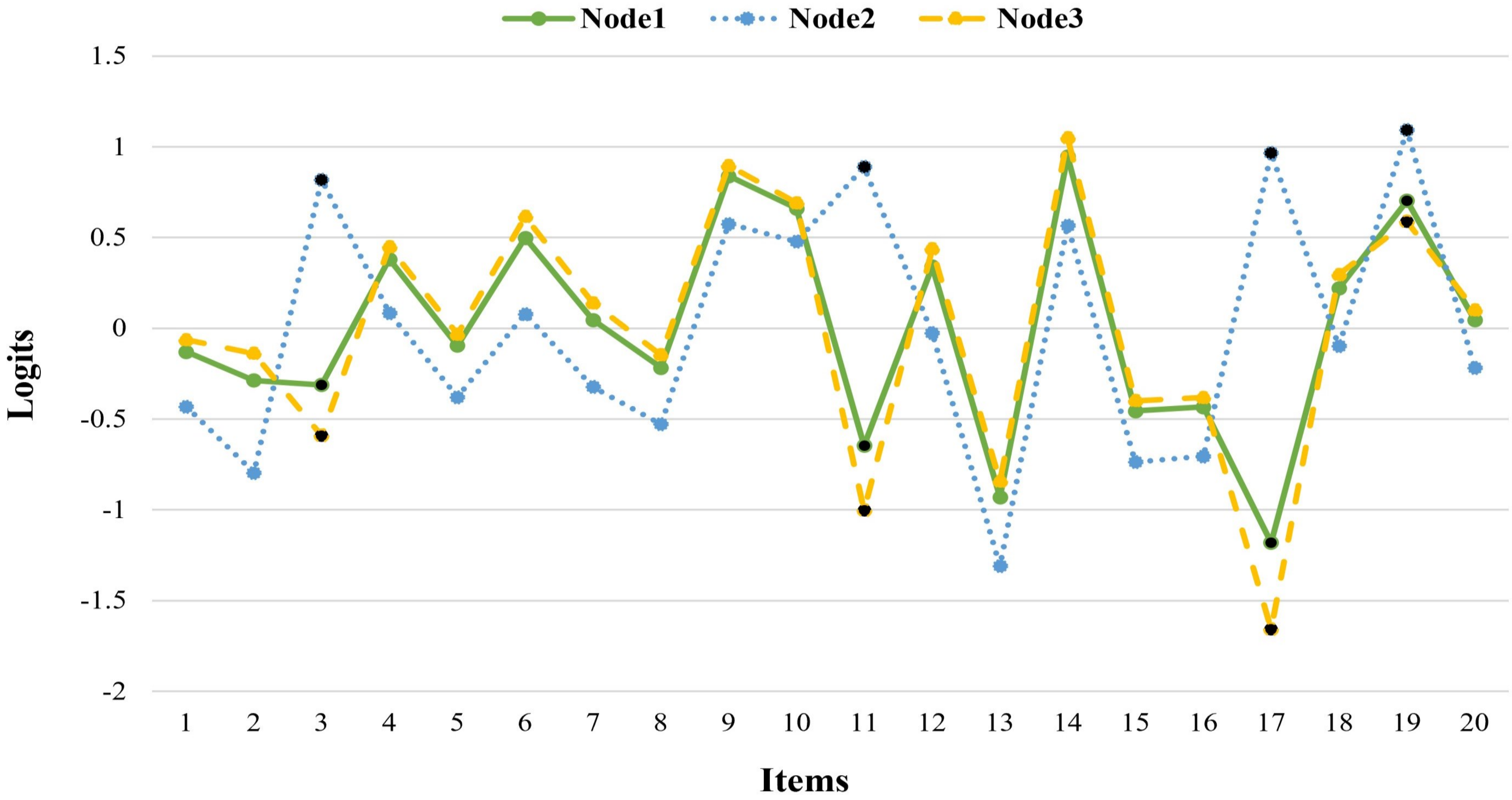
Results: Rasch Tree Model



Items with large DIF:

Items 3, 11, 17, and 19

Results: Item Difficulties across the Nodes



Discussion

Justification for all the items exhibiting DIF

- Test takers’ performance on Booklet 1 (Node 2) was likely influenced by the positioning of distractors.
- When correct answers appeared next to highly plausible distractors, cognitive load increased, making it harder to distinguish between similar options—especially under timed conditions.
- Consequently, because correct answers were consistently placed next to distractors in Booklet 1, it might cause more second-guessing or lead to incorrect selections, especially with more plausible distractors.

Discussion

Justification for Items 3, 11, and 17

- Test takers often form expectations about option formats.
- The items had different writing formats, which may have disrupted these expectations.
- Such variations can affect how options are processed. For example, longer or differently formatted options might be harder to scan quickly, leading to more mistakes or longer response times, particularly under test conditions.

Discussion

- Most items have the following format:

Vertically listed options

N– XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX …………….. XXXXXX.

a) ___________

b) ___________

c) ___________

d) ___________

Discussion

- The three DIF items have the following format:

Horizontally listed options

1–XXXXXXXXXXX................... XXXXXXXXXXXXXXXXXXXX.

a) ____________ b) ____________ c) ____________ d) ____________

Discussion

Justification for Items 3, 11, and 17

- When answer options look or sound similar at the start, it can disrupt visual and cognitive processing.
- Due to perceptual priming, test takers may skim familiar-looking options and miss key differences.
- Similar starting letters or sounds can also cause lexical interference, leading to confusion or second-guessing.
- Under time pressure, this overlap—combined with the serial position effect—can impair recall and increase errors, especially when relying on partial memory.

Discussion

Justification for Item 19

- The correct answer requires an inference from the passage, which can be harder under cognitive load.
- Options 3 and 4 both relate to astronaut well-being (physical/emotional vs. psychological), creating conceptual overlap.
- This similarity, combined with fatigue or stress—especially later in the test—can lead to confusion and increase the chance of misinterpretation or error.

Implications

- The findings suggest that booklet design, though often considered neutral when content is the same, can favor certain subgroups, particularly through interactions with gender.

- Test developers should consider not only what items assess but also how they are presented.

Limitations

- Only two covariates were incorporated into the analysis.
- The findings may not generalize to other domains (e.g., math, science).
- Although booklet content was the same, contextual or order effects across booklets (e.g., fatigue, engagement) could still influence performance and were not fully isolated from positional effects.
- Gender differences in performance might be related to factors not measured (e.g., reading fluency, test anxiety).

Thank you for your attention!