Pearson

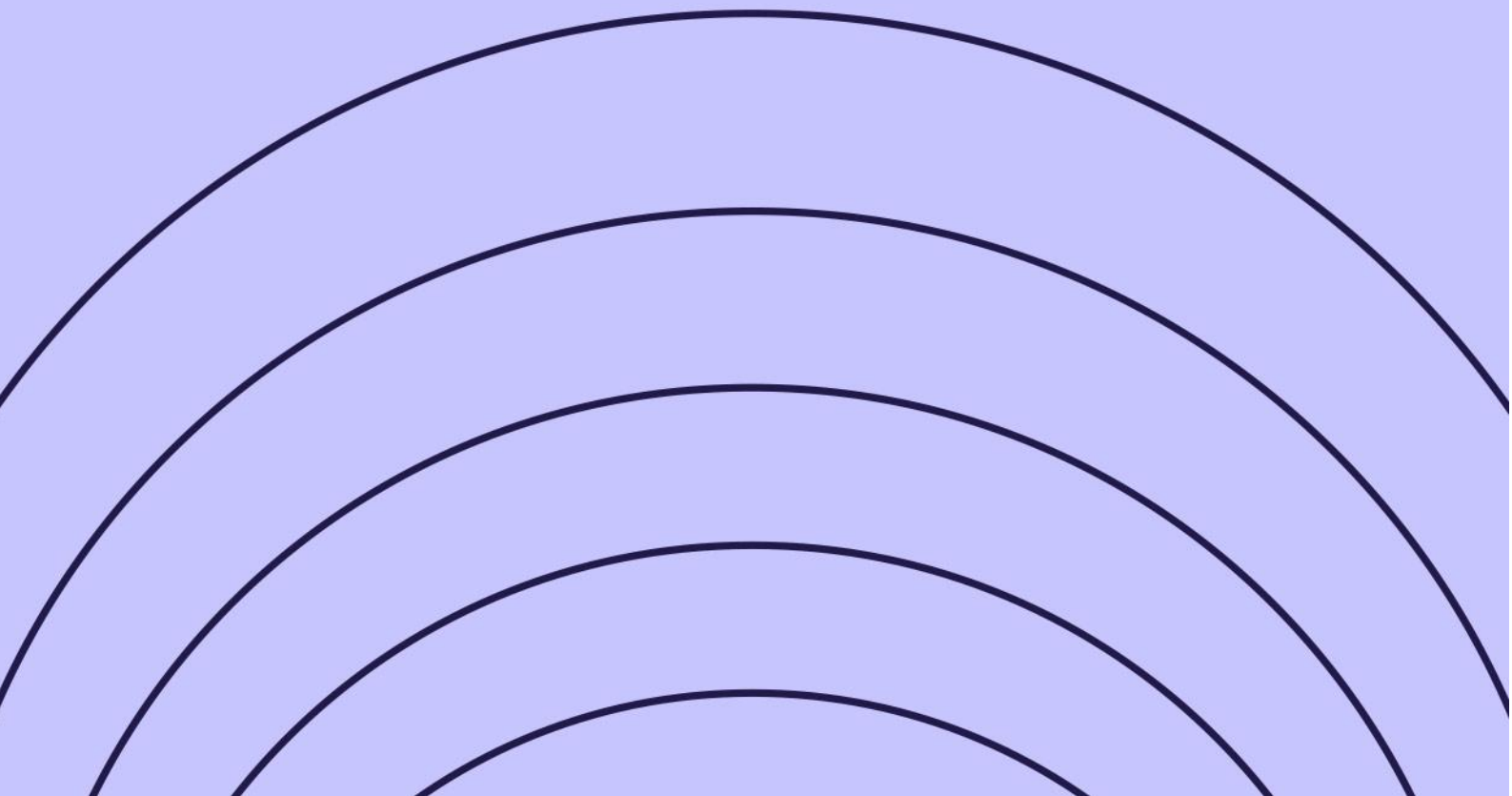
Research note:

Using Product Data to Investigate the Internal Structure and Measurement Invariance of Pearson Test of English Core (PTE Core) Reading Items

Summary paper

Esra Sözer Boz, Canakkale Onsekiz Mart University

Derya Akbaş, Aydın Adnan Menderes University



Introduction

The PTE Core is a computer-based international test assessing English skills such as writing, reading, listening, and speaking (Pearson, 2024). Since most test-takers are second-language users from diverse backgrounds, ensuring that scores are comparable across groups is essential for fair and valid interpretations (Kane, 2013). This requires measurement invariance (MI), meaning the test measures the same construct consistently across groups (Putnick & Bornstein, 2016; Vanderberg & Lance, 2000). In this study, the factor structure of the reading items in the PTE Core test and the extent to which measurement invariance across different groups was investigated using Bayesian Structural Equation Modeling (BSEM) (Muthen & Asparouhov, 2012).

BSEM offers benefits over frequentist methods, including computational efficiency, incorporation of prior information, and easier interpretation of results (Smid et al., 2020). Unlike frequentist approaches that require large samples and may perform poorly with small ones (Muthen & Asparouhov, 2012), Bayesian methods provide reliable estimates based on posterior distributions (Chen et al., 2024). This is particularly useful when sample sizes are limited due to small populations (Egberts et al., 2016), hard-to-reach groups, or test designs like Linear-on-the-Fly Test Assembly (LOFT), which generates unique test forms for each taker to ensure security but results in missing data (Becker & Bergstrom, 2013; Pearson, 2024). Given these advantages, BSEM was applied to evaluate the PTE Core reading items in this study.

For measurement invariance (MI), Muthén and Asparouhov (2013) proposed Bayesian approximate measurement invariance (AMI) as an alternative to exact MI testing, in which parameter differences across groups are constrained to zero. Unlike this approach, which often results in poor model fit, especially with many groups (Kim et al., 2017), Bayesian AMI allows for small, substantively negligible differences in parameters across groups (Leitgöb et al., 2023). This method permits minor variations in factor loadings and intercepts while still enabling valid comparisons (van de Schoot et al., 2013). It assumes these differences follow a normal distribution centered at zero with a predefined variance, keeping deviations minimal to maintain approximate invariance (Wurster, 2022). In this study, Bayesian AMI was applied to test the invariance of reading items in the PTE Core test.

This study explored the factor structure of reading items in the PTE Core test and investigated whether the underlying construct (i.e., reading skills) is measured in the same way across different groups of test-takers. The factor structure of the reading items was investigated using Bayesian Exploratory and Confirmatory Factor Analysis (EFA and CFA). Bayesian Approximate Measurement Invariance (AMI) was used to assess item-level comparability across groups (e.g., gender, language family, and test center location). It was hypothesized that full measurement invariance would hold for gender, given comparable educational and language learning opportunities, whereas full invariance was not expected for language family and test center location due to substantial cross-linguistic and contextual diversity.

Method

Participants and Measure

The dataset provided by Pearson contained PTE Core reading responses and background information for 42,879 test-takers (52.6% male, 47.4% female; M_{age}= 30.90 years, *SD* = 7.10). The PTE Core reading section includes five item types (Pearson, 2024): (1) reading and writing: fill in the blanks (GAPS RW), (2) multiple choice – multiple answers (MAMC), (3) reorder paragraph (DRDR), (4) fill in the blanks (GAPS), and (5) multiple choice – single answer (SAMC). The first item type was excluded because it assesses integrated reading and writing skills, allowing us to focus on items assessing reading skills only when evaluating unidimensionality. Data were provided for 107 reading items. MAMC, DRDR, and GAPS are scored using partial credit, while SAMC is scored dichotomously (Pearson, 2024). Under Pearson's linear-on-the-fly testing (LOFT) design, each test-taker receives a unique set of items drawn from a large item pool (eight items in the present dataset).

Data Analysis

Data analysis was conducted in three stages: exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and Bayesian approximate measurement invariance (AMI). All analyses were performed using M*plus* 8.11 software (Muthén & Muthén, 1998-2017). First, we conducted a Bayesian exploratory factor analysis (EFA), evaluating both one- and two-factor solutions. The full sample was randomly divided into two subsamples: Sample 1 (odd registration IDs) for EFA and Sample 2 (even registration IDs) for CFA. Sample 1 included 21,424 test-takers (52.5% male, 47.5% female; M_{age} = 30.90 years, *SD* = 7.16). Then, a Bayesian Confirmatory Factor Analysis (CFA) was conducted using Sample 2, which comprised 21,455 test-takers (52.7% male, 47.3% female; M_{age} = 30.91 years, *SD* = 7.05). Based on EFA results, a one-factor model was estimated to evaluate the fit of the unidimensional structure. As no prior information was available for the model parameters, non-informative priors (i.e., N (0, ∞) (Asparouhov & Muthen, 2021)) were used. Finally, we applied the Bayesian approximate measurement invariance (AMI) approach using the data of the full sample (n = 42,879) to assess the measurement invariance across gender, language family, and test center location. For language family, languages were categorized into four language families based on Barkaoui (2019): Indo-European (e.g., English, German; 25.6%), Indo-Iranian (e.g., Farsi, Hindi; 48.2%), Dravidian (e.g., Kannada, Tamil; 8.6%), and other languages (17.5%). For test center location, centers were grouped according to Kachru's (1992) concentric circles of English: (1) inner circle (e.g., Australia, Canada; 73.1%), and (2) outer and expanding circles (e.g., India, Malta; 26.9%). For Bayesian AMI, we tested five models with informative priors (mean = 0) and variances of 0.001, 0.005, 0.01, 0.05, and 0.10 for the differences in factor loadings and intercepts (i.e., scalar invariance) across each grouping variable separately.

Results

EFA results showed that eigenvalues were 28.25 for the first factor and 3.18 for the second, indicating a sharp drop and the presence of a dominant general factor. One-factor model demonstrated a good fit with a posterior predictive *p*-value (*ppp*) of 0.494 and a 95% confidence interval for the difference between observed and replicated chi-square values

of [-288.317, +307.118], consistent with the criteria suggested by Muthén and Asparouhov (2012). This result suggests that the reading items in the PTE Core test have a unidimensional factor structure. Factor loadings ranged from 0.12 to 0.78, and items with loadings below 0.32 were retained for further analysis.

CFA results indicated that one-factor CFA showed an acceptable fit, with a *ppp* of 0.476 and a 95% confidence interval (CI) of [-310.773, +325.363]. Loadings ranged from 0.15 to 0.79. While some items had lower loadings, most contributed meaningfully to the latent factor. Overall, the findings support a well-fitting one-factor structure for reading items in the PTE Core test, despite some variability in item contributions.

AMI results indicated that all models demonstrated good fit, with posterior predictive *p*-values (*ppp*) close to 0.50 and 95% confidence intervals that included zero near their centers (Table 1). Among the models, the model with the lowest Deviance Information Criterion (DIC) value was selected for further interpretation. The selected models had prior variances of V= 0.001 for gender, V = 0.01 for language family, and V = 0.005 for test center location. Overall, the findings provided some support for approximate scalar MI. However, item-level deviations in factor loadings and intercepts indicated that, for gender, most items showed minimal differences, although a few exhibited notable deviations. For language family, most items showed deviations in either factor loadings or intercepts in at least one group, but not both across all groups. For test center location, some items showed no deviations, whereas many exhibited deviations in one or both parameters.

Table 1. Model fit for Bayesian AMI across gender, language family, and test center location

Prior	DIC	*ppp*	95% CI
Gender			
M = 0, V = 0.001	767036.268	0.456	[-386.171, +381.929]
M = 0, V = 0.005	767049.620	0.464	[-386.611, +387.468]
M = 0, V = 0.01	767073.242	0.480	[-405.969, +465.207]
M = 0, V = 0.05	767090.476	0.479	[-388.774, +401.285]
M = 0, V = 0.10	767093.885	0.478	[-388.032, +401.751]
Language family			
M = 0, V = 0.001	762714.241	0.380	[-455.846, +721.799]
M = 0, V = 0.005	761903.018	0.473	[-536.637, +676.425]
M = 0, V = 0.01	761835.116	0.490	[-544.893, +672.474]
M = 0, V = 0.05	761884.059	0.461	[-627.119, +608.931]
M = 0, V = 0.10	761907.215	0.461	[-624.339, +609.926]
Test center location			
M = 0, V = 0.001	762784.879	0.456	[-408.979, +454.690]
M = 0, V = 0.005	762670.735	0.483	[-419.232, +440.186]
M = 0, V = 0.01	762678.014	0.473	[-421.450, +452.171]
M = 0, V = 0.05	762702.558	0.483	[-418.205, +438.157]
M = 0, V = 0.10	762708.201	0.496	[-418.824, +439.336]

Note. DIC=deviance information criterion; *ppp*=posterior predictive *p*-value; CI=confidence interval. M = prior mean; V = prior variance. All priors follow a normal distribution.

Conclusion

This study examined the factor structure of reading items in the PTE Core test and tested whether they measure reading skills consistently across groups. Given the high-stakes nature of the test, evaluating the validity of its scores is essential. Overall, the study aimed to support the ongoing validation of the PTE Core reading assessment.

The Bayesian EFA results indicated that factor loadings were generally strong and credible intervals were narrow, indicating a stable solution for the one-factor model. This result suggested that reading items in the PTE Core Test had a one-factor structure, measuring the reading skills of test-takers. The Bayesian CFA results supported the one-factor structure of the reading test.

The Bayesian AMI results indicated that scalar invariance could be supported for gender, language family, and test center location, with best-fitting prior variances of 0.001, 0.01, and 0.005, respectively. Item-level deviation patterns for gender suggest that gender does not introduce substantial measurement bias, which is plausible given that males and females are generally exposed to comparable educational contexts and are not systematically differentiated in terms of language learning opportunities. Item-level deviations for language families may reflect genuine cross-linguistic differences in how test items are processed, since test-takers' first language (L1) structures (e.g., sentence construction, word formation patterns, and word order) and literacy practices differ widely across language families. Such structural and educational variations can naturally lead to differences in response patterns even when the underlying construct is the same. Deviations for test center location groups might mirror the diverse sociolinguistic and educational contexts represented by test centers, which range from settings with high daily exposure to English (inner circle countries) to those where English is primarily learned in formal instruction (outer or expanding circle contexts). These contextual differences can affect the degree of familiarity with item formats and test-taking practices, producing more frequent deviations at the item level. Overall, these findings support approximate scalar invariance, although caution is warranted regarding latent mean comparisons, especially across language families and test centers due to more frequent item-level deviations.

In conclusion, this study offers important evidence for the validity of reading items in the PTE Core. Evaluating factor structure and invariance is crucial for tests used with diverse populations to ensure fair and meaningful cross-group comparisons. Given that PTE Core assesses English proficiency internationally, ensuring score comparability is essential for its valid and equitable use.

References

Asparouhov, T., & Muthen, B. (2021). *Bayesian Analysis of Latent Variable Models using Mplus.* Retrieved from Mplus: https://www.statmodel.com/download/BayesAdvantages18.pdf

Barkaoui, K. (2019). Examining sources of variability in repeaters' L2 writing scores: The case of the PTE Academic writing section. *Language Testing, 36*(1), 3-25. https://doi.org/10.1177/0265532217750692

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research & Evaluation, 18*(14).

Chen, Q., Su, K., Feng, Y., Zhang, L., Ding, R., & Pan, J. (2024). A tutorial on Bayesian structural equation modelling: Principles and applications. *International Journal of Psychology, 59*(6), 1326-1346. https://doi.org/10.1002/ijop.13258

Egberts, M. R., van de Schoot, R., Boekelaar, A., Hendrickx, H. G., & Van Loe, N. E. (2016). Child and adolescent internalizing and externalizing problems 12 months postburn: The potential role of preburn functioning, parental posttraumatic stress, and informant bias. *European Child & Adolescent Psychiatry, 25*, 791-803. https://doi.org/10.1007/s00787-015-0788-z

Kachru, B. (1992). *The other tongue: English across cultures (2nd ed.).* Urbana, IL: University of Illionis Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 524-544. https://doi.org/10.1080/10705511.2017.1304822

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., Roover, K. D., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research, 110*. https://doi.org/10.1016/j.ssresearch.2022.102805

Muthen, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313-335. https://doi.org/10.1037/a0026802

Muthén, B., & Asparouhov, T. (2013). *BSEM measurement invariance analysis.* Retrieved from Mplus Web Note: No. 17: https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L., & Muthén, B. (1998-2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Pearson. (2024). *PTE Core Test taker score guide.* Pearson.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A

systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 131-161. https://doi.org/10.1080/10705511

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

Vanderberg, R. J., & Lance, C. E. (2000). A Review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

Wurster, S. (2022). Measurement invariance of non-cognitive measures in TIMSS across countries and across time. An application and comparison of Multigroup Confirmatory Factor Analysis, Bayesian approximate measurement invariance and alignment optimization approach. *Studies in Educational Evaluation, 73*. https://doi.org/10.1016/j.stueduc.2022.101143