Pearson

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

Esra Sözer Boz, Canakkale Onsekiz Mart University

Derya Akbaş, Aydın Adnan Menderes University

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

Abstract

The PTE Core is an international computer-based test designed to evaluate English language skills among test-takers. Most of the test-takers are second-language learners and come from different backgrounds. Therefore, comparability of the PTE Core scores is crucial for validating interpretations and ensuring fair use of the test. This study aimed to examine the factor structure of the reading items (focusing on single-skill items to provide unidimensionality) in the PTE Core test and investigate whether the underlying construct (i.e., reading skills) is measured in the same way across different groups of test-takers (i.e., measurement invariance). Data consisted of 42,879 test takers who had completed background questionnaires and reading items. The PTE Core has different item types and scoring types. Each test-taker responded to a total of eight reading items, which are one item from multiple-choice multiple answers, two items from reorder paragraphs, four items from fill-in-the-blank, and one multiple-choice single answer. Bayesian exploratory factor analysis and confirmatory factor analysis were conducted to explore the factor structure of the reading items. Bayesian approximate measurement invariance was conducted to test whether measurement invariance is established across different groups (e.g., gender, language family, and test center location). The factor analysis results indicated that the PTE Core reading items had a unidimensional factor structure. The results of the measurement invariance showed that it is reasonable to compare the latent means of PTE Core reading items across gender, language families, and test center locations, but latent mean comparisons should be interpreted with some caution. Consequently, this study, by providing empirical evidence on the invariance of the PTE Core, contributes to the broader body of research on the validity of Pearson's English language assessments.

Keywords: measurement invariance, PTE Core, validity, reading, internal structure

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). Without MI, differences in scores may reflect bias rather than true ability differences (Brown et al., 2017; Leitgöb et al., 2023). For example, if items function differently for males and females, test bias occurs, leading to unfair outcomes (Brown, 2015). To make valid and trustworthy comparisons, it needs to be confirmed that the variations in scores reflect true differences in the construct of interest, not influenced by the measurement tool. 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).

Bayesian Structural Equation Modeling (BSEM)

Bayesian analysis is often used in educational and psychological measurements (Chen et al., 2024) when models are too complex for frequentist methods, sample sizes are small, prior knowledge needs to be incorporated, or researchers prefer Bayesian outputs (Hoofs et al., 2018; Depaoli et al., 2020).

Bayesian Structural Equation Modeling (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.

Measurement Invariance

Multi-Group Confirmatory Factor Analysis (MGCFA) is the standard method for testing measurement invariance (MI) (Vanderberg & Lance, 2000). It assesses different levels of MI: configural (same factor structure), metric (equal factor loadings), and scalar invariance (equal factor loadings and intercepts) across groups. While metric invariance allows comparison of relationships between variables, scalar invariance is needed to compare latent means (Cieciuch et al., 2014). However, achieving full scalar invariance is often challenging (Byrne & van de Vijver, 2010). Partial invariance, which relaxes constraints on a few parameters, can be used instead (Byrne et al., 1989). Still, both full and partial invariance approaches have limitations, and approximate measurement invariance, which allows small differences in parameters, offers a more flexible and unbiased alternative (Winter & Depaoli, 2019).

Muthén and Asparouhov (2013) proposed Bayesian approximate measurement invariance (AMI) as an alternative to strict MI testing. Unlike full MI, 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—known as "wiggle room"—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.

The Present Study

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. To evaluate the test's factor structure and its measurement invariance, two research questions were addressed:

1. What is the factor structure of reading item responses on the PTE Core test?
2. To what extent are reading item responses on the PTE Core test comparable across different groups of test-takers (i.e., gender, language family, and test center location)?

The first research question was investigated using Bayesian Exploratory and Confirmatory Factor Analysis (EFA and CFA) to explore the dimensional structure of the reading items. The second question involved applying Bayesian Approximate Measurement Invariance (AMI) 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

PTE Core

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). We excluded the first type, which assesses integrated skills, and focused solely on the items assessing only reading skills to evaluate the unidimensionality. Due to Pearson's LOFT design, each test taker receives a unique set of items drawn from a large item pool. There are several items for each item type in PTE Core's item bank, with test takers being allocated different subsets of these items each time a test is taken. Each test taker responds to eight reading items: one MAMC, two DRDR, four GAPS, and one SAMC. Data was provided for 107 reading items. MAMC, DRDR, and GAPS are scored using partial credit, while SAMC is scored dichotomously (Pearson, 2024).

Dataset

Pearson provided test data, background, and technical information for 42,879 test-takers (52.6% male, 47.4% female; M_{age} = 30.90 years, SD = 7.10). Participants came from 155 countries, mainly India (n = 24,191, 56.4%), China (n = 3,451, 8.0 %), and Nepal (n = 1,796, 4.2%), and reported speaking 105 languages, most commonly English (n = 9,385, 21.9%), Punjabi (n = 7,990, 18.6%), and Hindi (n = 4,534, 10.6%). The test was administered in 105 countries, with most taking place in Canada (n = 27,792, 64.8%), India (n = 5,454, 12.7%), and the U.S. (n = 1,725, 4.0%). The most common occupational fields were the health and

social services sector (*n* = 6,266, 14.6%), education (*n* = 4,750, 11.1%), and technical/scientific fields (*n* = 4,496, 10.5%).

The overall and reading scores range from 10 to 90. The mean overall score was 63.69 (*SD* = 15.03), and the mean reading score was 63.58 (*SD* = 15.43). Due to the LOFT design, response counts per item varied between 1,858 and 4,163. Item score ranges were 0–1 for SAMC, 0–2 for MAMC, 0–3 for DRDR, and 0–5 for GAPS. In this study, item-level scores were used to conduct factor analysis and test measurement invariance across gender, language family, and test center location.

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).

Exploratory Factor Analysis

To explore the underlying structure of the PTE Core reading items, 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 confirmatory factor analysis (CFA). Sample 1 included 21,424 test-takers (52.5% male, 47.5% female; M_{age} = 30.90 years, *SD* = 7.16). Participants came from 145 countries, mainly India (*n* = 12,060), China (*n* = 1,788), and Nepal (*n* = 889), and reported speaking 92 languages, most commonly English (*n* = 4,706), Punjabi (*n* = 4,019), and Hindi (*n* = 2,317). Testing took place in 96 countries, most frequently in Canada (*n* = 13,884), India (*n* = 2,700), and the U.S. (*n* = 904). For Sample 1, the mean overall test score was 63.67 (*SD* = 15.14), and the mean reading score was 63.55 (*SD* = 15.56).

Bayesian EFA was performed using the Bayes estimator and Gibbs sampling with two MCMC chains and 50,000 iterations, applying Geomin (oblique) rotation. Model fit was evaluated using the posterior predictive *p*-value (*ppp*), which reflects the proportion of iterations where the replicated χ^2 exceeds the observed χ^2 (Lee, 2007). A good fit is indicated by a *ppp* around 0.50 (Muthen & Asparouhov, 2012), while values below 0.05 or 0.01 suggest a poor fit (van de Schoot et al., 2014). The posterior predictive check also produces a 95% confidence interval for the discrepancy between the observed and

replicated data. A well-fitting model is indicated when the lower bound of this interval is negative and the interval includes zero at its center (Muthen & Asparouhov, 2012).

Confirmatory Factor Analysis

To examine the factor structure of the PTE Core reading items, 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). A one-factor model was estimated to evaluate the fit of the unidimensional structure. Participants came from 140 countries, mainly India (*n* = 12,131), China (*n* = 1,663), and Nepal (*n* = 907), and reported speaking 98 languages, most commonly English (*n* = 4,679), Punjabi (*n* = 3,971), and Hindi (*n* = 2,217). The test was completed in 95 countries, most frequently in Canada (*n* = 13,908), India (*n* = 2,754), and the U.S. (*n* = 821). For Sample 2, the mean overall test score is 63.71 (*SD* = 14.92), and the mean reading score was 63.60 (*SD* = 15.30).

Bayesian CFA was conducted using the Bayes estimator and Gibbs sampling algorithm, using two MCMC chains with 50,000 iterations to estimate the posterior distribution. As no prior information was available for the model parameters, non-informative priors were used. In Mplus, the default priors for intercepts and factor loadings are specified as N (0, ∞) (Asparouhov & Muthen, 2021). Model fit was assessed using the posterior predictive *p*-value (*ppp*) and the 95% credibility interval, following the same procedure as in the Bayesian EFA.

Bayesian Approximate Measurement Invariance

To assess the measurement invariance of the PTE Core reading items across gender, language family, and test center location, we applied the Bayesian approximate measurement invariance (AMI) approach using the data of the full sample (n = 42,879). Based on Barkaoui (2019), test-takers were categorized into four language families: Indo-European (e.g., English, German, Dutch; *n* = 10,986, 25.6%), Indo-Iranian (e.g., Farsi, Hindi, Urdu; *n* = 20,680, 48.2%), Dravidian (e.g., Kannada, Tamil, Telugu; *n* = 3,709, 8.6%), and other languages (*n* = 7,504, 17.5%). Test center locations were grouped according to Kachru's (1992) concentric circles of English: (1) inner circle (Australia, Canada, UK, USA, Ireland, New Zealand; n = 31,339, 73.1%), (2) outer circle (e.g., India, Malta, Philippines; n = 9,200, 21.5%), and (3) expanding circle (all other countries; n = 2,340, 5.5%). Due to convergence issues stemming from the small sample size in the expanding circle, the

outer and expanding circles were combined into one group (n = 11,540; 26.9%). The descriptive statistics of overall scores and reading items for each group are given in Table 1.

Table 1. The overall and mean scores of reading items for each group

Groups	Overall scores		Reading items	
	Mean	Standard Deviation	Mean	Standard Deviation
Gender				
Female	64.84	14.60	64.65	14.97
Male	62.65	15.33	62.60	15.76
Language families				
Indo-European	66.67	16.08	66.41	16.42
Indo-Iranian	62.06	13.98	61.81	14.36
Dravidian	65.83	14.18	65.92	14.74
Other languages	59.76	16.77	60.45	17.36
Test center locations				
Inner circle	65.60	14.24	65.32	14.73
Outer-expanding circles	58.60	15.12	58.92	17.23

The Bayesian AMI approach consists of two steps: (1) setting different informative priors for the cross-group differences in loadings and intercepts, and (2) releasing equality constraints (on loadings and intercepts) that are not supported by the data (Zercher et al., 2015). Muthén and Asparouhov (2012) and later (Asparouhov, Muthén, & Morin, 2015) proposed a strategy that begins with very small prior variances (e.g., 0.001) and gradually increases them. Following this strategy and previous studies (Muthén & Asparouhov, 2013; van de Schoot et al., 2013; Pokropek et al., 2020), 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.

The analysis was conducted using the Bayes estimator and Gibbs sampling algorithm with two MCMC chains and 50,000 iterations. We estimated models with informative priors for differences between factor loadings and intercepts, ranging from N (0, 0.001) to N (0, 0.10). Model fit was assessed using the posterior predictive *p*-value (*ppp*) and the 95% confidence interval (CI), consistent with the approach applied in earlier analyses. We also used the Deviance Information Criterion (DIC), a comparative fit index for Bayesian analysis (Spiegelhalter et al., 2002), with lower values indicating better

model fit. Non-invariance in parameters is identified when a parameter in a specific group deviates from the overall group average, and this difference is considered significant if zero falls outside the 2.5% and 97.5% quantiles of the posterior distribution (Seddig & Leitgöb, 2018).

Results

Exploratory Factor Analysis Results

Models specifying one- and two-factor solutions were tested. The 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. Both models demonstrated a good fit: the one-factor model had a posterior predictive *p*-value (*ppp*) of 0.494 with a 95% confidence interval (CI) of [-288.317, +307.118], while the two-factor model yielded a *ppp* of 0.485 and a 95% CI of [-309.131, +293.918]. Based on the steep eigenvalue drop, interpretability, and parsimony, the one-factor solution was retained. This result suggests that the reading items in the PTE Core test have a unidimensional factor structure. Factor loadings ranged from 0.12 [95% CI: 0.05, 0.18] to 0.78 [95% CI: 0.75, 0.81], but 12 items had loadings below the commonly accepted threshold of 0.32 (Tabachnick & Fidell, 2013). However, they were retained for further analysis.

Confirmatory Factor Analysis Results

A one-factor CFA model was estimated. The model showed an acceptable fit, with a posterior predictive *p*-value (*ppp*) of 0.476 and a 95% confidence interval for the difference between observed and replicated chi-square values of [-310.773, +325.363]. Table 2 displays the posterior median estimates of factor loadings along with their 95% credibility intervals. Loadings ranged from 0.15 [95% CI: 0.07, 0.22] to 0.79 [95% CI: 0.75, 0.81]. While some items had lower loadings, most contributed meaningfully to the latent factor. Importantly, none of the credibility intervals included zero, confirming that all items significantly loaded onto the 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.

Table 2. Standardized factor loading estimations and associated 95% credibility intervals

Item	Factor loading	95% CI	Item	Factor loading	95% CI
1	0.40	[0.35, 0.47]	55	0.64	[0.60, 0.67]
2	0.15	[0.07, 0.22]	56	0.47	[0.43, 0.52]
3	0.52	[0.47, 0.56]	57	0.62	[0.59, 0.66]
4	0.45	[0.39, 0.51]	58	0.68	[0.64, 0.72]
5	0.29	[0.25, 0.33]	59	0.72	[0.69, 0.75]
6	0.52	[0.48, 0.56]	60	0.40	[0.35, 0.45]
7	0.34	[0.29, 0.39]	61	0.58	[0.53, 0.61]
8	0.47	[0.43, 0.51]	62	0.65	[0.62, 0.69]
9	0.27	[0.20, 0.33]	63	0.32	[0.26, 0.39]
10	0.30	[0.25, 0.35]	64	0.52	[0.46, 0.57]
11	0.37	[0.31, 0.43]	65	0.49	[0.43, 0.54]
12	0.59	[0.56, 0.62]	66	0.46	[0.40, 0.52]
13	0.47	[0.43, 0.50]	67	0.39	[0.32, 0.44]
14	0.42	[0.38, 0.46]	68	0.44	[0.40, 0.48]
15	0.48	[0.42, 0.53]	69	0.46	[0.42, 0.49]
16	0.42	[0.36, 0.47]	70	0.38	[0.34, 0.42]
17	0.47	[0.41, 0.54]	71	0.26	[0.21, 0.31]
18	0.44	[0.36, 0.50]	72	0.58	[0.54, 0.62]
19	0.47	[0.42, 0.52]	73	0.69	[0.66, 0.72]
20	0.42	[0.37, 0.48]	74	0.67	[0.64, 0.71]
21	0.58	[0.53, 0.63]	75	0.60	[0.56, 0.64]
22	0.33	[0.29, 0.37]	76	0.59	[0.55, 0.62]
23	0.17	[0.12, 0.22]	77	0.62	[0.58, 0.66]
24	0.43	[0.39, 0.47]	78	0.45	[0.40, 0.50]
25	0.38	[0.34, 0.43]	79	0.69	[0.66, 0.72]
26	0.25	[0.20, 0.29]	80	0.73	[0.69, 0.76]
27	0.32	[0.28, 0.37]	81	0.66	[0.62, 0.69]
28	0.34	[0.30, 0.38]	82	0.61	[0.57, 0.65]
29	0.39	[0.35, 0.44]	83	0.64	[0.61, 0.68]
30	0.33	[0.29, 0.38]	84	0.62	[0.58, 0.65]
31	0.44	[0.40, 0.48]	85	0.60	[0.55, 0.63]
32	0.29	[0.24, 0.33]	86	0.34	[0.27, 0.40]
33	0.54	[0.51, 0.58]	87	0.53	[0.47, 0.58]
34	0.49	[0.45, 0.53]	88	0.46	[0.39, 0.52]
35	0.60	[0.56, 0.63]	89	0.43	[0.37, 0.49]
36	0.58	[0.52, 0.62]	90	0.22	[0.17, 0.27]
37	0.48	[0.44, 0.52]	91	0.33	[0.29, 0.38]
38	0.66	[0.62, 0.69]	92	0.48	[0.44, 0.52]
39	0.63	[0.59, 0.66]	93	0.51	[0.47, 0.55]
40	0.63	[0.60, 0.67]	94	0.51	[0.47, 0.54]
41	0.49	[0.45, 0.53]	95	0.70	[0.66, 0.73]
42	0.63	[0.59, 0.66]	96	0.68	[0.64, 0.71]
43	0.66	[0.62, 0.70]	97	0.64	[0.60, 0.67]
44	0.60	[0.55, 0.64]	98	0.71	[0.68, 0.74]
45	0.49	[0.43, 0.53]	99	0.60	[0.56, 0.63]
46	0.65	[0.62, 0.68]	100	0.55	[0.51, 0.59]
47	0.58	[0.54, 0.62]	101	0.66	[0.62, 0.69]
48	0.71	[0.68, 0.75]	102	0.63	[0.59, 0.66]
49	0.69	[0.66, 0.72]	103	0.72	[0.69, 0.75]
50	0.78	[0.75, 0.80]	104	0.74	[0.70, 0.77]
51	0.79	[0.75, 0.81]	105	0.62	[0.58, 0.66]
52	0.63	[0.59, 0.66]	106	0.45	[0.41, 0.49]

53	0.64	[0.60, 0.68]	107	0.59	[0.55, 0.62]
54	0.59	[0.55, 0.63]			

Note. CI= Credibility Interval.

Approximate Measurement Invariance Results

Table 3 summarizes the model fit indices for Bayesian approximate measurement invariance (AMI) across gender, language family, and test center location. For gender, all models demonstrated good fit, with posterior predictive *p*-values (*ppp*) close to 0.50 and 95% confidence intervals included zero near the center. The model with a prior variance of $\mathcal{V}$ = 0.001 had the lowest DIC and was therefore selected, indicating that approximate scalar MI could be established across gender. For the language family, all models showed acceptable fit, but the model with $\mathcal{V}$ = 0.01 produced the lowest DIC value. Thus, the model with $\mathcal{V}$ = 0.01 was chosen, suggesting that approximate scalar MI could be supported. Finally, for test center location, all models similarly fit the data well, with *ppp* values near 0.50 and confidence intervals including zero at the center. The model with $\mathcal{V}$ = 0.005 yielded the lowest DIC and was selected, indicating that approximate scalar MI could also be established across test center locations.

Table 3. Model fit for Bayesian approximate scalar measurement invariance across gender, language family, and test center location

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

Note. DIC=deviance information criterion; *ppp*=posterior predictive *p*-value; CI=confidence interval

Table 4 presents the deviations of factor loadings and intercepts from the prior-defined values across female and male test-takers. Most item parameters did not show significant deviations, but for some items, either the factor loading or the intercept differed significantly. Notably, only two items (items 79 and 103) showed significant deviations in both parameters, suggesting non-invariance. Despite these item-level deviations, the overall model fit met acceptable standards. These findings support the conclusion that approximate scalar measurement invariance holds across gender, indicating that comparisons of latent means between female and male test-takers are supported.

Table 5 summarizes the deviations of parameters from the defined priors across language family groups. For most items, either the loading or intercept deviated in at least one group. No item showed significant deviations in both parameters across all groups. Still, only a few items remained invariant across all groups. Although model fit was acceptable, the high number of deviations suggests that while latent mean comparisons are feasible, interpretations should account for group-specific patterns.

Table 6 provides the deviations of parameters across test center location groups. While some items showed no significant differences from the prior-defined parameters, several (e.g., items 12 and 19) exhibited deviations in either loadings or intercepts. 16 items showed significant deviations in both parameters, indicating non-invariance. Despite the model's acceptable fit, the extent of these deviations across inner and outer-expanding circle groups suggests that approximate scalar measurement invariance can be assumed, though latent mean comparisons should be interpreted with some caution.

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 confirmed the one-factor structure of the reading test.

The Bayesian AMI results indicated that scalar invariance was achieved for gender, language family, and test center location, with best-fitting prior variances of 0.001, 0.01, and 0.005, respectively. For gender, most item loadings and intercepts showed minimal differences, though a few items had significant deviations. This pattern suggests 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.

For the language family, most items showed deviations in either loading or intercept within at least one group, but not both across all groups. These deviations 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.

For the test center location, while some items showed no deviations, many differed in either or both parameters. This heterogeneity likely mirrors 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 suggest approximate scalar invariance, supporting latent mean comparisons, though caution is needed, especially for language family and test center comparisons 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 measurement 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.

Table 4. Deviations of loadings and intercepts from prior defined parameters (mean = 0, variance = 0.001) for gender groups

	Female		Male			Female		Male	
Item	Lo	Int	Lo	Int	Item	Lo	Int	Lo	Int
1					55				
2					56		x		x
3					57				
4					58	x		x	
5					59				
6					60	x		x	
7		x		x	61				
8		x		x	62				
9					63		x		x
10					64				
11		x		x	65				
12	x		x		66				
13					67				
14					68				
15					69				
16					70		x		x
17					71				
18		x		x	72		x		x
19					73		x		x
20					74				
21					75				
22					76		x		x
23					77				
24					78				
25					79*	x	x	x	x
26					80				
27					81				
28					82				
29					83		x		x
30					84				
31		x		x	85				
32					86		x		x
33		x		x	87				
34					88	x		x	
35		x		x	89				
36					90		x		x
37					91				
38					92				
39		x		x	93				
40					94		x		x
41		x		x	95				
42					96				
43					97				
44					98				
45					99				
46					100				
47		x		x	101				
48					102				
49					103*	x	x	x	x
50					104				
51					105				
52					106				
53					107				
54									

Note. Lo = loading; Int = intercept; x=deviation of a given parameter in a given group from the defined priors (mean = 0, variance = 0.001); *=Non-invariant item.

Table 5. Deviations of loadings and intercepts from prior defined parameters (mean = 0, variance = 0.01) for language family groups

	Indo-European		Indo-Iranian		Dravidian		Other languages			Indo-European		Indo-Iranian		Dravidian		Other languages	
Item	Lo	Int	Lo	Int	Lo	Int	Lo	Int	Item	Lo	Int	Lo	Int	Lo	Int	Lo	Int
1									55			x	x	x	x	x	x
2			x	x				x	56				x		x	x	x
3								x	57		x		x	x	x	x	x
4				x				x	58		x			x	x		
5	x	x	x	x		x	x	x	59			x				x	
6	x	x	x	x		x			60	x		x	x			x	
7				x			x		61			x	x				
8			x					x	62		x				x		
9			x			x		x	63			x	x				x
10	x		x	x		x		x	64		x						
11			x						65								
12		x				x			66	x			x				x
13			x				x		67			x	x				
14									68				x				x
15								x	69	x		x	x	x		x	x
16									70		x		x				
17			x	x					71	x	x	x			x	x	
18				x					72		x	x					x
19						x		x	73						x		x
20						x		x	74								
21				x		x		x	75				x				x
22	x	x	x	x		x		x	76	x							
23				x					77						x		x
24	x	x	x	x		x		x	78		x		x			x	x
25	x	x		x		x		x	79	x	x	x	x		x		x
26	x						x	x	80				x				x
27		x	x			x		x	81								
28		x		x					82				x				x
29		x	x	x		x		x	83	x	x						
30	x	x	x	x		x		x	84				x		x		x
31	x					x	x	x	85								
32							x	x	86	x							
33	x	x						x	87								
34		x		x		x		x	88			x	x				x
35				x				x	89				x				x
36			x	x					90	x		x	x				x
37									91						x		x
38								x	92								
39	x	x		x				x	93			x					
40	x								94	x			x		x		
41		x				x		x	95								
42								x	96								
43	x						x	x	97		x				x		x
44							x	x	98								
45		x	x	x		x		x	99				x				x
46			x			x	x	x	100			x	x				
47			x				x		101								
48	x			x		x		x	102				x				

49	x	x	x	x	x	x	103	x	x	x	x
50	x	x		x			104		x		x
51			x	x		x	105				
52	x	x					106		x	x	
53				x	x	x	107			x	x
54						x					

Note. Lo = loading; Int = intercept; x=deviation of a given parameter in a given group from the defined priors (mean = 0, variance = 0.01).

Table 6. Deviations of loadings and intercepts from prior defined parameters (mean = 0, variance = 0.005) for test center location groups

	Inner circle		Outer-expanding circle			Inner circle		Outer-expanding circle	
Item	Lo	Int	Lo	Int	Item	Lo	Int	Lo	Int
1					55*	x	x	x	x
2	x		x		56	x		x	
3					57	x		x	
4	x		x		58*	x	x	x	x
5	x		x		59				
6*	x	x	x	x	60*	x	x	x	x
7					61*	x	x	x	x
8*	x	x	x	x	62*	x	x	x	x
9					63	x		x	
10					64	x		x	
11					65				
12		x		x	66	x		x	
13*	x	x	x	x	67		x		x
14		x		x	68				
15					69	x		x	
16					70		x		x
17	x		x		71		x		x
18					72*	x	x	x	x
19	x		x		73				
20					74				
21		x		x	75				
22		x		x	76				
23					77	x		x	
24	x		x		78				
25		x		x	79*	x	x	x	x
26		x		x	80				
27	x		x		81	x		x	
28*	x	x	x	x	82				
29	x		x		83*	x	x	x	x
30	x		x		84	x		x	
31					85				
32					86	x		x	
33					87				
34					88				
35					89				
36		x		x	90	x		x	
37					91	x		x	
38					92				
39					93				
40*	x	x	x	x	94*	x	x	x	x
41	x		x		95	x		x	
42					96				

43*	x	x	x	x	97				
44		x		x	98				
45		x		x	99				
46					100		x		x
47		x		x	101				
48		x		x	102				
49*	x	x	x	x	103	x		x	
50	x		x		104		x		x
51		x		x	105				
52					106				
53					107				
54									

Note. Lo = loading; Int = intercept; x=deviation of a given parameter in a given group from the defined priors (mean = 0, variance = 0.005); *=Non-invariant item.

References

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

Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. *Journal of Management, 41*, 1561–1577. doi:10.1177/0149206315591075

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).

Brown, T. A. (2015). Confirmatory factor analysis for applied research. Guilford.

Brown, G. T., Harris, L. R., O' Quin, C., & Lane, K. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research and Method in Education, 40*(1), 66-90. https://doi.org/10.1080/1743727X.2015.1070823

Byrne, B. M., & van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*(2), 107-132. https://doi.org/10.1080/15305051003637306

Byrne, B. M., Shavelson, R., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466. https://doi.org/10.1037/0033-2909.105.3.456

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

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*, 982. https://doi.org/10.3389/fpsyg.2014.00982

Depaoli, S., Winter, S. D., & Visser, M. (2020). The importance of prior sensitivity analysis in Bayesian statistics: Demonstrations using an interactive Shiny App. *Forntiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.608045

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

Hoofs, H., van de Schoot, R., Jansen, N. W., & Kant, I. (2018). Evaluating model fit in Bayesian Confirmatory Factor Analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement, 78*(4), 537-568. https://doi.org/10.1177/0013164417709314

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

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach.* Chichester, England: Wiley.

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., Roover, K. D., . . . Schoot, R. V. (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.

Seddig, D., & Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: Concept and application with panel data. *Survey Research Methods, 12*(1), 29-41. https://doi.org/10.18148/srm/2018.v12i1.7210

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583–639. https://doi.org/10.1111/1467-9868.00353

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th edition ed.). Pearson Education.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*(3), 842-860. https://doi.org/10.1111/cdev.12169

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

Winter, S. D., & Depaoli, S. (2019). An illustration of Bayesian approximate measurement invariance with longitudinal data and a small sample size. *International Journal of Behavioral Development, 44*(4), 371-382. https://doi.org/10.1177/0165025419880610

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

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance. *Frontiers in Psychology, 7*, 733. https://doi.org/10.3389/fpsyg.2015.00733