Psychological Test and Assessment Modeling, 2025.67:127-140
DOI: https://doi.org/10.2440/001-0021

Modeling Local Item Dependence in PIRLS 2016: Comparing ePIRLS and Paper-Based PIRLS Using the Rasch Testlet Model

Purya Baghaei[1], *Hamdollah Ravand*[2] & *Rolf Strietholt*[1]

[1] International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany
[2] Vali-e-Asr University of Rafsanjan, Iran

Abstract:
Testlets are groups of items linked by a common theme or stimulus, such as a reading passage or a diagram. While testlets enhance testing efficiency and contextual relevance, they can also violate the local independence assumption underlying item response theory models. Such violations may introduce bias in parameter estimates and inflate reliability coefficients artificially. This study investigates and compares testlet effects in the PIRLS and ePIRLS 2016 assessments. A total of eleven tasks—comprising both ePIRLS and paper-based PIRLS items—were analyzed using data from seven countries. For each country, both the standard unidimensional Rasch model and the Rasch testlet model were fitted separately. The results indicated that testlet variances were generally low across all seven countries. Nonetheless, model fit indices, including the deviance statistic and information criteria, consistently favored the Rasch testlet model over the unidimensional model. The reliability of the general dimension was lower in the testlet model which is consistent with expectations when accounting for local dependence. Importantly, correlations between item difficulty and person ability estimates from the two models were uniformly high (r = .99), suggesting that the unidimensional model still yields reasonable approximations in the presence of moderate testlet effects. Additionally, the findings revealed that paper-based tasks generated higher levels of local dependence than ePIRLS tasks. Potential explanations for this difference, along with the broader implications for test design and validity, are discussed.

Keywords: PIRLS, local dependence, testlet, Rasch testlet model

Correspondence: Purya Baghaei; International Association for the Evaluation of Educational Achievement (IEA), Überseering 27, 22297 Hamburg, Germany, purya.baghaei@iea-hamburg.de

Introduction

Testlets are a cluster of items around the same input or stimuli (Rosenbaum, 1988). For example, to assess reading comprehension a text passage with some questions based on it are administrated. This set of items is considered a testlet. Testlets are very common in educational testing as they help to simulate a more naturalistic and contextualized testing environment. Furthermore, testlets lead to more time-efficient and economical assessment and make the measurement of higher-order skills, which are hard to measure with atomistic decontextualized items, possible (Wainer et al., 2007). Therefore, testlets improve the validity of tests and of any uses and interpretations that are made based on test scores.

Although testlets have many advantages and are commonly employed in assessments, they create certain technical issues as far as unidimensional item response theory (IRT) models are concerned. Dependence of items on a shared prompt may lead to local item dependence (LID) which is a violation of one of the basic assumptions of IRT models. The assumption of conditional independence in IRT models stipulates that items should be independent after the influence of the target dimension is accounted for. Dependence between items, above and beyond, the latent trait is an indication that the test measures a secondary dimension which is not intended to be measured (Baghaei 2010; Eckes & Baghaei, 2015). This secondary dimension could be prior or special knowledge such as the theme of a text passage which enables some examinees to perform better than others with the same level of ability (Li, 2017). In other words, the violation of the local independence assumption is a form of the violation of unidimensionality (Baghaei & Christensen, 2023; Edwards et al., 2018). According to Marais and Andrich (2008), LID is generated in two ways. First, there is a secondary dimension that is measured by the test in addition to the target latent trait. This type of LID is more likely to occur when items are clustered under shared stimuli. Second, responses to initial items may affect responses to the items that follow later. This type of LID may occur even in the case of allegedly independent items.

Application of unidimensional IRT models to testlet-based items may violate the conditional independence assumption of IRT models. Research shows that LID can result in biased parameter estimates, errors in test equating, and spuriously high reliability coefficients (Eckes, 2014; Sireci et al., 1991; Wang & Wilson, 2005a, 2005b). Zhang (2010) in the context of a large-scale EFL (English as a Foreign Language) certification test demonstrated that testlet effects heavily influenced the accuracy of the language proficiency classifications.

To account for LID, some multidimensional IRT models, known as testlet response theory models (TRT), have been developed (Bradlow et al., 1999; Wainer et al., 2000; Wang & Wilson, 2005a). In these models, testlets are considered as specific dimensions and the interaction between examinees and testlets are modeled as random variables. TRTs are, in fact, bifactor IRT models where the specific factors are testlets. The testlets' variances show the strength of the local dependence (testlet effect) within each testlet. Numerous simulation and empirical data studies have shown that TRT

models have good model fit and recover the parameters of testlet-based items with adequate accuracy compared with unidimensional IRT models (DeMars, 2006; Li, 2017; Jiao et al., 2005; Wainer et al., 2007).

Rasch Testlet Model

Wang and Wilson (2005a) extended the testlet response theory to the family of Rasch models (Rasch 1960/1980) and introduced the Rasch testlet model (RTM) for both dichotomous and polytomous items. They argued that Rasch models provide important measurement advantages, including the existence of a sufficient statistic and the capacity to yield robust estimates with small samples. RTM is formally presented as shown in Eq. 1, where θn is the ability of person n, bi is the difficulty of item i, and γnd(i) refers to the random effect for person n on testlet d(i) and describes the interaction between persons and items (local item dependence) within the testlet (Wang & Wilson, 2005a). If the items are independent and no local dependence is assumed, i.e., $y_{nd(i)} = 0$, the model in Eq. 1 reduces to the standard dichotomous Rasch model.

In RTM, items are fixed effects and, therefore, need no distributional assumptions. The ability parameter and the testlet variables are treated as random effects and are assumed to be normally distributed but no distributional assumptions are made for the variance of the testlet effects. TRT is a special case of a multidimensional Rasch model known as the multidimensional random coefficients multinomial logit model (Adams et al., 1997).

$$p_{ni1} = \frac{\exp(\theta_n - b_i + \gamma_{nd(i)})}{1 + \exp(\theta_n - b_i + \gamma_{nd(i)})}.$$

Progress in International Reading Literacy Study

The Progress in International Reading Literacy Study (PIRLS) is an international assessment designed to measure the reading ability of fourth graders. PIRLS is administered every five years by the International Association for the Evaluation of Educational Achievement (IEA). The goal of the study is to provide information on students' reading achievement and their ability to apply what they understand from texts in real-life contexts or projects. PIRLS focuses on two broad reading purposes referred to as "literary experience" (reading for pleasure) and reading to "acquire and use information" (Mullis & Martin, 2015). Four fine-grained reading and hierarchically ordered subskills are included within each of these two broad purposes for reading, namely, focus on and retrieve explicitly stated information, make straightforward

inferences, interpret and integrate ideas and information, and evaluate and critique content and textual elements (Mullis & Martin, 2015).

The fourth cycle of PIRLS was administered in 2016 and featured the ePIRLS, an electronic reading assessment that measures fourth-grade students' ability to comprehend and interpret online informational texts. It is an extension of traditional PIRLS, which assesses reading comprehension in printed formats. The ePIRLS evaluates how well students can read, interpret, and understand digital texts similar to websites. It includes hyperlinks, pop-ups, and multimedia elements to simulate real-world digital reading experiences. ePIRLS tests students' ability to navigate and extract information from digital environments. It was added to the PIRLS in response to the growing importance of digital literacy in education. It provides insights into how well students are prepared for the increasing shift toward online learning and digital information consumption.

Participation in ePIRLS was voluntary for the participating countries. Fifty countries took part in PIRLS 2016 while 14 of them chose to participate in ePIRLS (alongside the conventional paper-and-pencil PIRLS which was mandatory). The scores for paper-and-pencil PIRLS and the ePIRLS for the participating countries were reported separately.

The ePIRLS was initiated in 2016 in response to the widespread availability of online reading material and the significance of digital literacies. The ePIRLS, in contrast to traditional printed reading material, contains nonlinear reading tasks and comprises reading passages that are scattered across several webpages. Examinees have to navigate back and forth across the pages while there is also irrelevant information in the form of popups, links, and advertisements.

While both PIRLS and ePIRLS assessment materials are administered in testlets, the way the items are related to each other differs. The ePIRLS texts and questions are presented to test-takers in a fragmented and interactive manner, unlike traditional reading comprehension tasks. Instead of providing an entire passage at once and then presenting all the related questions together, ePIRLS structures the assessment into multiple steps, each focusing on a specific portion of the text. This makes the ePIRLS assessment somewhat different from the standard reading comprehension tests where an entire passage is presented in its entirety in a single block of text. Nevertheless, the ePIRLS passages are on the same theme and altogether make a coherent passage.

The PIRLS and ePIRLS assessments are composed of several testlets, each consisting of a text passage followed by 12 to 24 related items. PIRLS 2016 contained 12 testlets (5 were ePIRLS testlets) and a total of 226 items (Mullis & Prendergast, 2017). The 20 paper-based passages or tasks were presented in a rotated booklet design where each booklet contained two testlets. Likewise, the five ePIRLS testlets were also distributed in 12 task combinations each containing two testlets (Mullis & Martin, 2015).

Since item response theory (IRT) is the primary methodology used to scale PIRLS (Martin et al., 2017), its testlet-based structure raises concerns about potential violations of the local independence assumption if not explicitly modeled. This study

applies a testlet modeling approach within the Rasch model framework to evaluate the presence and magnitude of LID in PIRLS 2016, including both paper-based and ePIRLS tasks. Rasch models provide key psychometric advantages, such as sufficient statistics for parameter estimation and fewer distributional assumptions. In the RTM, items are treated as fixed effects, while person abilities (θ) and testlet effects (γ) are modeled as random variables with normally distributed values. No assumptions are made about the variances of the testlet effects ($\sigma^2\gamma$) (Wang & Wilson, 2005a).

We analyze data from seven countries—Denmark, Georgia, Italy, Portugal, Taiwan, UAE, and the USA—selected to reflect diverse regions and to assess the generalizability of findings. The central aim of the study is to compare the extent of LID generated by PIRLS and ePIRLS formats. Given that ePIRLS was first introduced in 2016, a detailed investigation of its testlet structure and associated LID is both timely and necessary for evaluating the validity of this digital assessment format.

Methodology

Data

Publicly available data from PIRLS 2016 were used for this study. PIRLS is an international large-scale assessment of reading comprehension of 4th grade students. Fifty countries and 11 benchmarking entities participated in PIRLS 2016. Fourteen countries and two benchmarking entities participated in the ePIRLS assessment while the rest of the countries participated only in the paper-based PIRLS (Mullis et al., 2017). For the purposes of this study, we focused on the five ePIRLS tasks and six paper-based tasks all measuring the 'acquire and use information' purpose of reading. Information about the texts is presented in Table 1. For each country a random sample of fourth grade students were drawn. The data from seven countries including Denmark (n=2506), Georgia (n=5557), Italy (n=3767), Portugal (n=4558), Taiwan (n=4326), United Arab Emirates (UAE, n=15566), and USA(n=4090) representing different geographical regions were analyzed. Items with partial credit scoring were recoded to dichotomous items (correct = 1; incorrect or partially correct = 0) to simplify the analyses.

Table 1
PIRLS 2016 Task/Testlet Information

Testlet	Text Name	No. Items	Word Count	Flesch-Kincaid	Mode
T1	Marse	24	–	–	Electronic
T2	Rainforests	24	–	–	Electronic
T3	Dr Elizabeth Blackwell	17	–	–	Electronic
T4	Zebra and wildebeest migration	24	–	–	Electronic
T5	The legend of Troy	22	–	–	Electronic
T6	Sharks	12	570	7.6	Paper
T7	Icelandic horses	15	870	5	Paper
T8	Leonardo da Vinci	12	869	5.1	Paper
T9	How did we learn to fly?	17	514	6.3	Paper
T10	Where is the honey?	16	870	3.2	Paper
T11	The green sea turtle	16	943	4	Paper

Note: The Flesch-Kincaid Grade Level Formula is a readability index that is based on average syllables per word and average sentence length and represents the US grade in which students can read the text. The table shows the index for the English version of the PIRLS test.

Analyses and Results

To examine testlet effect in the PIRLS 2016 reading assessment, two models were fitted to the data: a unidimensional dichotomous Rasch model (Rasch, 1960/1980) and a Rasch testlet model (RTM: Wang & Wilson, 2005a). In the unidimensional model, local dependence was ignored while in the RTM it was assumed that the items nested within a testlet generate LID. The data of the seven countries were analyzed separately using the 'TAM' package (Robitzsch et al., 2022) in R (R Core Team, 2022).

Table 2 shows the information criteria for the two estimated models across the seven countries. Deviance statistic or -2loglikelihood (-2LL), Akiak Information Criterion (AIC, Akaike, 1974), and the Bayesian Information Criterion (BIC, Schwartz, 1978) are presented. Models with smaller information criteria are selected as the better

model. As Table 2 shows, in all the seven countries the testlet model is preferred according to the deviance statistic and information criteria. Table 2 also shows that the reliability of the general dimension in the testlet model is relatively small compared to the reliability of the unidimensional Rasch model in all the countries. Correlation coefficients between the item difficulty parameters and the person parameters from the two models were .99 in all the countries. The mean of the absolute differences between the item parameters from the two models across the seven countries was .03 with a range of 0–.20.

Table 2
Information Criteria for the Two Models across Countries

	RTM					RM				
Country	-2LL	AIC	BIC	Par	Rel.	-2LL	AIC	BIC	Par	Rel.
Denmark	128312.3	128686	129776	187	.862	128547.2	128899	129925	176	.893
Georgia	288848.1	289220	290452	186	.825	290412.6	290763	291922	175	0.88
Italy	200139.3	200513	201679	187	.831	200957.2	201309	202406	176	.877
Portugal	240695.6	241070	242271	187	.848	241066.3	241418	242549	176	.884
Taiwan	220152.3	220526	221717	187	.860	220403.9	220756	221876	176	.889
UAE	791222.9	791597	793028	187	.888	793935.8	794288	795635	176	.918
USA	214440.4	214814	215996	187	.862	215037.1	215389	216501	176	.902

Note. One item in Georgia had a maximum score of 0 and had to be removed for the analyses.

Table 3 depicts the variances of the general reading dimeson (G) and the testlet dimensions (T1-T11) across the seven countries. To evaluate the strength of LID in a testlet, its variance should be compared with the variance of the general dimension. There are no commonly accepted cut-offs to judge the strength of the testlet variance. Nevertheless, in simulation studies, researchers have considered values of .25, .50, and .75 as small, moderate, and large testlet effects, respectively (Jiao et al., 2013; Wang & Wilson, 2005a, 2005b). Therefore, testlet variances smaller than .25 are considered negligible in this study (Zhang, 2010). Since the testlet variances are construct-irrelevant variances, the ideal is to have small testlet variances and a large general dimension variance. As is evident from Table 3, the general dimension has the largest variance in all the countries, except for Georgia. The variances of the testlet dimensions range from .05 to .93, which cover a wide range from very small to very high. Compared to the variance of the general dimension, some testlets have large variances which are nontrivial. For example, testlet 11 in Georgia has a variance of .93 which is greater than the variance of the main reading dimension. Testlet 11 has generated a large testlet effect in six countries which could be related to the

peculiarities of this text. Overall, the variances of other testlets are small with only a few testlets having a variance greater than .25 in some of the countries. Table 3 also shows that paper-based passages have generated a greater level of local dependence than ePIRLS passages. When the testlet effects for the tasks are summed across the countries and the mean of the sums for the paper-based testlets (M=1.71, SD=.83) and the ePIRLS testlets (M=1.15, SD=.71) are compared, a relatively large difference is observed.

Table 3
Variance Estimates of the Theta and Random Testlet Variables in RTM

Dimension	Denmark	Georgia	Italy	Portugal	Taiwan	UAE	USA	
G	1.052	0.899	0.858	0.911	1.006	1.496	1.137	Sum
T1	0.099	0.092	0.137	0.059	0.05	0.134	0.147	.71
T2	0.233	0.378	0.296	0.229	0.166	0.39	0.212	1.90
T3	0.298	0.423	0.268	0.157	0.219	0.353	0.251	1.96
T4	0.05	0.111	0.05	0.057	0.05	0.078	0.05	.44
T5	0.113	0.197	0.101	0.089	0.083	0.113	0.091	.78
T6	0.05	0.114	0.147	0.118	0.088	0.071	0.168	.75
T7	0.142	0.165	0.173	0.16	0.112	0.153	0.253	1.15
T8	0.149	0.302	0.251	0.258	0.162	0.233	0.11	1.46
T9	0.216	0.307	0.211	0.158	0.176	0.353	0.224	1.64
T10	0.116	0.43	0.38	0.24	0.234	0.295	0.416	2.11
T11	0.324	0.936	0.411	0.263	0.48	0.47	0.255	3.13
Mean (SD)	.16 (.09)	.31 (.24)	.22 (.11)	.16 (.07)	.16 (.12)	.24 (.13)	.19 (.10)	1.45 (.79)

Note: T1-T5 are electronic testlets and T6-T11 are paper-based testlets. 'Sum' shows the sum of the testlet effects across the seven countries (rows).

Discussion and Conclusion

In this research, local dependence in PIRLS 2016 was modeled using the Rasch testlet model (Wang & Wilson, 2005a). The TAM package (Robitzsch, 2022) in R (R Core Team, 2024) was employed to estimate the Rasch testlet model and the unidimensional dichotomous Rasch model. Findings revealed that the magnitude of LID is relatively small in PIRLS and ePIRLS 2016. However, the Rasch testlet model, where LID is accounted for, fitted better than the unidimensional Rasch model where LID is

ignored. The testlet model showed that testlet effect values were small (<.25) for most of the passages. A few PIRLS and ePIRLS testlets exhibited testlet effects greater than the critical value recommended in the literature. The means of the testlet variances across countries ranged from .16 to .31. This is slightly greater than the lowest critical value of .25 and suggests an overall small testlet effect for PIRLS and ePIRLS.

The findings showed that only one paper-based testlet—Testlet 11—exhibited a very high level of LID in Georgia and exceeded the critical threshold of 0.25 in the other countries as well. This testlet contained the longest passage and the second easiest passage among the six paper-based testlets. This suggests that longer and easier passages may generate more local dependence. Testlet 6, which is the hardest and the shortest of paper PRLS tasks generated a low level of LID which strengthens the hypothesis that LID could be a function of text length and difficulty. It seems that easier and longer items generate more LID, but this conjecture needs further investigations. Our analysis is based on existing data, and the testlets may vary not only in terms of difficulty and length but also in terms of other, unobserved characteristics. As such, we exercise caution in interpreting these results and refrain from drawing definitive conclusions at this stage. Unfortunately, no information about the length and difficulty of the ePIRLS tasks is recorded in the PIRLS documentation to examine this hypothesis across the ePIRLS tasks too.

An important qualification to the length-and-difficulty hypothesis arises when comparing Testlet 10 to Testlet 11. Despite being easier and almost as long as Testlet 11, Testlet 10 produced moderate LID values (above .25) in four countries, yet never approached the magnitude observed for Testlet 11. This discrepancy suggests that text length and surface difficulty alone are not sufficient to induce high LID. Instead, it is likely that Testlet 11 possesses greater semantic cohesion, referential overlap, or structural redundancy—features that encourage students to cross-reference and integrate responses across items, a core mechanism of response dependence (Marais & Andrich, 2008; Wang & Wilson, 2005). In contrast, Testlet 10, although easier, may have featured more modular item design with less overlap in content, thereby promoting independent answering. This finding refines the hypothesis, emphasizing that LID emerges not just from text length or difficulty, but from the interaction of cognitive demands, textual structure, and item interdependence (Chen et al., 2023). Findings also suggest that testlet effect is independent of the number of items nested within a testlet. Testlet 1, 2, and 4 have the highest number of items each with 24 items. Testlets 1 and 4 have generated the least amount of LID across the countries.

Another finding of this research is that paper-based PIRLS tasks have generated more LID than ePIRLS tasks. We argue that ePIRLS reduces LID because it fragments the text and presents questions in separate stages, making each question more isolated from the others. On the other hand, traditional paper-based reading tests promote more interdependence because examinees have access to the entire passage and all the questions at once, which might lead them to approach the questions in a more interconnected manner. In ePIRLS, each question is tied to a specific portion of the text. Since students only see one portion of the text at a time, and each question is directly linked

to that portion, they do not need to consider the broader passage. This compartmentalization could reduce item dependence because each question is mostly self-contained. In the paper-based PIRLS, questions are answered with full context in mind. Examinees see the entire passage and all the questions at once, allowing them to cross-check between the items. If a student realizes their answer to one question contradicts another, it is easier for them to go back and adjust their responses, increasing interdependence among the items.

Marais and Andrich (2008) identified two types of local item dependence: *trait dependence* and *response dependence*. Trait dependence, often referred to as multidimensionality in the literature, arises when a test measures an additional trait or dimension beyond the intended one. In contrast, response dependence, also known as the *item chaining effect* (Wang & Wilson, 2005a), occurs when response to one item influences how subsequent items are answered. The ePIRLS' structure and stepwise presentation could reduce local item dependence because the possibility of item chaining is eliminated or mitigated to a great extent. Meanwhile, traditional reading tests might intensify LID since all questions are presented together, leading to more interdependent answering strategies or response dependence.

This interpretation aligns with broader cognitive theories. For example, the construction-integration model (Kintsch, 1988) suggests that segmented processing lowers cognitive burden, which in turn reduces the probability of strategic inter-item reasoning. Interface constraints in ePIRLS (e.g., lack of backtracking, isolated screens) further restrict students from revisiting earlier items, thereby limiting opportunities for LID to occur.

Findings showed that the reliability of the single reading factor in the unidimensional Rasch model, where LID is ignored, is greater than the reliability of the general reading factor in the testlet model in all the countries. This is evidence that local dependence among the items spuriously increases test reliability and gives a distorted picture of the test's precision.

Differences in parameter estimates were also examined across the countries. Correlation coefficients between the item difficulty parameters and ability parameters from the two models were .99 in all the countries. The mean of the absolute differences between the item parameters from the two models across the seven countries was .03 with a range of 0–.20. This finding suggests that the item difficulty and person ability parameters are not affected very much by local dependence. It seems that although moderate local dependence for some testlets dramatically impacts test reliability, it has a minimal effect on the item and person parameters. The findings are in line with Chang and Wang (2010) who examined local dependence in PIRLS 2006 for Taiwan using the 3-PL testlet response theory (Wainer et al., 2000). Their findings revealed that testlet effects ranged from .168 to .489. They also showed that while item difficulty and person parameters are not affected by LID, there is bias in discrimination and guessing parameters when LID is ignored. This agrees with research (outside the context of PIRLS) on the impact of LID on item and person parameters when the

Rasch and 2-PL testlet response models are used (Baghaei & Ravand, 2016; Eckes & Baghaei, 2015).

An interesting finding in this study is that the passages have almost the same level of testlet effects across the countries. That is, a testlet with a small testlet effect, like Testlet 4 with testlet variance of .05 in Denmark, has also very small effects in other countries and a testlet like Testlet 11 with testlet variance of .93 in Georgia has also very large testlet effect values in other countries. This implies that there must be some textual features possibly including narrative structure, discourse cohesion, or content familiarity, as sources of LID. Such a hypothesis could be further examined through testlet response models with covariates (Wainer et al., 2007).

Future Directions

To deepen understanding of local item dependence (LID) and mitigate its impact, several promising research avenues are proposed. First, linguistic and cognitive profiling of high-LID testlets is essential to isolate specific textual features—such as lexical density, referential cohesion, and semantic overlap—that drive inter-item associations (Graesser et al., 2004; McNamara et al., 2014). The importance of these textual properties is underscored by Kintsch's (1988) construction-integration model, which emphasizes coherence as a key determinant in reading comprehension. Second, student response process data should be incorporated to reveal how test-takers engage with texts and items. Methods such as eye-tracking, think-aloud protocols, and digital navigation logs can offer insight into when and how response chaining or item referencing occurs. Third, the use of Natural Language Processing (NLP) tools, such as Coh-Metrix, can help operationalize textual complexity indices and include them as covariates in testlet models, providing a more systematic way to quantify and predict LID. Finally, it is crucial to compare LID patterns across other large-scale international assessments such as PISA and NAEP. These comparisons would help assess the generalizability of LID effects under different testlet designs and delivery modes.

Conclusion

This study affirms that local item dependence is a significant measurement issue in reading assessments involving shared stimuli. While most PIRLS and ePIRLS tasks exhibited minor LID, select passages—particularly longer, easier, and more cohesive texts—produced substantial inter-item dependencies. The case of Testlet 11 in Georgia underscores how linguistic, instructional, and cognitive factors can jointly amplify LID. The study also confirmed that digital assessments like ePIRLS mitigate LID through segmented design and constrained navigation. Modeling LID explicitly through testlet models is essential not only for improving reliability estimation but also for understanding the cognitive and textual mechanisms behind inter-item dependencies. Integrating text analysis, response behavior data, and cross-platform

comparisons offers a comprehensive path forward for improving the validity, fairness, and psychometric robustness of international reading assessments.

References

Adams, R. J., Wilson, M. R., & Wang, W. L. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–24. http://dx.doi.org/10.1177/0146621697211001

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705

Baghaei, P., & Christensen, K. B. (2023). Modeling local item dependence in C-tests with the loglinear Rasch model. *Language Testing*, *40*(3), 820–827. https://doi.org/10.1177/02655322231155109

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37*(1), 85–104. https://www.uv.es/psicologica/articulos1.16/5BAGHAEI.pdf

Chang, Y., & Wang, J. (2010, July 1–3). *Examining testlet effects on the PIRLS 2006 assessment*. [Conference presentation]. The 4th IEA International Research Conference, University of Gothenburg, Sweden. https://www.iea.nl/publications/presentations/examining-testlet-effects-pirls-2006-assessment

Chen, Q., Zheng, H., Fan, H. & Mo, L. (2023). Construction of a reading literacy test item bank for fourth graders based on item response theory. *Frontiers in Psychology, 14*:1103853. https://doi.org/10.3389/fpsyg.2023.1103853

DeMars, C.E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145–168. https://doi.org/10.1111/j.1745-3984.2006.00010.x.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39–61. https://doi.org/10.1177/0265532213492969

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education, 28*(2), 85–98. https://doi.org/10.1080/08957347.2014.1002919

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods, 23*(1), 138–149. https://doi.org/10.1037/met0000121

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202. https://doi.org/10.3758/BF03195564

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*(2), 186–203. http://www.jstor.org/stable/24018106

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, *6*, 311–321.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Li, F. (2017). An information-correction method for testlet-based test analysis: From the perspectives of item response theory and generalizability theory. *ETS Research Report Series*, 2017(1), 1–25. https://doi.org/10.1002/ets2.12151

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(2), 105–124.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and procedures in PIRLS 2016*. IEA. https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd Ed.). IEA. https://timssandpirls.bc.edu/pirls2016/framework.html

Mullis, I. V. S., & Prendergast, C. O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 1.1–1.29). IEA. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-1.html

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. IEA. http://timssandpirls.bc.edu/pirls2016/international-results/

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chicago: The University of Chicago Press, 1980).

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules*. R package version 4.1-4. https://CRAN.R-project.org/package=TAM

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*(3), 349–359. https://doi.org/10.1007/BF02294217

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464. https://doi.org/10.1214/aos/1176344136

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL model useful in testlet-based adaptive testing. In W. J. Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Springer. https://doi.org/10.1007/0-306-47531-6_13

Wang, W.-C., & Wilson, M. (2005a). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149. https://doi.org/10.1177/0146621604271053

Wang, W.-C., & Wilson, M. (2005b). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*, 296–318. https://doi.org/10.1177/0146621605276281

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), 119–140. https://doi.org/10.1177/0265532209347363