Education and Science

Original Article

2025 Supplement 1 111-127

Using Testlets in Education: eTIMSS 2019 as an Example *

Kübra Atalay Kabasakal [1], Sebahat Gören [2]

Abstract

A testlet refers to groups or clusters of items that are linked to a common stimulus such as a text, graphic or table. Due to the shared stimulus among these items, there is a high likelihood of inter-item dependency within the responses, which violates the assumption of local independence in Item Response Theory (IRT). This violation results in local dependence among the items within the testlets. Therefore, this study employed IRT and the Testlet Response Theory (TRT) models to assess the impact of local dependence stemming from testlets on item and ability parameter estimations, classification accuracy, and Differential Item/Bundle Functioning (DIF/DBF), and compared the findings obtained from these models. Responses to three testlets that were both in booklets 13 and 14 in the eTIMSS 2019 mathematics subtest were analysed using the *mirt* package in R software. The analysis revealed a moderate degree of local dependence in the testlets. Additionally, a very high correlation was observed between the item and ability parameter estimations derived from both models. Regarding classification accuracy, the IRT and TRT models demonstrated equivalent performance. When items were analysed both independently and as part of testlets, no items exhibited evidence of DIF/DBF based on gender. The findings indicate that IRT can tolerate the effects of testlets when the degree of local dependence is low to moderate.

Keywords

Testlet

Testlet response theory

Item parameter estimation

Ability estimation

Classification accuracy

Differential item/bundle functioning

Article Info

Received: 10.05.2024

Accepted: 12.30.2024

Published Online: 03.03.2025

DOI: 10.15390/EB.2025.14104

Introduction

The assessment of student performance has long been a central focus of educational research. Accurate estimation of item difficulty and student ability is crucial for effective teaching and assessment practices. However, the estimation of item difficulty and ability parameters can vary depending on the type of the test, the nature of the items included, the structure of the test, and the group to which it is administered. Testlets or clusters, which consist of items sharing a common stimulus such as a reading comprehension passage, a figure or a graph, are referred to as testlets (Wainer & Kiely, 1987). Testlets allow for a more detailed modelling of factors influencing student responses such as contextual information and cognitive processes. Koziol (2016) states that the purpose of using testlets in a test is to

* A part of this study was presented at the International Symposium on Measurement, Selection and Placement held between 4-6 October 2024 as an oral presentation.

[1] Hacettepe University, Faculty of Education, Department of Educational Sciences, Türkiye, kkatalay@gmail.com

[2] Kütahya Dumlupınar University, Faculty of Education, Department of Educational Sciences, Türkiye, sebahatgoren@gmail.com

capture performance beyond what is explained by the latent trait of interest. Furthermore, the use of testlets aims to better assess higher-order skills (DeMars, 2010; Wainer & Wang, 2000).

From the perspective of test development, testlets not only bring together more complex and interrelated items but also contribute to improving test efficiency (Thissen, Steinberg, & Gerrard, 1986). Specifically, groups of items organized within a testlet allow test participants to respond to multiple items linked to a common stimulus, providing time and effort efficiency (Ho & Dodd, 2012; Wainer & Wang, 2000). Furthermore, testlets can help reduce issues related to variance irrelevant to the construct by anchoring item responses to a shared stimulus, potentially enhancing the validity of inferences made about test-takers' abilities. This is because individuals only need to evaluate the content associated with the shared stimulus once and can then use this information across all items in the testlet. For these reasons, the use of multiple-choice test formats incorporating one or more sets of testlets referencing a common text is particularly prevalent in assessing foreign language skills. In Türkiye, national standardized exams are administered by the Centre for Assessment, Selection and Placement (ÖSYM). Due to the advantages they offer, testlets are frequently employed in both exams conducted by ÖSYM (e.g., ALES, KPSS, YDS, e-YDS, YKS, YÖKDİL) and international assessments (e.g., GRE, IELTS, SAT, TOEFL). Integrating similar testlet applications into K-12 test development processes in Türkiye could enhance both the validity and efficiency of measurement and evaluation practices. Considering that primary and middle school students have limited attention spans, testlets based on a common stimulus could save time and enable students to proceed through tests in a more focused manner.

Testlets, due to their frequent use and the advantages they offer, have necessitated the application of various methods based on Item Response Theory (IRT) in addition to Classical Test Theory (CTT) to examine the validity and reliability of test scores. IRT analyses have gained increasing importance both nationally and internationally as a robust tool for enhancing the validity and reliability of tests. This is because IRT models not only evaluate test scores but also model the latent ability levels underlying responses to individual items, thus helping to estimate students' abilities more accurately (Embretson, 2010; Hambleton & Rogers, 1989). Weiss (1982) highlighted that the use of different IRT methods has grown with the widespread adoption of Computerized Adaptive Testing (CAT) designs, which adapt the difficulty level of test items to an individual's level of ability, allowing for more precise estimations of ability. Although IRT provides a powerful statistical framework for modeling the relationship between an individual's ability and their responses to test items, its application relies on several fundamental assumptions (Embretson, 2010). IRT models are primarily categorized into unidimensional and multidimensional models. The first assumption of unidimensional IRT is that the latent trait being measured is unidimensional. Secondly, unidimensional IRT assumes that the probability of a correct response to an item is a monotonically increasing function of an individual's level of ability. This function, known as the item characteristic curve, describes the relationship between the latent trait and the probability of a specific response (Hambleton & Rogers, 1989). The third assumption of unidimensional IRT is that responses to different test items are locally independent, meaning that the probability of responding correctly to one item is unaffected by responses to other items, given the individual's ability level. When these assumptions are met, unidimensional IRT enables the estimation of item parameters (such as difficulty and discrimination) and individual ability parameters, providing more accurate and reliable test scores and supporting the development of adaptive tests (Hambleton & Rogers, 1989). Despite these appealing features, testlets can lead to violations of the local independence assumption in unidimensional IRT during parameter estimations. Local independence is a critical assumption in unidimensional IRT models; however, in practice, items within a testlet often exhibit correlated responses even after the latent trait is controlled (Koziol, 2016). IRT tends to overestimate the precision of ability estimations and test reliability when applied to testlets, leading to biased estimations of item difficulty and discrimination parameters (Eckes & Baghaei, 2015;

Sireci, Thissen, & Wainer, 1991). Furthermore, when the local independence assumption is not controlled in tests composed of testlets, there may arise errors in test equating, linking and classification accuracy (Keller, Swaminathan, & Sireci, 2003; Lee, Kolen, Frisbie, & Ankenmann, 2001; Li, Bolt, & Fu, 2006). Additionally, the use of testlets in CAT applications allows for the control of contextual and ordering effects (Wainer, Bradlow, & Wang, 2007). However, when unidimensional IRT models are applied to CAT systems with testlets, violations of the local independence assumption can lead to an overestimation of item/testlet information functions (Thissen, Steinberg & Mooney, 1989). For this reason, in CAT applications composed of testlets, the use of Testlet Response Theory (TRT) models is more appropriate to achieve greater accuracy in ability estimation and measurement precision.

In tests containing testlets, two methods have been proposed to address violations of the local independence assumption. One method involves treating all items within a testlet as polytomous items (super-item) and using a model suitable for unidimensional polytomous items for parameter estimation (Cook, Dodd, & Fitzpatrick, 1999; Sireci et al., 1991; Yen, 1993; Wainer, 1995). This method is appropriate in situations where the degree of local dependence among items in a testlet is moderate, and the test predominantly consists of independent items (Wainer, 1995). However, this approach is impractical because the number of possible response patterns increases geometrically with the number of items in a testlet, making it rarely used in practice (Thissen et al., 1989). Additionally, since the total score of the items in the testlet is considered, it may result in a loss of information (Wainer & Lewis, 1990). An alternative method is to account for the effects of testlets by incorporating specific dimensions in addition to the general dimension within IRT models. Such multidimensional IRT models are frequently employed by researchers. These include bifactor models (Gibbons & Hedeker, 1992) and random-effects testlet response models (Bradlow, Wainer, & Wang, 1999; Wainer et al., 2007). Li and others (2006), Rijmen (2010), and Min and He (2014) have noted that random-effects testlet models can be used as a special case of bifactor models. This is achieved by constraining the loadings on the specific dimension to be proportional to the loadings on the general dimension within each testlet. In summary, tests consisting of testlets require the use of complex models such as the Testlet Response Theory (TRT), which incorporate an additional parameter that accounts for the level of local dependence and specifies the individual-specific amount of local dependence within each testlet (Wainer, Bradlow, & Du, 2000).

Researchers have often focused on parameter and ability estimation using unidimensional IRT, TRT and bifactor models in tests comprising testlets (Baghaei & Ravand, 2016; DeMars, 2006; Soysal & Yılmaz Koğar, 2022; Yılmaz Koğar, 2021). The number of studies in which DBF analysis was performed on testlets is quite small (Paek & Fukuhara, 2015; Tasdelen Teker & Dogan, 2015; Wainer, 1995). In this study, we examined parameter estimation, accurate classification performance of test participants, and the presence of Differential Item Functioning (DIF) or Differential Bundle Functioning (DBF) within items or testlets according to IRT and TRT models. Accurate classification of students is particularly critical in educational settings. Hence, this study investigated how local dependence in tests comprising testlets affects classification accuracy and how this accuracy varies across binary and multi-category classifications. Another crucial issue considered is the potential for test participants from specific subgroups to exhibit differing performance patterns on testlets due to factors beyond item difficulty. If not accounted for, this phenomenon, known as Differential Item Functioning (DIF), can lead to biased parameter estimates and misclassification of student proficiency. Therefore, conducting DIF studies that account for testlets and analysing this issue comparatively are expected to contribute significantly to the literature. Furthermore, the modelling of tests comprising testlets has predominantly been investigated using data from assessments such as PISA, SAT or simulation studies (Chang & Yang, 2010; Koziol, 2016; Yılmaz Koğar, 2021). However, applying these models to diverse real-world datasets could provide researchers with more detailed insights and greater opportunities for comparisons. Accordingly, this study employed the TIMSS-2019 dataset to perform parameter estimation using

traditional IRT and TRT models to analyse and compare classification accuracy and DIF results. Existing studies on classification accuracy have primarily focused on binary classifications such as pass-fail outcomes, yet their number remains limited (Koziol, 2016; Zhang, 2010). Comparing multi-category classifications and DIF outcomes in testlets across different models (IRT-TRT) using the TIMSS dataset is expected to provide a unique contribution to the literature. Additionally, discussing the findings of this study in light of other studies in the field will help establish a broader perspective on classification accuracy and DIF analyses in tests comprising testlets. To achieve these objectives, the study explored the research questions by analysing responses to three testlets included in both Booklet 13 and Booklet 14 of the e TIMSS 2019 assessment in the Turkish sample:

1. What is the level of local dependence among the three testlets in the mathematics subtest?
2. How is the relationship between the item and ability parameters obtained from the 2PL-IRT and 2PL-TRT models in the mathematics subtest consisting of testlets?
3. How is the classification accuracy obtained from the 2PL-IRT and 2PL-TRT models in the mathematics subtest consisting of testlets?
4. Are there any items containing gender based DIF/DBF in the mathematics subtest consisting of testlets?

Method

Research Type

In this study, we analyzed a test consisting of testlets in terms of parameter estimation, classification accuracy and DIF/DBF using different IRT models. This study is descriptive research that provides more information about the current situation by thoroughly comparing the results obtained from different methods (Creswell, 2014; Karasar, 2016).

Study Group

Trends in International Mathematics and Science Study (TIMSS) is an achievement monitoring study conducted every four years by the International Association for the Evaluation of Educational Achievement (IEA). TIMSS, first conducted in 1995, is carried out at four-year intervals and is an important international research study. TIMSS aims to assess the achievements of students in mathematics and science at the fourth and eighth grade levels. In the 2019 cycle, the sample from Türkiye, one of the 39 participating countries at the eighth-grade level, consisted of 4.077 students from 181 schools. In the 2019 TIMSS application, there was a shift to computer-based assessment (eTIMSS). This study used three testlets that were both in Booklets 13 and 14, from the Turkish sample of eTIMSS 2019. These testlets consist of two, four and six items respectively. In the analysis, we used listwise deletion for 89 participants with missing data. After the deletion of missing data, we used the responses of a total of 503 students, of which 47.9% were female and 52.1% were male.

Data Analysis

We examined the model-data fit within both IRT and TRT frameworks and found out that the best fit was achieved with the 2PL model. The results showed that the item discrimination and intercept parameters in the 3PL and 3PL TRT models exhibited unusual values. As a result, the analyses proceeded with the 2PL IRT and 2PL TRT models. For the three testlets in both Booklets 13 and 14 in the TIMSS 2019 application, we made estimations of and compared item and ability parameter values according to the 2PL-TRT (Wainer et al., 2007) and 2PL-IRT (Birnbaum, 1968) models. In this study, we considered item parameters such as slope and intercept coefficients. We obtained ability estimates using the Expected A. Posteriori (EAP) method. We conducted the item and ability parameter estimates and DIF analyses using the *mirt* package (Chalmers, 2012) in R software, while we assessed classification accuracy using the *cacIRT* package (Lathrop, 2015).

In this study, we calculated the standard errors corresponding to the item and ability parameter estimates obtained from different IRT models and then compared these models. Additionally, we used

the Spearman Rank-Order Correlation Coefficient to examine the relationship between item and ability parameter values. We employed the Rudner method based on IRT to calculate classification accuracy. In determining DIF/DBF, we used the SIBTEST method, which allows both individual items and testlets to be examined independently.

Testlet Response Theory

Testlet Response Theory (TRT) addresses local dependence on items that are part of testlets. If local dependence is not properly accounted for, the psychometric results of the test can be adversely affected. Over the past twenty years, various methods have been proposed to model testlet structures in order to capture local item dependence from different perspectives. Bradlow et al. (1999) and Wainer et al. (2000) extended traditional Item Response Theory (IRT) models by incorporating random effect parameters to explain the interaction between testlets and individuals. In this context, the 2PL-TRT model, which includes the random item effect parameter γ accounting for the local dependence levels among items within a testlet, consists of the following parameters: a (item discrimination), b (item difficulty) and γ (random testlet effect parameter)

$$P(\theta_i, \alpha_i, b_i) = \frac{exp\ (\alpha_i(\theta_j - b_i - \gamma_{jd(i)}))}{1 + \ exp\ (\alpha_i(\theta_j - b_i - \gamma_{jd(i)}))}$$

The testlet effect parameter ($\gamma_{jd(i)}$) is a parameter specific to both the individual and the testlet. When the local independence assumption is met, the value of this parameter is zero, i.e., $\gamma_{jd(i)} = 0$ for all individuals, and in this case, the TRT model transforms into the unidimensional IRT model. The variance of $\gamma_{jd(i)}$ is typically estimated for each item set and is used as an indicator of the degree of local dependence among the items within the testlet. The variances of the testlet effects vary across testlets. Additionally, when the guessing parameter (c_i) is included, the 3PL-TRT model becomes a special case of the 2PL-TRT model. However, since the 3PL-TRT model contains more parameters than other TRT models, the computational algorithms are more complex.

Li and others (2006) proposed a general two-parameter normal ogive testlet response theory model (2PNOTRT) from a multidimensional perspective. In this multidimensional model, each item response is dependent on both the primary dimension and the secondary testlet dimension. Both TRT models are constructed within the probit link function framework. Based on this, Zhan, Li, Wang, and Bian (2015) introduced the concept of within-item multidimensional testlet effect. Lu, Zhang, Zhang, Xu, and Tao (2021) proposed a new testlet discrimination parameter based on the logit link function for dichotomously scored items. This parameter has been applied in large-scale language assessments (Eckes, 2014; Rijmen, 2010; Zhang, 2010), hierarchical data analyses (Jiao, Kamata, Wang, & Jin, 2012), and cognitive diagnostic assessments (Zhan et al., 2018) in fields such as education and psychology.

One of the most commonly used estimation methods for TRT models is the marginal maximum likelihood method through the Expectation-Maximization (EM; Dempster, Laird, & Rubin, 1977) algorithm (Bock & Aitkin, 1981; Glas, Wainer & Bradlow, 2000; Mislevy, 1986; Wang & Wilson, 2005). Ability parameters and testlet effects are viewed as unobserved latent variables, and then the marginalization of the full data likelihood (responses and unobserved data) can be computed based on the unobserved data. However, the marginal maximum likelihood estimation of TRT models is hindered by the fact that the calculations often involve analytically challenging high-dimensional integrals, making it difficult to obtain the maximum likelihood estimates of the parameters. More specifically, when integrals over the latent variable distributions are evaluated using Gauss quadrature (Bock & Aitkin, 1981), the number of relevant computations increases exponentially with the number of latent variable dimensions. Although the number of quadrature points per dimension can be reduced by using adaptive Gauss quadrature (Pinheiro & Bates, 1995), the total number of points still grows exponentially with the number of dimensions. Naturally, all of this TRT analysis process encompasses a more complex and time-consuming procedure.

Classification Accuracy

Classification accuracy refers to the degree to which decisions made based on test scores align with decisions that would be made if the scores were free of measurement error (Hambleton & Novick, 1973). Since all measurements in education and psychology involve some degree of measurement error, it is necessary to determine classification accuracy. A misclassification of a test participant indicates a classification error. Classification errors occur when a test participant is assigned to a higher ability category than their true proficiency level or to a lower category than their true ability. Classification accuracy is crucial in high-stakes assessments, as it can have significant implications for students' futures. Additionally, it provides valuable insights into the strengths and weaknesses of an assessment, helping educators and policymakers make data-driven, informed decisions. Classification accuracy is critical for guiding instructional decisions, evaluating program effectiveness, and ensuring that students receive the necessary support to succeed (Cizek & Bunch, 2007).

Both criterion-referenced and norm-referenced assessments involve classification. An example of a two-category classification could be "pass" and "fail," while multiple classification categories might include levels such as "basic," "sufficient" and "advanced." An example of classification used in criterion-referenced tests is the language score categorization (A, B, C, D) in the Foreign Language Exam (YDS) conducted by ÖSYM. In norm-referenced assessments, an example of classification would be the Higher Education Institutions Exam (YKS), which evaluates student performance by comparing it to the performance of other candidates. In the calculation of classification accuracy, the first methods developed were based on two applications. However, due to the challenges associated with two applications, there has been a growing effort to develop classification accuracy indices based on a single application. These methods can generally be classified into two categories: methods based on Classical Test Theory (CTT) (Hanson & Brennan, 1990; Huynh, 1976; Lee & Song, 2004; Livingston & Lewis, 1995; Subkoviak, 1976) and methods based on Item Response Theory (IRT) (Lee, 2010; Rudner, 2001, 2005). Within the IRT framework, the point estimate of ability can be treated as the latent true score. Rudner (2001, 2005) proposed a method for evaluating decision accuracy by calculating the expected probability of classifications. He developed indices based on the ability (θ) scale. In this method, classification accuracy is obtained by calculating the expected likelihood of classifications. For example, let the cutoff score be θ_c, the true ability of candidate A be θn, and the true ability of candidate B be θ_m. Since $\theta_m > \theta_c > \theta_n$, candidate A should be classified as "failed" in all predictions, while candidate B should be classified as "passed." However, due to the error in ability estimation, a conditional distribution accompanies each true ability (θ). This means that candidate A might be classified as "passed" or "successful" by chance. This occurs when the estimated ability of the candidate exceeds the cutoff score θ_c. In classification terminology, this chance refers to the probability of making a false positive error, where a failure or non-expert individual is incorrectly classified as successful or an expert.

In this study, we classified students' possible scores into two categories based on proficiency level cutoff scores: below and above average. Subsequently, we performed a multi-category classification for four proficiency levels (lower, middle, upper, and advanced). For this comparable classification among different IRT models, we first determined cutoff scores and the corresponding ability levels for these cutoff scores.

Differential Item and Differential Item Functioning

Differential Item Functioning (DIF) occurs when individuals from different groups with the same ability level have different probabilities of answering an item correctly (Clauser & Mazor, 1998). Understanding and addressing DIF is crucial for ensuring fairness and validity in assessment practices. Identifying DIF is a critical step in evaluating fairness. Among the methods for detecting DIF are the Mantel-Haenszel method (Holland & Thayer, 1988), standardization method (Dorans & Kulick, 1986), logistic regression method (Swaminathan & Rogers, 1990), SIBTEST method (Shealy & Stout, 1993), Lord's (1980) chi-square test (Wright & Stone, 1979), likelihood ratio test (Thissen, Steinberg, & Wainer, 1988; Wang & Yeh, 2003), and the Multiple Indicators Multiple Causes (MIMIC) model (Finch, 2005;

Oort, 1998). In addition, there have been studies in which DIF was detected with methods based on Cognitive Diagnostic Models in recent years (Eren, Gündüz, & Tan, 2023; Ma, Terzi, & de la Torre, 2021).

Differential Item Functioning (DIF) has also been defined within a multidimensional framework, based on the fact that it may arise due to certain characteristics of test items that are unrelated to the construct being measured. This framework assumes that all tests are, to some extent, multidimensional. In a test, there may be a primary dimension related to the construct being measured, as well as other dimensions that generate variance unrelated to the construct. For example, in a problem-based math test, in addition to primary dimensions reflecting mathematical ability, there may be secondary dimensions such as reading comprehension or verbal ability. These other dimensions are typically correlated with the primary dimension. From this perspective, DIF is thought to arise from dimensions that are different from the primary construct of the test. Ackerman (1992) has discussed the foundation of the multidimensional framework in depth. Shealy and Stout (1993) developed a DIF statistic called SIBTEST within this framework. SIBTEST allows for the examination of multiple dimensions as sources of DIF. Since this method involves a type of factor analysis, it also allows for the examination of item groups rather than individual items. Because the method permits grouping items for DIF analysis, it is particularly useful for making more robust generalizations about the sources of DIF (Gierl, Bisanz, Bisanz, & Boughton, 2003; Mendes-Barnett & Ercikan, 2010).

Unlike traditional DIF detection methods such as logistic regression and Mantel-Haenszel, SIBTEST has the advantage of examining DIF at both the individual item level and the testlet/bundle level. SIBTEST, which stands for Simultaneous Item Bias Test, is a regression-based method that evaluates the degree of different functioning of an item or item group across two or more subgroups of the test after controlling for the measured underlying ability. The method primarily involves predicting the relationship between the total test scores of the focus and reference groups and then testing for deviations from this general relationship for the specific item or item group. In this study, we used the SIBTEST method, which can work at both the item and item group level, to identify DIF and DBF. To interpret the β index for DIF detection based on the SIBTEST method, Roussos and Stout (1996) proposed a classification. A β index value smaller than 0.059 indicates no DIF, a β index smaller than 0.088 indicates moderate DIF, and a β index equal to or greater than 0.088 indicates high-level DIF. While a classification for effect size is possible for DIF, no such classification exists for DBF.

Results

In response to the first research question, "What is the level of local dependence among the three testlets in the mathematics subtest?", the analyses showed that the local dependence levels of the testlets were moderate ($\sigma > 0.5$). Specifically, the two-item testlet 1, the four-item testlet 2, and the six-item testlet 3 all exhibited moderate levels of local dependence. The local item dependence caused by the testlet is referred to as the "testlet effect" (Wainer & Kiely, 1987). As variance increases, the effect created by testlets also increases (Wainer & Wang, 2000). The variance values are interpreted as follows: 0 indicates "no testlet effect," 0.5 indicates "moderate testlet effect," and 1 indicates "large testlet effect" (Wang, Bradlow ve Wainer, 2002; Wang & Wilson, 2005). In this study, using the 2PL-TRT model, the testlet effects for the three testlets were found to be 0.575, 0.505 and 0.612 respectively. Although the highest local dependence appeared in the third testlet, overall, the local dependencies in all three testlets were moderate. As a result, none of the testlets exhibited a significant testlet effect.

For the second research question, "How is the relationship between the item and ability parameters obtained from the 2PL-IRT and 2PL-TRT models in the mathematics subtest consisting of testlets?", the item and ability parameters were first calculated using the 2PL-IRT and 2PL-TRT models. The item parameters and their standard error values are presented in Table 1. In Table 1, α represents the slope parameter, and δ represents the intercept parameter.

Table 1. Item Parameters and Standard Error Values

		2PL-IRT				2PL-TRT			
Testlet	**Items**	**α**	**α_{se}**	**δ**	**δ_{se}**	**α**	**α_{se}**	**δ**	**δ_{se}**
Testlet I	ME72041A	4.06	0.77	-0.40	0.06	4.44	1.63	-0.62	0.39
	ME72041B	4.40	0.87	-1.06	0.06	5.91	2.89	-1.84	0.93
Testlet II	ME72081A	1.03	0.16	1.24	0.18	1.57	0.38	1.63	0.31
	ME72081B	0.68	0.13	0.52	0.19	0.98	0.19	0.61	0.13
	ME72081C	0.73	0.13	-0.48	0.17	0.75	0.18	-0.49	0.11
	ME72081D	0.72	0.14	1.12	0.29	0.79	0.20	1.20	0.14
Testlet III	ME72140A	1.76	0.26	1.89	0.12	1.93	0.35	2.33	0.31
	ME72140B	1.84	0.34	3.21	0.19	2.13	0.47	4.03	0.61
	ME72140C	1.64	0.25	2.20	0.15	1.98	0.37	2.85	0.39
	ME72140D	1.02	0.21	2.35	0.38	1.32	0.25	2.83	0.29
	ME72140E	0.75	0.14	0.91	0.23	0.68	0.14	0.94	0.12
	ME72140F	1.48	0.25	2.48	0.20	1.44	0.31	2.74	0.29

The item slope parameter (α) is interpreted as the item discrimination parameter. Higher values indicate that the item is more discriminative (Baker, 2001). Upon examining the parameter values for both models, we concluded that for the IRT model the discrimination parameters ranged from 0.68 to 4.40 with standard errors varying between 0.13 and 0.87. For the TRT model, the discrimination values ranged from 0.68 to 5.91, with standard errors varying from 0.14 to 2.89. The item intercept parameter (δ) is interpreted as item difficulty, which is the inverse of the item difficulty parameter. A higher value indicates that the item is easier (Reckase, 2009). For the intercept parameter, in the IRT model, the values ranged from -1.06 to 3.21, with standard errors varying from 0.06 to 0.38. In the TRT model, these values ranged from -1.84 to 4.03, with standard errors ranging from 0.11 to 0.93. Table 2 below presents the mean, minimum and maximum values for these item parameter estimates.

Table 2. Descriptive Statistics of Estimated Item Parameter Values

Parameters	**2PL-IRT**			**2PL-TRT**		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
Slope (α)	1.68	0.68	4.40	1.99	0.67	5.91
α_{se}	0.30	0.13	0.87	0.61	0.14	2.89
Intercept (δ)	1.17	-1.06	3.21	1.35	-1.84	4.03
δ_{se}	0.19	0.06	0.38	0.34	0.11	0.93

As is seen in Table 2 presenting the slope parameters and standard errors for both models, the 2PL-IRT model shows lower values. This difference in the parameter values is accompanied by smaller differences in the standard errors. Similarly, examining the item intercept parameters results in the same pattern: the 2PL-IRT model results in lower parameter and standard error values. Additionally, the standard errors for the item intercept parameters are found to be lower than those for the discrimination parameters. The ability parameters and their corresponding standard errors are presented in Table 3.

Table 3. Descriptive Statistics Regarding the Ability Parameters and Standard Errors

Model	**Ability Parameter Value (θ)**			**Standard Error (se)**		
	Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
2PL-IRT	0.00	-2.33	1.31	0.45	0.34	0.63
2PL-TRT	0.00	-1.96	1.25	0.59	0.45	0.70

As is seen in Table 3, the ability parameters and their corresponding standard error values are similar for both models. However, the range of the parameter values for the 2PL-IRT model is wider,

while the standard errors are lower compared to the 2PL-TRT model. The relationship between the ability parameters obtained from both models is illustrated in the scatter plot shown in Figure 1.

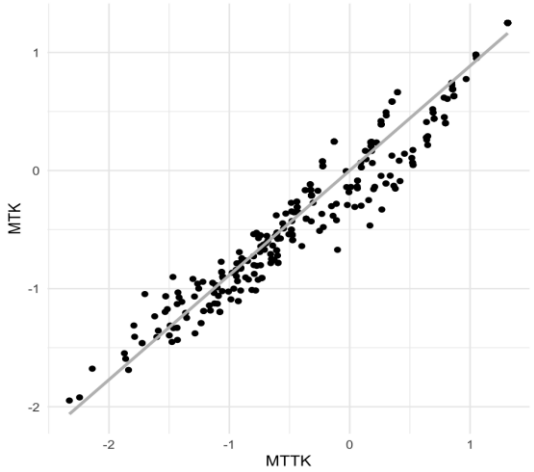

Figure 1. Scatter Plot Graph Regarding Ability Parameter Estimation

Figure 1 presents the scatter plot of ability estimates obtained from the IRT and TRT models. As is seen in the scatter plot, the ability parameter estimates from both models are very similar. While there is a high correlation between the slope and intercept parameters, a graphical comparison of these values is not included since the estimated values for both models do not lie on the same scale. The correlation values between the parameters obtained from the IRT and TRT models are provided in Table 4.

Table 4. Correlation Values Regarding the Parameters Estimated via IRT and TRT Models

Testlet	Number of Item	α	δ	θ
Testlet I	2			
Testlet II	4	0.983	0.997	0.976
Testlet III	6			

As is seen in Table 4, all the correlation values obtained are quite high ($r > .95$). Specifically, the correlation values are as follows: 0.983 for the slope parameter (α), 0.997 for the intercept parameter (δ), and 0.976 for the ability estimates (θ). These high correlation values suggest a strong similarity between the parameters obtained from the IRT and TRT models. We first determined the cutoff scores for the third research question, "How is the classification accuracy obtained from the 2PL-IRT and 2PL-TRT models in the mathematics subtest consisting of testlets?" The cutoff score calculation was based on the possible values for the students' performance in the mathematics test and the four different proficiency levels (low, medium, high, and advanced) defined in the TIMSS application to represent the students' behavioural indicators of success. To compare the two IRT models, we determined the proficiency levels corresponding to the determined cutoff score. Those below the determined proficiency level were classified as "below the level," and those above were classified as "above the level." The classification was made based on four different proficiency levels, which are "below/above low," "below/above medium," "below/above high," and "below/above advanced" as per the TIMSS. The findings related to classification accuracy are presented in Table 5.

Table 5. Values regarding the Two-Category Classification Accuracy

Model	2PL-IRT				2PL-TRT			
Level	**Low**	**Medium**	**High**	**Advanced**	**Low**	**Medium**	**High**	**Advanced**
Cutoff score (θ)	-0.74	-0.32	0.40	1.31	-0.64	-0.21	0.39	1.25
Classification Accuracy	0.94	0.90	0.93	0.95	0.90	0.93	0.94	0.91
Classification Consistency	0.91	0.86	0.91	0.94	0.86	0.90	0.93	0.86

As is seen in Table 5, the cutoff scores for the 2PL-IRT and 2PL-TRT models increase as expected for each level. For the 2PL-IRT model, the cutoff scores are -0.74 for the low level, -0.32 for the medium level, 0.40 for the high level, and 1.31 for the advanced level. For the 2PL-TRT model, the cutoff scores are -0.64 for the low level, -0.21 for the medium level, 0.39 for the high level, and 1.25 for the advanced level. In both models, the cutoff scores increase across levels as expected. This indicates that as students progress from the low to the advanced levels, their knowledge increases and the application of this knowledge becomes more complex. For instance, a student at the low level is expected to have basic knowledge of mathematics, while a student at the advanced level should be able to solve more complex problems and provide justifications for their solutions. Therefore, students at the advanced level are expected to possess higher abilities. Regarding classification accuracy, slightly higher accuracy and consistency were found for the 2PL-TRT model in the two-category middle and higher levels. However, for the lower and advanced levels, the 2PL-IRT model showed higher accuracy. Overall, the classification accuracy for both models is high and similar, with small differences observed across some levels.

Table 6. Values Regarding Multi-Category Classification Accuracy

Model	**2PL-IRT**	**2PL-TRT**
Classification Accuracy	0.74	0.73
Classification Consistency	0.68	0.64

As is seen in Table 6, for multi-category classification, when the four levels (lower, middle, higher and advanced) are considered together, the classification accuracy and consistency of the 2PL-IRT model are slightly higher than those of the 2PL-TRT model.

To answer the fourth research question, "Are there any items containing gender based DIF/DBF in the mathematics subtest consisting of testlets?", we used the SIBTEST function from the *mirt* package in the R software. The results for each item and item group are presented in Table 7.

Table 7. Results of Differential Item and Differential Bundle Functioning

		DIF		**DBF**	
	Items	**β**	**p**	**β**	**p**
Testlet I	ME72041A	-0.012	0.797	-0.164	0.056
	ME72041B	-0.065	0.136		
Testlet II	ME72081A	-0.062	0.171	-0.037	0.741
	ME72081B	-0.009	0.855		
	ME72081C	0.064	0.175		
	ME72081D	0.033	0.468		
Testlet III	ME72140A	0.016	0.719	0.137	0.38
	ME72140B	-0.007	0.872		
	ME72140C	0.056	0.269		
	ME72140D	-0.052	0.156		
	ME72140E	-0.055	0.244		
	ME72140F	0.074	0.107		

As is seen in Table 7, according to the SIBTEST method, none of the items in the testlets exhibited DIF ($p > .05$), and no item group displayed DBF either.

Discussion, Conclusion and Suggestions

In this study, we modelled the local dependence of three testlets that were both in Booklets 13 and 14 in the eTIMSS 2019 application for the Türkiye sample using the TRT method (Wainer et al., 2007). Grouping related items in testlets can reduce cognitive load during the test, which may be particularly beneficial for younger age groups (Yen, 1993). Testlets can also provide more detailed information about students who struggle in specific areas, potentially enhancing targeted educational opportunities. The primary point to examine in such items, which are increasingly used both nationally and internationally, is the degree of local independence that reveals the relationship between the items. Therefore, the first research question pertains to the degree of local dependence that may exist within each testlet. During the analysis, we calculated the effect sizes for the testlets as 0.575, 0.505 and 0.612 respectively. These values fall within the critical range identified in the literature as indicators of moderate local dependence (Li et al., 2006; Wang & Wilson, 2005). Murphy, Dodd, and Vaughn (2010) found that IRT and TRT models showed similar performance in testlet-based computer-adaptive tests with low to moderate testlet effects. This study suggests that traditional IRT methods can be used effectively in situations where the testlet effect is moderate and that these methods may serve as a more practical alternative in terms of time and effort.

The second research question involved a comparative examination of the parameter estimates (slope, intercept and ability parameters) obtained from the 2PL-IRT and 2PL-TRT models. The correlation between the ability parameter estimates based on the TRT model and the IRT model was found to be very high, with a value of 0.976. However, examining the standard errors of the ability estimates showed that the IRT model had lower error values. When local independence is violated in IRT models, ability estimates are typically obtained with lower standard errors (Chang & Wang, 2010; Eckes, 2014; Koziol, 2016; Wainer & Wang, 2000). Additionally, we concluded that the relationship between the slope and intercept parameters was also quite high ($r > 0.98$). The moderate level of local dependence between the items within the testlets may have contributed to the high alignment between the parameter estimates derived from the models. In conclusion, a comparative examination of the analyses performed using the IRT and TRT models allows for a more effective design of testlets, especially in K-12 assessments.

In the third research question, we examined classification accuracy results using the IRT and TRT models. In this study, we performed two-category classifications, such as “below advanced level” and “above advanced level” for each level. For classifications requiring four levels—low, medium, high and advanced—we defined cutoff scores for each level. The cutoff scores for the two-category classification increased as expected in each model with the increase in the level. While cutoff scores in studies examining classification accuracy with different IRT models have been found to be quite similar (Lee, 2010; Zhang, 2010), in this study, the cutoff scores of both models differ from each other, except at the high level. Overall, we found out that the classification accuracy and consistency obtained from both two- and multi-category classifications were very high, and both models produced highly consistent results. For two-category classifications, the analysis based on the cutoff scores for medium and high levels showed better classification accuracy with the 2PL-TRT model. However, multi-category classification accuracy and consistency were lower than the two-category classification values. This is because the number of defined levels is an important criterion in calculating classification accuracy and consistency (Lathrop & Cheng, 2014). The current study findings indicate that the TRT model performs better than or equivalently to other approaches in terms of classification accuracy, especially when strong item set effects are present in the data (Keller et al., 2003; Zhang, 2010). However, in Koziol’s (2016) study, although IRT and TRT models yielded similar performances under small testlet effects, the classification accuracy percentages were found to be lower under large testlet effects. Therefore, in this study, similar and high classification accuracy percentages may have been obtained due to the absence of significant local dependence within the testlets.

In the fourth research question, we examined whether testlets demonstrate Differential Item Functioning (DIF) based on gender both at the item level and the testlet level using the SIBTEST method. The reason for selecting this method is that IRT-based approaches are suggested to be more robust than CTT-based approaches (Hambleton & Swaminathan, 2013). However, IRT-based approaches require the assumption of local item independence, which is increasingly difficult to meet as modern test design increasingly shifts towards the use of testlets (Ferne & Rupp, 2007; Wainer & Lukhele, 1997). In such cases, one of the applicable methods, the SIBTEST method, was used to apply DBF analysis. This study aimed to compare the results when testlets are considered both as independent items and as item groups, which is why the SIBTEST method was used for this comparison. When items were considered independently and as item groups, no items or testlets demonstrating DIF or DBF were found. This may be due to the presence of low to moderate item set effects within the testlets.

In conclusion, the correlation of parameter estimates, classification accuracy and DIF results in this study were found to be quite similar in both standard IRT and testlet-based IRT models. However, the IRT models yielded lower standard errors and higher classification accuracy percentages. Therefore, it is essential to continue the analysis by determining whether the conceptual advantages of testlets outweigh the statistical disadvantages. In general, if the performance decline is negligible, the advantages of these tests may outweigh their disadvantages. The literature includes several studies comparing parameter estimation in one-dimensional IRT and TRT models (Bradlow et al., 1999; Glas et al., 2000; Wainer et al., 2000; Wainer & Wang, 2000). However, there is a lack of studies investigating classification accuracy and DIF between these models. Therefore, future research should focus on methods for determining DIF in testlets and classification accuracy when high testlet effects are present. Particularly, studies can explore classification accuracy for situations with high testlet effects and items exhibiting DIF. Moreover, the observed similarities in classification accuracy between IRT and TRT models may be limited to the specific cutoff scores used. As the cutoff score approaches the extreme end of the latent ability distribution, classification accuracy may decrease. Therefore, studies can also be conducted to assess the impact of the cutoff score on the relative performance of the models. In K-12 classroom assessments, determining students' skill levels accurately is critical for teachers to develop individualized teaching strategies. The use of testlets can offer a more detailed assessment of various skill levels in the classroom. For instance, classifying a student as "below advanced" or "above advanced" can help the teacher identify which areas require additional support. Additionally, the use of multi-category classifications in K-12 classroom assessments can allow for a more detailed tracking of student progress, although it may require a more careful analysis in terms of classification accuracy.

Testlets at the K-12 level offer significant advantages, such as increasing measurement accuracy, shortening test duration, ensuring uniqueness, aligning with learning objectives and enhancing reliability. These benefits are supported by various studies that highlight the potential of testlets to both support learning and assessment. While taking advantage of these benefits, it is essential to consider their impact on psychometric analyses. Therefore, it is believed that this study will serve as a guide for future research on similar topics.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

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.

Baker, F. B. (2001). *The basics of item response theory.* Retrieved from https://files.eric.ed.gov/fulltext/ED458219.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison- Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. doi:10.1007/BF02293801

Bradlow, E. T., Wainer, H., & Wang, X. (1999). Bayesian random effects model for testlets. *ETS Research Report Series, 1998*(1). doi:10.1002/j.2333-8504.1998.tb01752.x

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6). doi:10.18637/jss.v048.i06

Chang, Y., & Wang, J. (2010). *Examining testlet effects on the PIRLS 2006 assessment.* Paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome Measurement, 3*(1), 1-20.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage.

DeMars, C. (2010). *Item response theory.* Oxford: Oxford University Press.

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. doi:10.1111/j.1745-3984.2006.00010.x

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement, 23*(4), 355-368.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61. doi: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. doi:10.1080/08957347.2014.1002919

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. Washington: American Psychological Association. doi:10.1037/12074-000

Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 76-94. doi:10.21031/epod.1218144

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113-148. doi:10.1080/15434300701375923

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436. doi:10.1007/BF02295430

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, *40*(4), 281-306

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271-287). Kluwer-Nijhoff.

Hambleton, R., & Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159-170.

Hambleton, R. K., & Rogers, H. (1989). Solving criterion-referenced measurement problems with item response models. *International Journal of Educational Research, 13*(2), 145-160. doi:10.1016/0883-0355(89)90003-7

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications.* Berlin: Springer Science & Business Media.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*(4), 345-359.

Ho, T.-H., & Dodd, B. G. (2012). Item selection and ability estimation procedures for a mixed-format adaptive test. *Applied Measurement in Education, 25*(4), 305-326. doi:10.1080/08957347.2012.714686

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Mahwah, NJ: Lawrence Erlbaum Associates.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 253-264.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82-100. doi:10.1111/j.1745-3984.2011.00161.x

Karasar, N. (2016). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayıncılık.

Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education, 16*(3), 207-222. doi:10.1207/S15324818AME1603_3

Koziol, N. A. (2016). Parameter recovery and classification accuracy under conditions of Testlet dependency: A comparison of the traditional 2PL, Testlet, and Bi-Factor models. *Applied Measurement in Education, 29*(3), 184-195. doi:10.1080/08957347.2016.1171767

Lathrop, Q. N. (2015). *cacIRT: Classification accuracy and consistency under item response theory.* Retrieved from https://CRAN.R-project.org/package=cacIRT

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. *Journal of Educational Measurement, 51*(3), 318-334. doi:10.1111/jedm.12048

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*(4), 357-372. doi:10.1177/01466210122032226

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653-686. doi:10.1207/s15327906mbr3904_4

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessment using item response theory. *Journal of Educational Measurement*, *47*(1), 1-17.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3-21. doi:10.1177/0146621605275414

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179-197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Lu, J., Zhang, J., Zhang, Z., Xu, B., & Tao, J. (2021). A novel and highly effective Bayesian sampling algorithm based on the auxiliary variables to estimate the testlet effect models. *Frontiers in Psychology,* 12. doi:10.3389/fpsyg.2021.509575

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement*, *45*(1), 37-53. doi:10.1177/0146621620965745

Mendes-Barnett, S., & Ercikan, K. (2010). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289-304.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453-477. doi:10.1177/0265532214527277

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-195. doi:10.1007/BF02293979

Murphy, D. L., Dodd, B. G., & Vaughn, B. K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, *34*(6), 424-437. doi:10.1177/0146621609349804

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*(2), 107-124.

Paek, I., & Fukuhara, H. (2015). An investigation of DIF mechanisms in the context of differential testlet effects. *British Journal of Mathematical and Statistical Psychology*, *68*(1), 142-157. doi:10.1111/bmsp.12039

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12-35. doi:10.1080/10618600.1995.10474663

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Rijmen, F. (2010). Formal relations and an empirical comparison among the Bi-Factor, the Testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372. doi:10.1111/j.1745-3984.2010.00118.x

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371. doi:10.1177/014662169602000404

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(1), 14. doi:10.7275/an9m-2035

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(1), 13. doi:10.7275/56a5-6b14

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DBF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

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

Soysal, S., & Yılmaz Koğar, E. (2022). Item parameter recovery via traditional 2PL, Testlet and Bi-factor models for Testlet-Based tests. *International Journal of Assessment Tools in Education, 9*(1), 254-266. doi:10.21449/ijate.948182

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265-275.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Tasdelen Teker, G., & Dogan, N. (2015). The effects of testlets on reliability and differential item functioning. *Educational Sciences: Theory and Practice*, *15*(4), 969-980. doi:10.12738/estp.2015.4.2577

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118-128.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260. doi:10.1111/j.1745-3984.1989.tb00331.x

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8*(2), 157-186. doi:10.1207/s15324818ame0802_4

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing, theory and practice* (pp. 245-270). Kluwer-Nijhoff.

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

Wainer, H., Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201. doi:10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14. doi:10.1111/j.1745-3984.1990.tb00730.x

Wainer, H., & Lukhele, R. (1997). Managing the influence of DIF from big items: The 1988 advanced placement history test as an example. *Applied Measurement in Education*, *10*(3), 201-215. doi:10.1207/s15324818ame1003_1

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220. doi:10.1111/j.1745-3984.2000.tb01083.x

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*(1), 109-128. doi:10.1177/014662160202600100

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296-318. doi:10.1177/0146621605276281

Wang, W. C., & Yeh, L. Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 479-498. doi:10.1177/0146621603259902

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492. doi:10.1177/014662168200600408

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213. doi:10.1111/j.1745-3984.1993.tb00423.x

Yılmaz Koğar, E. (2021). Comparison of testlet effect on parameter estimates using different item response theory models. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 254-266. doi:10.21031/epod.948227

Zhan, P., Li, X., Wang, W.-C., & Bian, Y. (2015). *The logistic testlet framework for within-item multidimensional testlet-effect.* Paper presented at the 2015 International Meeting of the Psychometric Society (IMPS), Beijing Normal University, Beijing, China.

Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology*, 9, 607. doi:10.3389/fpsyg.2018.00607

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), 119-140. doi:10.1177/0265532209347