Language Testing in Asia

RESEARCH

Open Access

Investigating the psychometric properties of the Qiyas for L1 Arabic language test using a Rasch measurement framework

Amjed A. Al-Owidha

Correspondence: alowidha@kfupm.edu.sa
Department of General Studies, King Fahd University of Petroleum and Minerals, P.O. Box 18, Dhahran 31261, Kingdom of Saudi Arabia

Abstract

Background: This study investigated the psychometric properties of the recently developed Qiyas for L1 Arabic language test using a Rasch measurement framework.

Methods: Responses from 271 examinees were analyzed in this study. The test is hypothesized to involve one dominant factor that assesses four skills: reading comprehension, rhetorical expression, structure, and writing accuracy.

Results: Fit statistics and reliability analysis, principal component analysis of Rasch residuals, and the results of differential item functioning supported the hypothesized structure of the Qiyas for L1 Arabic language test. However, the results of a person-item map analysis suggested that the content aspect validity of the Qiyas for L1 Arabic language test lacked representation to some extent.

Conclusion: The initial findings of the Rasch analysis indicated that the Qiyas for L1 Arabic language test maintains satisfactory psychometric properties. However, these findings should be interpreted with caution given the limitations of the sample population used. Continued investigation of the psychometric proprieties of the test is necessary to ensure its appropriate use as a tool of assessment for modern Arabic language.

Keywords: Language testing, Arabic language test, Rasch model, Differential item functioning

Introduction

The Qiyas for L1 Arabic language test is a standardized test recently developed by the National Center for Assessment (NCA) in Riyadh, Saudi Arabia. The lack of and need for a high-quality measurement tool that produces dependable estimations of the language skills of L1 speakers in the Arab world motivated the NCA to pioneer the development of such a test. The Qiyas for L1 Arabic language test is hypothesized to involve one dominant factor that assesses four skills: reading comprehension, rhetorical expression, structure, and writing accuracy. The test was developed primarily for the purpose of selection, where the intended population is people seeking jobs in schools, public relations, TV, radio stations, or other types of local or international communication that use modern Arabic as the main language. In addition, the test serves as a tool for selecting students for Arabic language programs in universities and/or programs that require higher language skills, like Islamic studies and law, and for diagnostic purposes, including placement of students at an appropriate level in university-level Arabic language programs and course waivers from some Arabic courses. Such language skills are expected to have been acquired by the targeted population throughout their education. The ultimate purpose of the Qiyas for L1 Arabic language test is to serve as a standardized tool that assesses modern standard Arabic language skills, not only in Saudi Arabia but throughout the Arab world. Because of its widespread use, it is critical that the NCA, as the developer and owner of this test, ensures that the Qiyas for L1 test maintains adequate psychometric properties.

One step of test construction and development is to check the quality of test items, to be sure they are functioning as expected. This process is called reliability and validity analysis. At this stage of development, test developers usually select and use stringent measurement models suited to the type of responses on the test. The purpose of this process is to ensure that the data under study are appropriately handled before validation takes place. The existing practice of the NCA in the field-testing stage involves the use of item response theory measurement models (IRT), in particular, a three-parameter logistic model that calibrates test items, generates item parameters, checks their appropriateness, and then utilizes the best item parameters to construct the test. This 3-IRT model requires a sample size of at least 1000 people to ensure sufficient and accurate stability in item parameter estimation (e.g., Lord 1980; Hutten 1981). Inaccuracy in item parameter estimation can affect the measurement invariance property of the IRT, which, in turn, would call into question test-score validation (de Jong and Stoyanova 1994). Such was the case for this specific Qiyas for L1 Arabic language test in its first field-testing implementation, which involved only 271 people; thus, a more robust and suitable IRT model is needed to validate this form. In this study, Rasch measurement was selected as the model of choice. One advantage of using this model over other IRT models is that it is usable and applicable with small sample sizes, while maintaining strong and restrictive assumptions. For instance, a sample size of between 25 and 50 subjects per response category is adequate to achieve stable and accurate item parameters when analyzing dichotomous data with the Rasch model (Linacre 1994). Rasch models have been used for validation purposes in the area of language testing since the early 1980s. For instance, De Jong used the Rasch model to assess the validity of a language test (De Jong 1983; McNamara and Knoch 2012). Nakamura (2007) also examined the psychometric properties of an in-house English placement test with the Rasch model. However, the NCA has not commonly used the Rasch model for test-validation purposes during the field-testing stage; to date, application of the model has been largely limited to in-house technical measurement reports, such as test equating and test bias. Accordingly, the purpose of this study was to illustrate the usability and applicability of the Rasch measurement framework during field testing. The objective of the study was to examine the psychometric properties of the field-testing version of the Qiyas for L1 Arabic language test using the Rasch model. Specifically, this study asked the following research question: Does the field-testing version of the Qiyas for L1 Arabic language test exhibit adequate psychometric properties according to the Rasch measurement framework?

To answer this question, quantitative analysis within the framework of Rasch measurement was conducted. The model is briefly introduced below, before the analysis is discussed.

The Rasch model

The Rasch model is an item-response model that provides a linear transformation of the ordinal raw scores to a linear logit scale (Boone, Staver, and Yale 2014). More specifically, Rasch measurement specifies the relationships between people and items on a test that measures one trait at a time, that is, the likelihood of a person's success will increase as the measurement of a trait increases. Conversely, the likelihood of failure increases when the trait is less measured. With the Rasch model, only the interaction between the person's position on the underlying ability being measured by a test and item difficulty are modeled. The model is expressed mathematically as follows (De Ayala 2009):

$$p\left(x_j = 1|\theta, \delta_\mathrm{j}\right) = \frac{e^{\left(\theta - \delta_j\right)}}{1 + \mathrm{e}^{\left(\theta - \delta_j\right)}}, \tag{1}$$

where $p(x_j = 1|\ \theta,\ \delta_j)$ is the probability of the response of 1, θ is the person location, δ_j is the item j's location, and e is a constant number whose value is 2.7183. In other words, Eq. 1 states that the probability of a person getting a correct response 1 on item j is a function of the distance between a person located on θ and the item located at δ.

Unlike other IRT models that focus on fitting the data given the constraints of the model, the Rasch model focuses on constructing the variable of interest (Andrich 1988; Wright and Masters 1982; Wright and Stone 1979). According to this perspective, the Rasch model represents the standard by which one can create a test for measuring the variable of interest; thus, the test data must meet this standard. The Rasch standard has an additive measurement form, where adding one more unit (defined by the logit value) adds the same extra amount regardless of how much was already there (Linacre 2012a, 2012b). However, the Rasch model requires that certain assumptions that must be met; not meeting those assumptions can compromise the usefulness of the model. These assumptions include unidimensionality, local independence, parallel item characteristic curves (ICCs), and measurement invariance. Unidimensionality assumes that the test data measure one ability or trait at a time (e.g., verbal ability). This assumption is never completely met; however, the presence of a dominant trait that influences the performance of the examinee on a test is necessary and can be met. Local independence is a related assumption that states that when a specified ability (e.g., verbal ability) influencing the performance of the examinee on a test has been partialled out, the responses of the examinee to any pair of items are statistically independent. The assumption of local independence can be met as long as the complete ability space has been accurately specified. It has been previously shown that if the assumption of unidimensionality is met, unidimensionality and local independence can be viewed as interchangeable concepts (Lord 1980; Lord and Novick 1968).

Another important assumption that is unique to and required by only the Rasch model is that test items that have ICCs must not intersect. Restrictions on item discrimination have made it difficult for some test data to fit expectations of the Rasch model unless discrimination indices are chosen to be equal (Birnbaum 1968). Stage (1996) also found it difficult to fit the Rasch model to some test data. Additionally, a study by Leeson and Fletcher (2003) has also shown evidence consistent with the findings of Birnbaum (1968) and Stage (1996). On the other hand, studies conducted by Wright and Panchapakesan (1969), Dinero and Haertel (1977), and Hambleton and Cook (1978) have indicated that the Rasch model is robust to heterogeneous item discrimination, that is, variation in item discrimination indices has little impact on the fit of the Rasch model. Measurement invariance is another property that is required by the Rasch model. This property assumes that a person's ability can be estimated independently of items on a particular test and that item indices can be estimated independently of the specific sample of people taking the test. Another feature of the Rasch model that distinguishes it from other unidimensional IRT models is that the total score provides sufficient statistics to model the data of interest. With the Rasch model, the total score contains all the information needed to estimate θ_i (Wright 1984; de Ayala 2009). The total score is sufficient for Rasch model estimation, but only if fit statistics conform to model expectations. Moreover, researchers can establish construct-related validity evidence if the data meet the requirements of the Rasch model. The Rasch model provides researchers with quality-control fit statistics and a principal component analysis of residuals (PCAR) that can be used to evaluate the internal structure of test scores. The outcome informs the researcher as to whether item responses follow a logical pattern. Items that do not fit a logical pattern are likely to be harmful to the construct under study and should be modified or deleted. Conversely, items that fit the logical pattern are likely to enhance the construct and should be retained.

Furthermore, differential item functioning (DIF) analysis within the Rasch measurement framework can be used to help detect bias or irrelevant factors. In the recently released *Standards for Educational and Psychological Testing* (American Educational Research Association 2014), the issue of "equivalence of the construct being assessed" was considered with respect to the importance of designing tests that produce scores that reflect only the ability that is being measured by the test, regardless of the identity of the subgroup that took it. Hence, the authors recommend the use of DIF analysis as an approach to investigate item bias. Differential item functioning occurs when examinees with the same level of ability from different subgroups (e.g., gender, language background, etc.) differ in their likelihood of answering an item correctly. Several DIF methods can detect item bias, including the Mantel-Haenszel test (Holland and Thayer 1988), logistic regression (Zumbo 1999), and IRT-based DIF methods (Thissen 1991; Thissen et al 1993; Wright and Stone 1979; Wright, Mead, and Draba 1976). The IRT-based DIF methods are relevant to this study, specifically Rasch-based DIF methods. According to Smith (2004), two approaches exist within the framework of the Rasch model: the separate *t* test approach and the between-group fit approach. The former is based on two separate *t* test calibrations of two or more subgroups of interest. The latter is based on a single calibration that involves one or more subgroups of interest. In this study, the former was used, and so it is briefly discussed below.

The *t*-statistic approach

The *t*-statistic approach is well documented in the Rasch model literature (Smith 2004). It is based on the differences between two separate calibrations of the same item of two or more subgroups of interest. Mathematically, the *t*-statistic approach is expressed as follows (Smith 2004):

$$t = \frac{d_{i1} - d_{i2}}{\left(s_{i1}^2 + s_{i2}^2\right)^{1/2}}, \tag{2}$$

where d_{i1} is the difficulty of item I based on the first subpopulation, d_{i2} is the difficulty of item i based on the second subpopulation, s_{i1} is the standard error of the estimate for d_{i1}, and s_{i2} is the standard error for d_{i2}. This method works only with pairwise comparisons; if there are more than two subpopulations included in the analysis, multiple comparisons must be made. A drawback of multiple comparisons based on one variable is that the type I error rate and the ability of this statistic to detect bias can be affected; however, this issue is of little concern here given that only pairwise comparisons were conducted. Furthermore, an essential requirement of this method is that any item that does not fit the Rasch model should be excluded. An item with poor fit can violate the fundamental assumption of the Rasch standard that ICCs do not cross, and the lower asymptote of ICCs must be zero (Smith 2004).

Methods

Participants

Data from the field-testing version of Qiyas for L1 Arabic language test used in this study were obtained from the NCA database. This test was administered by the NCA in January 2017 in Riyadh, Saudi Arabia. The test included binary-scored responses where 1 = correct and 0 = incorrect for 271 examinees of both genders (male and female).

Measurement

The Qiyas for L1 Arabic language test is a newly developed standardized test designed to measure the extent of Arabic language skills in L1 speakers and classify them at the appropriate level. More specifically, it aims to measure Arabic language skills in L1 speakers starting at approximately 11th grade and continuing through to university graduates for educational and professional purposes. Table 1 defines the skills measured by the Qiyas for L1 Arabic test.

Table 1 shows that the Qiyas for L1 Arabic language test is composed of the following skills:

Reading comprehension: Reading passages are of various lengths, classified as short (40–50 words), medium (51–200 words), and long (more than 201 words). Items in this skill

Table 1 Test skills and number of items for Qiyas L1 Arabic test

Skills	Explanations	No. of items
Reading comprehension	Questions related to understanding, analysis, and synthesis	14
Rhetorical expression	Questions related to situational, stylistic, and figurative language use	6
Structures	Questions related to structural correctness	16
Writing accuracy	Questions related to writing correctness	14
Total		50

target higher-order reading comprehension abilities, including inference and understanding of subtle meaning; text analysis; synthesis and abridgement; and summary.
Rhetorical expression: Items in this skill target situational, stylistic, and figurative language use. Punning, hyperbole, and metaphor are parts of speech that are used to measure pragmatic uses of language. Speech ornaments are common in the Arabic language. They are used to measure the extent of stylistic appreciation of language.
Structure: Items in this skill target all forms of structural correctness, including correct syntactic constructions like predication, attribution, coordination, conjunction, and adjectival and adverbial constructions. Structural correctness is highly linked to actual language use, rather than the perceptual understanding of grammar.
Writing accuracy: Items in this skill target correct written communication, which is intimately related to writing technique and spelling.

In addition, the Qiyas for L1 Arabic language test was developed to estimate four levels of language attainment, as follows:

1. Below medium: Examinees at this level are able to write texts that show basic language structure but lack expository narrative writing skills. Examples of the latter skills include the development of ideas, cohesion, and organization. Basic spelling mistakes are rampant. Examinees at this level are able to determine the main messages of reading and listening passages. They are also able to recognize explicit and direct ideas and some common words. Examinees at this level are not expected to distinguish rhetorical expressions and/or comprehend their significance.
2. Medium: Examinees at this level are able to write texts that show some detailed language structure beyond basic. They may also demonstrate some expository narrative writing skills related to idea development, cohesion, and organization. Some fine spelling mistakes may be present. Examinees in this level are able to determine the main messages of reading and listening passages. They are also able to recognize explicit ideas and some inexplicit ideas, as well as the textual meanings of common words. Examinees at this level are able to distinguish easier rhetorical expressions and/or comprehend their significance.
3. High-medium: Examinees at this level are able to write texts that show correct basic, and some more detailed, language structure. They may also show greater expository narrative writing skills related to idea development, cohesion, and organization. Occasional spelling mistakes may be present. Examinees in this level are able to determine the main messages of reading and listening passages. They are also able to recognize explicit ideas and some inexplicit ideas, as well as the textual meanings of common words. Examinees at this level are able to distinguish a significant number of rhetorical expressions and comprehend their uses and/or significance.
4. High: Examinees at this level are able to write texts that demonstrate correct basic and detailed language structure. They also show optimal expository narrative writing skills related to idea development, cohesion, and organization. Examinees at this level utilize creative, persuasive, and polemic writing techniques. No spelling mistakes are present. Examinees in this level are able to determine inferred messages in reading and listening passages. They are also able to recognize explicit and implicit ideas and the textual meanings of some common words. Examinees at

this level are able to distinguish almost all rhetorical expressions and comprehend their uses and/or significance.

Data analysis

To evaluate the psychometric properties of the Qiyas for L1 Arabic test, Winsteps® version 3.75.1 (Linacre 2012a, 2012b) was used. Two stages of Rasch analysis were carried out in this particular study. The first stage of analysis compared the suitability of Qiyas L1 test items against the Rasch standard. Various Rasch statistical indices were examined (e.g., fit statistics, person and item reliability indices, the person-item map, and point-measure correlation indices). The second stage of analysis investigated the structural aspect of the Qiyas for L1 Arabic language test using PCAR and DIF analysis, to examine whether irrelevant factors might be interfering with the main construct under study.

Results

Stage 1

Fit statistics and reliability analysis

Qiyas for L1 Arabic language test data were fitted to the Rasch model. Table 2 shows that the overall mean infit and outfit were 1.00 and 1.02, respectively, with a mean standardized infit and outfit of 0.0 and 0.1, respectively. This result suggests that overall, the Qiyas for L1 Arabic language test data fit the Rasch model reasonably well. The extra 0.02 in the overall mean outfit represented a small amount of unmodeled noise in the Qiyas for L1 Arabic language test data. Table 2 also showed that the reliability of the Qiyas for L1 Arabic language test was 0.86, supporting the notion that the ordering of persons along the construct is replicable given similar items measuring the same trait, that is, the Qiyas for L1 Arabic language

Table 2 Overall person- and item-fit statistics and reliability analysis of the Qiyas for L1 Arabic language test

Total score	Count	Measure	Model error	Mosq	Infit ZSTD	Mnsq	Outfit ZSTD	
Summary of 271 measured persons								
Mean	28.3	50.0	0.33	0.33	1.00	0.0	1.02	0.1
SD	8.7	0.0	0.92	0.03	0.13	0.9	0.26	1.1
Max.	46.0	50.0	2.78	0.54	1.43	2.5	2.29	3.4
Min.	9.0	50.0	− 1.75	0.31	0.69	− 2.7	0.51	− 2.6
Real RMSE = 0.34		True SD = 0.85		Separation = 2.48		Person reliability = 0.86		
Model RMSE = 0.33		True SD = 85		Separation = 2.55		Person reliability = 0.87		
SE of person mean = 0.06								
Summary of 50 measured items								
Mean	153.1	271.0	0.00	0.14	0.99	0.0	1.02	0.1
SD	46.5	0.0	0.90	0.01	0.10	1.8	0.20	2.0
Max.	229.0	271.0	2.17	0.18	1.26	4.4	1.62	4.5
Min.	46.0	271.0	− 1.63	0.13	0.78	− 4.1	0.72	− 3.7
Real RMSE = 0.15		True SD = 0.89		Separation = 6.10		Item reliability = 0.97		
Model RMSE = 0.14		True SD = 0.89		Separation = 6.22		Item reliability = 0.97		
SE of item mean = 0.13								

test had adequate test score reliability. The separation index per person was 2.48. This index measures the spread of examinee scores along a logit interval scale. Separation greater than 1 suggests that the data are sufficiently broad in terms of position (Frantom and Green 2002).

Table 2 also demonstrates that the item reliability index was 0.97. This indicates that the Qiyas for L1 Arabic language test items were reasonably well dispersed along the interval logit scale, suggesting an adequate breadth of position on the linear continuum from persons who were less skillful to more skillful in Arabic language proficiency. The person-item map, as depicted in Fig. 1, provides a clear picture of the linear continuum of the performance of persons in comparison to the Qiyas for L1 Arabic language test items. The left side represents people on the interval logit scale continuum. The upper left quadrant represents people who were more skillful in Arabic, whereas the lower left quadrant indicates people who were less skillful. The right side of the map represents Qiyas for L1 Arabic language test items. More difficult items are located closer to the top and easier items are located closer to the bottom of the graph. The letter "M" is the distribution mean for both items and persons, and "SD" is one standard deviation. The

```
MEASURE    Person - MAP - Item
               <more>|<rare>
    3                +
                     |
                     |
                     |
                  .  |
                     |
                 .#  |
                  # T|  RE5
    2                +
                 .#  |T RE3
                 ##  |  WR_EI6
              .####  |
               ####  |
               .###  |  RE1
               #### S|  ST_SC6
               .###  |
    1     #########  +  RC11    RC8     RE6     ST_EI1
             .#####  |S ST_EI8  WR_EI2
                ###  |  RC14    ST_EI5
        ###########  |  ST_SC4
               ####  |
               #### M|  RC3     RC4     RE2
            .######  |
            #######  |  RC9     ST_EI4  ST_SC3  ST_SC7  WR_SC8
    0         .####  +M RC2     RC5     ST_EI7  ST_SC1  WR_SC4
              .####  |  ST_EI2  ST_SC5
           .#######  |  RC12    RE4     ST_EI6
               .###  |  RC10    ST_EI3  WR_EI4  WR_SC2
                .##  |  WR_SC7
                ### S|
             ######  |  ST_SC2  WR_EI3
                 .#  |S WR_EI1  WR_EI5
   -1            ##  +  WR_SC1
                 .#  |  RC1     RC13    WR_SC3
                 .#  |  WR_SC5
                 ##  |  ST_SC8  WR_SC6
                .## T|  RC6
                  #  |
                 .#  |  RC7
                     |T
   -2                +
               <less>|<frequent>
```

Fig. 1 Person-item map for the Qiyas L1 Arabic test

symbol "T" represents two standard deviation units. The mean item distribution was set to 0. Figure 1 indicates that the average person distribution was 0.33 logits higher than the distributions of items by one-third SD and was negatively skewed. This suggests that the Qiyas for L1 Arabic language test items were slightly easier for this group. Some gaps in the item location distribution exist in the map, particularly in the middle and at the top right side of the map. This finding indicates that people in the middle and upper levels of the distribution were not reasonably targeted by the Qiyas for L1 Arabic language test items, that is, the content aspect of the construct under study lacked some representation, potentially compromising the validity of the test (Messick 1989). Therefore, a future version of Qiyas for L1 should include more items that accurately represent the targeted group level.

The Qiyas for L1 Arabic language test items were examined using mean square outfit statistics for the item and point-measure correlations (see Appendix 1). The item mean square outfit ranged between 0.7 and 1.73. The item mean square outfit is a Rasch-based model with standardized residuals used for assessing item fit. The item mean square outfit statistic is relatively more sensitive to patterns of misfit far from the person trait level. The expected value of this index is 1. However, dataset items that adequately fit the Rasch model are expected to range between 0.7 and 1.30 (Wright and Linacre 1994). Table 6 (see Appendix 1) shows that only five out of 50 items in the Qiyas for L1 test failed to fit Rasch model expectations. These items were as follows: items 19 and 20 in rhetorical expression, items 26 and 28 in structure, and item 50 in writing accuracy. These five items should be inspected carefully and then either modified or removed from the Qiyas L1 test because they contain construct-irrelevant variance that threatens the structural validity of the construct (Messick 1989). Additionally, inspection of the point-measure correlations index indicates that all but two Qiyas for L1 test items were positively correlated with the construct. Point-measure correlations in the Rasch model are analogous to point biserial correlation in classical test theory and describe how well each item contributes to the total test score. For example, Table 6 (see Appendix 1) shows that items 19 and 20 associated with rhetorical expression had point-measure correlations of 0.08 and 0.09, respectively. This small positive correlation suggests that these two particular items were either not functioning as intended or did not contribute adequately to the construct. A common practice in a Rasch investigation is to modify or delete any negative or close-to-zero point-measure correlations because they contradict the construct of interest (Linacre 1998; Bond and Fox 2007). Thus, all five items that misfit the Rasch standard, including those with close-to-zero point-measure correlations, were removed for further investigation of the internal structure of the Qiyas for L1 Arabic language test.

Stage 2

Unidimensionality analysis

The structural aspect of validity of the Qiyas for L1 Arabic language test was further investigated using PCAR. The PCAR is a factor analysis of residuals, after the Rasch model is applied to the data. The factor analysis of these residuals is used to identify common variance shared among data that is unexplained by the Rasch model. If a dominant measure not explained by the Rasch model is found among the items, then it

Table 3 PCARs of Qiyas for L1 Arabic language test data

Standardized residual variance (in Eigenvalue units)		Observed (%)		Expected (%)
Total raw variance in observations	59.4	100.00		100.00
Raw variance explained by measures	14.4	24.20		24.30
Raw variance explained by persons	6.1	10.30		10.40
Raw variance explained by items	8.3	13.90		13.90
Raw unexplained variance (total)	45	75.80	100.00	75.70
Unexplained variance in first contrast	2.3	3.80	5.00	
Unexplained variance in second contrast	1.9	3.30	4.30	

can be inferred that a dimension other than the intended dimension has interfered with the test data. This calls into question the structural aspect of validity of the measure. Linacre (2012a, 2012b) argued that for test data to be unidimensional, the smallest eigenvalue for the contrasts in the residuals is 2 items in unit strength. Simulation studies have shown that eigenvalues might reach 2 accidentally (Raîche 2005; Linacre 2012a, 2012b). Table 3 displays the results of the Rasch PCAR analysis after removing misfit items. The total variance explained by the Rasch model was 14.4. The small percentage of explained variance could be due to narrow ranges of ability in examinees or the difficulty level of some items. In other words, similar abilities among examinees and equal difficulty of test items could have caused the small total explained variance observed. Table 4 shows that the first contrast remaining after Qiyas for L1 Arabic language test data were fit to the Rasch model and had strength of 2.3 out of 45 items. This contrast explained 3.8% of the variance, which exceeded the benchmark of 2 eigenvalue units suggested by Linacre (2012a, 2012b) and could be indicative of multidimensionality.

Inspection of Table 4 indicates that items 44 and 37 in writing accuracy and item 31 in structure all had large positive factor loadings of 0.40. The clustering of these three items is notable because it suggests that they have a common meaning that is different from the yardstick of Rasch measurement (Bond and Fox 2007). This finding is evidence of a secondary dimension with a small influence. In general, any item loading ≥ 0.40 should be investigated (Bond and Fox 2007).

Because the results indicated the presence of an influential three-item cluster on the Qiyas for L1 Arabic language test, two separate Rasch calibration analyses were conducted to determine if person measures were severely affected by the secondary dimension (Wright and Stone 1979; Linacre 2012a, 2012b). A confirmatory finding would indicate that those items should be either modified or removed from the Qiyas for L1 Arabic language test, because the construct runs a risk of being distorted by this irrelevant

Table 4 Factor loadings of Qiyas for L1 Arabic language test items that signify multidimensionality

Contrast	Loading	Measure	Infit MnSq	Outfit MnSq	Item	Loading	Measure	Infit MnSq	Outfit MnSq	Item
1	0.62	− 0.3	0.81	0.72	44	− 0.36	0.12	1	0.97	2
1	0.53	− 1.12	0.86	0.76	37	− 0.36	0.52	1.16	1.15	3
1	0.43	0.05	0.83	0.77	31	0.33	1.55	1.05	1.1	15
1	0.35	− 0.39	0.78	0.71	39	− 0.3	0.16	1.03	1.03	5

sub-dimension. The first Rasch analysis targeted only the three-item cluster with positive loadings (writing accuracy). The second Rasch calibration targeted items with negative loadings. If the Qiyas for L1 Arabic language test fits the Rasch standard, then the person measures should remain invariant, allowing for a reasonable number of errors, that is, the person measures obtained from the two calibrations should fall within the 95% two-sided confidence interval (Bond and Fox 2007; Linacre 2012a, 2012b).

As depicted in Fig. 2, one or two person measures fell outside the specified 95% confidence interval, and the correlation between the two sets of person measures was 0.66. This result implies that the Qiyas L1 Arabic language test is unidimensional.

DIF analysis

DIF analysis was also performed to determine whether any irrelevant factors interfered with the construct under study, and an analysis of test and item bias within the framework of the Rasch model was applied to the Qiyas for L1 Arabic language test data. A review of the literature of test bias (Crocker and Algina 1986) indicated that bias exists when test outcomes or results reflect irrelevant factors or characteristics outside the construct of interest (e.g., demographic variables). By this definition, bias would impair the construct validity by means of test score interpretation. Therefore, to investigate whether the Qiyas for L1 Arabic language test data produced assessment scores that reflected only the construct of interest, a Rasch analysis of uniform differential test and item functioning by gender (e.g., male versus female) was implemented. Items that misfit the Rasch model were first excluded. After removing five misfit items, a DIF analysis by gender was performed on the remaining 45 items to determine whether the Qiyas for L1 Arabic language test produced invariant scores in the cross-gender subgroup classification. Bond and Fox (2007) suggested that item measures obtained from two Rasch analyses must fall within the 95% two-sided

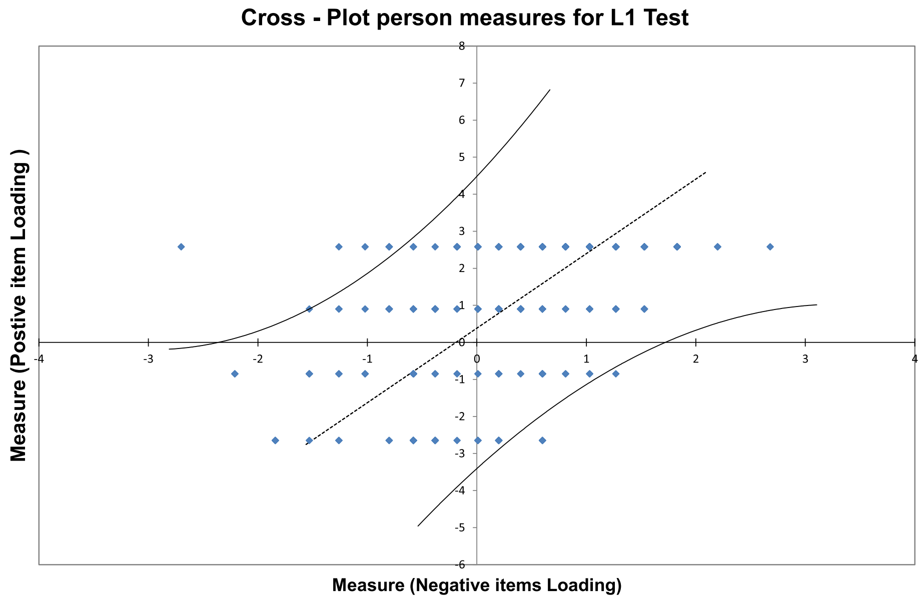

Fig. 2 Cross-plot person measures for the Qiyas L1 Arabic test

confidence interval to be invariant. Thus, two separate Rasch analyses for male and female subgroups were conducted. The two item measures obtained from each analysis were then cross-plotted, as shown in Fig. 3.

The dashed line in the middle of Fig. 3 represents the Rasch model best-fit line. The two curved solid lines surrounding the best-fit line represent the 95% confidence interval. The plot shows that, except for a few items, the majority of the Qiyas for L1 Arabic language test data fall within the 95% confidence interval and cluster around the Rasch best-fit line across male and female subgroups. This result implies that the Qiyas for L1 Arabic language test produces item measures that are invariant with respect to gender.

In the next step, DIF analysis at the item level was implemented to determine the specific items that exhibited bias across males and females. Linacre (2012a, 2012b) recommended two criteria: first, the probability of the item DIF should be small, that is, the probability of the item DIF must be statistically significantly different, with $p \leq 0.05$. Second, the DIF contrast must be at least 0.5 logit to merit a noticeable DIF difference. Test data were subjected to a uniform DIF analysis by gender (male versus female). Table 5 displays the results of DIF analysis for items that exhibited a significantly noticeable DIF. Items 1, 6, 9, and 10 in reading comprehension, item 16 in rhetorical expression, and item 36 in writing accuracy were statistically significant at an alpha level of 0.05 and had DIF contrasts at or above the required 0.5 logit. The same findings are depicted in Fig. 4. The gendered responses on items 1, 6, and 36 with logit values of 0.85, 0.95, and 0.66, respectively, indicated that those items were more difficult for males than for females. Conversely, items 9, 10, and 16 with logit values of − 0.66, − 0.63, and − 1.00, respectively, were easier for males than for females. Those flagged items would benefit from further investigation by the test developers and subject-matter reviewers of the Qiyas for L1 Arabic test, to determine why they differed

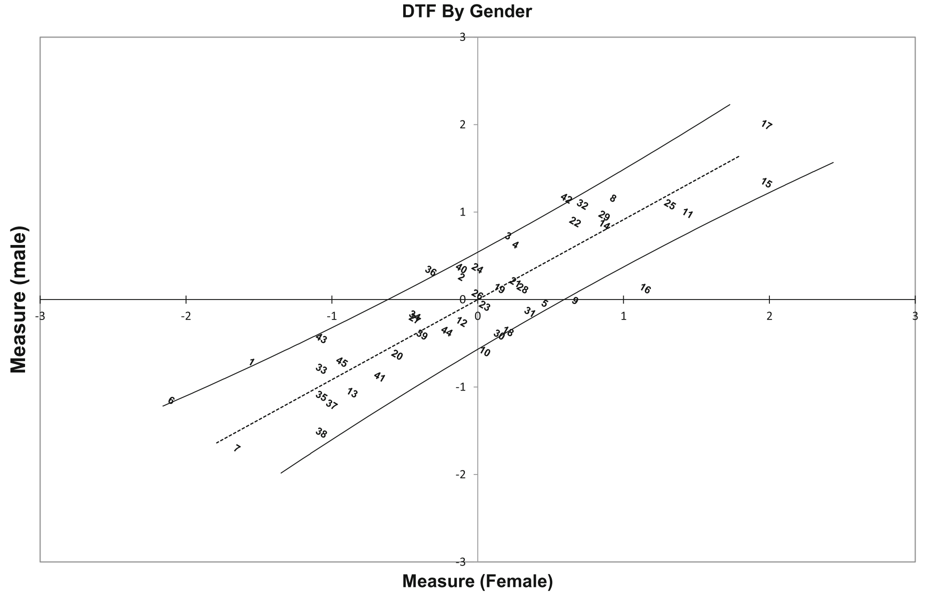

Fig. 3 DTF by gender for the Qiyas L1 Arabic language test

Table 5 DIF analysis of the Qiyas for L1 Arabic language test at the item level

DIF measure		DIF contrast	Rasch–Welch's *t*	*p*	Item
Male	Female				
1	2	0.85	2.32	0.02	1 RC
1	2	0.95	2.24	0.03	6 RC
1	2	− 0.66	− 2.26	0.03	9 RC
1	2	− 0.63	− 2.09	0.04	10 RC
1	2	− 1.00	− 3.35	0.00	16 RE
1	2	0.66	2.22	0.03	36 WR

between males and females. Overall, the results of the uniform DIF analysis within the framework of the Rasch measurement model lent support to the structural aspect of validity of the Qiyas for L1 Arabic language test from the perspective of gender. The complete DIF analysis is given in Appendix 2.

Discussion

The purpose of this study was to investigate the psychometric properties of the field-testing version of the Qiyas for L1 Arabic language test using Rasch analysis. To this end, several Rasch quality-control fit-statistic indices were used. At the test level, the Qiyas for L1 Arabic language test mean square outfit statistic for the test level was 1.02, indicating that it fit the Rasch standard reasonably well. However, at the item level, mean square outfit statistics indicated that five out of 50 items on the Qiyas L1 Arabic language test were misfits, according to the Rasch standard. Those items would benefit from modification or should be deleted from the Qiyas for L1 test because they add irrelevant variance that could distort the precision of measurement of the construct and compromise the structural aspect of validity. The Rasch reliability analysis of the

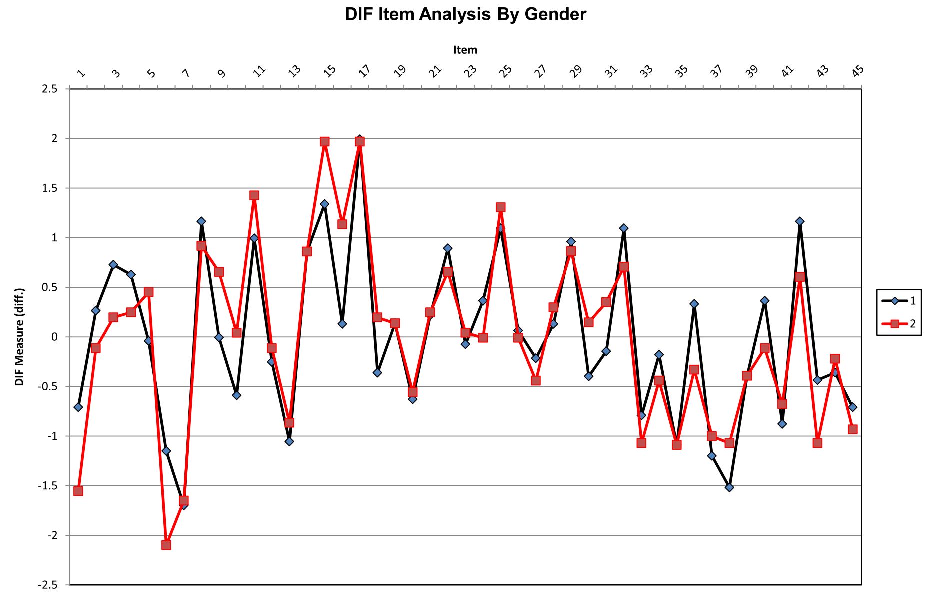

Fig. 4 DIF analysis by gender

Qiyas for L1 Arabic language test was 0.86, indicating that the test is sufficiently reliable in test scores. The person-item map of the Qiyas for L1 test indicated that the average person location is higher than the average item location by one-third SD, suggesting that the Qiyas for L1 Arabic language test items are slightly easier compared with the ability of those taking the test. In fact, the small mismatch between the locations of persons and items could be explained by the gaps at the top right and middle of the item locations. Those gaps clearly illustrate that there was a content representation deficiency in the Qiyas for L1 Arabic language test, which suggests that the construct of interest is under-represented. Construct Under-representation could threaten the content aspect validity of the construct, for example, if not enough items target the average- and higher-ability groups. This issue of construct underrepresentation could be resolved by instructing the Qiyas for L1 Arabic language test developer to add more items targeting specified groups. This modification would likely improve the quality and content representation of the Qiyas for L1 Arabic language test. However, it should be emphasized that the addition of more items should be driven by practical applications of measurement. If important decisions are being made about performance on a scale, then the scale should have sufficient item coverage. If a particular group or subset of a group is being targeted, then that targeted group should have sufficient item coverage on the scale.

The structural aspect of the construct under study was further investigated using PCAR and DIF methods in the second-stage analysis, after removing unwanted items. Findings from the PCAR analysis lent support to the assumption of unidimensionality of the Qiyas for L1 Arabic language test for assessing four skills, whereas the DIF analysis flagged six items that exhibited significant DIF and merit further investigation. Those identified items should be reviewed because they contain construct-irrelevant variance that could alter the precision of measurement and threaten the structural aspect of validity of the construct. The results of the Rasch-based DIF analysis could inform the developers and content experts of the Qiyas for L1 Arabic language test in determining what caused those particular six items to differ between males and females, before concluding that they are biased items.

Overall, the findings of this study indicate that the field-testing version of the Qiyas for L1 Arabic language test has satisfactory psychometric properties. However, two limitations of the study should be noted. First, the data used in this study were not collected from real job applicants and other high school- and university-level students across Saudi Arabia. Instead, they were collected from students in high schools and with college-level education in Riyadh and may therefore not be representative of the intended population of the Qiyas for L1 Arabic language test. Thus, the generalizability of the findings of this study is limited; additional investigations using more representative samples are warranted. Second, the sample size used to conduct the DIF analysis was small. Smith (2004) noted that Rasch-based DIF methods lack the power to detect biased items of less than 0.5 logits when the sample size is smaller than 500 people in each subpopulation. Moreover, DIF studies commonly produce nonreplicable results (Linacre 2012a, 2012b). Consequently, it is possible that the six items flagged in this study would not be flagged in a different study. Follow-up studies using greater sample sizes are needed to confirm the results reported here.

Last, the findings of this study suggest new venues for future research studies. First, taking into account the representative sample targeted in the Qiyas for L1 Arabic language test, it would be beneficial to cross-validate this study using different measurement models (e.g., classical test theory models, item response theory models, structural equation modeling) and compare the results of those models with those obtained using the Rasch measurement framework. Such comparisons would fill some gaps in knowledge associated with using a Rasch measurement framework alone. For instance, validity is typically viewed as a unitary concept that embodies all evidence that supports the intended interpretation of test scores for the proposed use (AERA, APA, and NCME, 2014). In this study, only psychometric features related to the structural and content aspects of validity were investigated. Validity arguments (Messick 1989) that include findings from different measurement and statistical models would add substantial value to the intended interpretation of the Qiyas for L1 Arabic language test score. Second, a standard-setting analysis is an integral component of test development, especially in testing situations related to education and licensing. Additional studies are needed to determine if the four-level categorization of language attainment, as specified by the Qiyas for L1 test developers, is reasonably defined in the field-testing version of the Qiyas for L1 Arabic language test.

Third, the ultimate objective of the Qiyas for L1 Arabic language test is to serve as a standardized tool that assesses modern standard Arabic language skills, not only in Saudi Arabia but throughout the Arab world. It would therefore be beneficial for the NCA to carry out cross-cultural studies of the Qiyas for L1 Arabic language test in other Arab countries. Such studies would strengthen the reliability and validity of the test and also broaden its usability, resulting in greater international recognition.

Conclusion

The overall aim of this study was to highlight the usability, applicability, and informative nature of the Rasch measurement framework in the field of language testing. The specific objective was to investigate the psychometric properties of the test during field-testing using a Rasch model. The initial findings of the Rasch analysis indicated that the Qiyas for L1 Arabic language test has satisfactory psychometric properties. However, this result should be interpreted with caution given the limitations of the sample population used. Thus, continued investigation of the psychometric proprieties of the test is necessary to ensure its appropriate use as a tool of assessment for modern Arabic language skills. Nevertheless, because the test data conformed to model expectations, developers of the Qiyas for L1 Arabic language test would likely benefit from these findings during field-testing for development and validation, particularly when the sample size is small. For instance, test developers could be guided by the results of the Rasch analysis in efforts to improve effectiveness of the assessment tool by adding, removing, or modifying some items. Test developers and measurement practitioners would also benefit from using Rasch analysis to evaluate the psychometric features of the test to assess construct-related validity (e.g., structural and content aspects of validity of test scores); this assessment would support the interpretation of test scores.

Appendix 1

Table 6 Summary of item measures, outfit indices, and PMCs of Qiyas for L1 Arabic language test data

Item	Measure	SE	Outfit MnSq	PMC
19	2.17	0.18	1.74*	0.08*
20	1.02	0.14	1.42*	0.09*
50	1.74	0.16	1.36*	0.21
26	1.3	0.15	1.33*	0.24
28	− 1.35	0.17	1.33*	0.36
17	1.85	0.17	1.29	0.18
24	0.68	0.14	1.29	0.16
6	− 1.55	0.18	1.23	0.21
36	0.82	0.14	1.2	0.28
4	0.36	0.14	1.15	0.28
3	0.4	0.14	1.13	0.27
14	0.74	0.14	1.13	0.29
9	0.15	0.14	1.12	0.32
11	1.02	0.14	1.12	0.32
8	0.94	0.14	1.1	0.29
15	1.42	0.15	1.07	0.3
40	− 0.03	0.14	1.07	0.34
38	− 0.38	0.15	1	0.34
25	− 0.13	0.14	1.04	0.38
31	− 0.41	0.15	1.03	0.35
46	0.82	0.14	1.03	0.34
18	− 0.24	0.14	1.02	0.37
5	0.05	0.14	1.02	0.37
33	0.8	0.14	1.02	0.37
7	− 1.76	0.19	0.91	0.31
21	0.03	0.14	0.99	0.39
44	0.07	0.14	1.01	0.39
16	0.4	0.14	1	0.4
10	− 0.43	0.15	0.94	0.41
2	0.01	0.14	0.95	0.42
37	− 0.99	0.16	0.98	0.41
1	− 1.1	0.16	0.88	0.41
23	0.11	0.14	0.95	0.44
32	0.09	0.14	0.96	0.43
47	− 0.77	0.15	0.93	0.41
45	− 0.89	0.16	0.88	0.42
13	−1.07	0.16	0.82	0.43
27	0.11	0.14	0.91	0.47
12	− 0.3	0.14	0.9	0.46
29	1.04	0.15	0.92	0.49
34	− 0.28	0.14	0.9	0.47

Table 6 Summary of item measures, outfit indices, and PMCs of Qiyas for L1 Arabic language test data *(Continued)*

Item	Measure	SE	Outfit MnSq	PMC
22	− 0.7	0.15	0.79	0.5
49	− 0.89	0.16	0.88	0.47
39	− 1.18	0.17	0.78	0.48
42	− 1.42	0.18	0.77	0.48
30	− 0.07	0.14	0.79	0.56
41	− 1.21	0.17	0.74	0.51
35	− 0.05	0.14	0.78	0.57
48	− 0.41	0.15	0.73	0.59
43	− 0.49	0.15	0.71	

*Indicates misfit items. Bold type indicates negative and small positive point-measure correlations

Appendix 2

Table 7 Summary of item bias analysis of Qiyas for L1 Arabic language test data by gender

DIF measure		DIF contrast	Rasch Welch's *t*	*p*	Item
Male	Female				
− 0.71	− 1.55	0.85	2.32	0.02	1
0.26	− 0.11	0.38	1.29	0.19	2
0.73	0.2	0.53	1.83	0.06	3
0.63	0.25	0.38	1.31	0.19	4
− 0.04	0.45	− 0.49	− 1.68	0.09	5
− 1.15	− 2.1	0.95	2.24	0.02	6
− 1.7	− 1.65	− 0.05	− 0.12	0.9	7
1.16	0.92	0.25	0.83	0.4	8
− 0.01	0.65	− 0.66	− 2.26	0.02	9
− 0.59	0.04	− 0.63	− 2.09	0.03	10
0.99	1.43	− 0.43	− 1.41	0.16	11
− 0.25	− 0.11	− 0.14	− 0.46	0.64	12
− 1.06	− 0.87	− 0.19	− 0.57	0.57	13
0.86	0.86	0	0	1	14
1.34	1.97	− 0.63	− 1.88	0.06	15
0.13	1.13	− 1	− 3.35	0	16
1.99	1.97	0.02	0.06	0.94	17
− 0.36	0.2	− 0.56	− 1.88	0.06	18
0.14	0.14	0	0	1	19
− 0.63	− 0.56	− 0.07	− 0.23	0.81	20
0.22	0.25	− 0.03	− 0.1	0.91	21
0.89	0.65	0.24	0.81	0.41	22
− 0.08	0.04	− 0.12	− 0.4	0.68	23
0.36	− 0.01	0.37	1.28	0.2	24
1.09	1.31	− 0.21	− 0.69	0.48	25
0.06	− 0.01	0.07	0.25	0.8	26
− 0.22	− 0.44	0.23	0.74	0.45	27

Table 7 Summary of item bias analysis of Qiyas for L1 Arabic language test data by gender *(Continued)*

DIF measure		DIF contrast	Rasch Welch's *t*	*p*	Item
Male	Female				
0.13	0.3	− 0.17	− 0.58	0.56	28
0.96	0.86	0.1	0.33	0.74	29
− 0.4	0.15	− 0.54	− 1.83	0.06	30
− 0.15	0.35	− 0.49	− 1.69	0.09	31
1.09	0.71	0.39	1.32	0.18	32
− 0.79	− 1.07	0.28	0.83	0.4	33
− 0.18	− 0.44	0.26	0.86	0.39	34
− 1.09	− 1.09	0	0	1	35
0.33	− 0.33	0.66	2.22	0.02	36
− 1.2	− 1	− 0.2	− 0.58	0.56	37
− 1.52	− 1.07	− 0.45	− 1.24	0.21	38
− 0.39	− 0.39	0	0	1	39
0.36	− 0.11	0.48	1.63	0.1	40
− 0.88	− 0.68	− 0.2	− 0.62	0.53	41
1.16	0.6	0.56	1.91	0.05	42
− 0.44	− 1.07	0.64	1.92	0.05	43
− 0.36	− 0.22	− 0.14	− 0.47	0.64	44
− 0.71	− 0.93	0.22	0.68	0.49	45

Abbreviations

DIF: Differential item functioning; ICCs: Item characteristic curves; IRT: Item response theory; NCA: National Center for Assessment; PCAR: Principal component analysis of residuals; PMC: Point-measure correlation; SD: Standard deviation; SE: Standard error

Acknowledgements

This research paper was supervised by the Department of Research and Studies at the National Center for Assessment, Saudi Arabia. Their cooperation was highly appreciated. Also, I sincerely express my deep gratitude to Dr. Abdul Rahman Shamrani, director of language testing at the National Center for Assessment, Riyadh, Saudi Arabia, for providing information with respect to test content and for his thoughtful review of the initial draft of this research paper.

Availability of data and materials

The datasets supporting the conclusions of this article are available from the National Center for Assessment (NCA), Riyadh, Saudi Arabia. Copyright for the data also belongs to the NCA.

Authors' contributions

Amjed Al-Owidha is the sole contributor to this research paper. The author read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 29 March 2018 Accepted: 7 June 2018
Published online: 20 July 2018

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Birnbaum, A (1968). Some latent trait models and their use in inferring an examinee's ability. In FM Lord, MR Novick (Eds.), *Statistical theories of mental tests*. Reading, MA: Addison-Wesley.

Bond, TG, & Fox, CM (2007). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Boone, WJ, Staver, JR, Yale, MS (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Crocker, L, & Algina, J (1986). *Introduction to classical and modern test theory*. Philadelphia: Harcourt Brace Jovanovich College Publishers.

De Ayala, RJ (2009). *The theory and practice of item response theory*. New York: Guilford Press.

De Jong, J.H.A.L. (1983). Focusing in on a latent trait: an attempt at construct validation by means of the Rasch model. In Van Weeren, J. (Ed.), Practice and problems in language testing 5. Non-classical test theory; final examinations in secondary schools. Papers presented at the International Language Testing Symposium (Arnhem, Netherlands, March 25–26, 1982) (pp. 11–35). Arnhem: Cito.

De Jong, J.H.A.L., & Stoyanova, F. (1994). Theory building: Sample size and data-model fit. Paper presented at the annual Language Testing Research Colloquium, Washington, DC.

Dinero, TE, & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement*, *1*(4), 581–592.

Frantom, C.G., & Green, K.E. (2002). Survey development and validation with the Rasch model. Paper presented at the international conference on questionnaire development, evaluation, and testing, Charleston, SC.

Hambleton, R.K., & Cook, L.L. (1978). Some results on the robustness of latent trait models. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Holland, WP, & Thayer, DT (1988). Differential item performance and the Mantel-Haenszel procedure. In H Braun Wainer, HI Braun (Eds.), *Test validity*, (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hutten, LR (1981). *The fit of empirical data to latent trait models (doctoral dissertation)*. Amherst, MA: University of Massachusetts.

Leeson, H, & Fletcher, R (2003). *An investigation of fit: comparison of the 1-, 2- ,3- parameter IRT models to the project TTle data*. Auckland, New Zealand: Paper presented at the Australian Association for Research.

Linacre, JM. (1994). Sample size and item calibrations stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, JM. (1998). Detecting multidimensionality: which residual works best. *Journal of Outcome Measurement*, *2*(3), 266–283.

Linacre, JM (2012a). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, OR: Winsteps.com.

Linacre, JM (2012b). *Winsteps® (Version 3.75.1) [Computer Software]*. Beaverton, Oregon: Winsteps.com.

Lord, FM (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, FM, & Novick, MR (1968). *Statistical theories of mental tests*. Reading, MA: Addison-Wesley.

McNamara, T, & Knoch, U. (2012). The Rasch wars: the emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–577.

Messick, S (1989). Validity. In RL Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). New York: Macmillan.

Nakamura, Y. (2007). A Rasch-based analysis of an in-house English placement test. http://hosted.jalt.org/pansig/2007/HTML/Nakamura.htm. Accessed 18 Feb 2018.

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, *19*(1), 1012.

Smith, RM (2004). Detecting item bias with the Rasch model. In S Jr., RM Smith (Eds.), *Introduction to Rasch measurement*, (pp. 391–418). Maple Grove, MN: JAM Press.

Stage, C. (1996). An attempt to fit IRT models to the DS subtest in the SweSAT. (Educational Measurement No 19). Umeå University, Department of Educational Measurement.

Thissen, D (1991). *MUTLTILOG user's guide: multiple categorical item analysis and test scoring using item response theory (Version 6.0)*. Chicago: Scientific Software.

Thissen, D, Steinberg, L, Wainer, H (1993). Detection of differential item functioning using the parameters of item response models. In PW Holland, H Wainer (Eds.), *Differential item functioning*, (pp. 67–113).

Wright, BD (1984). *MESA Research [Memorandum 41]*. Hillsdale: Lawrence Erlbaum Associates http://www.rasch.org/memo41.htm. Accessed 18 Oct 2006.

Wright, BD, & Masters, GN (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, BD, & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*(1), 23–48.

Wright, BD, & Stone, MH (1979). *Best test design*. Chicago: MESA Press.

Wright, BD, Mead, RJ, Draba, R (1976). *Detecting and correcting test item bias with a logistic response model (Research Memorandum 22)*. Chicago: University of Chicago, MESA Psychometric Laboratory.

Zumbo, BD (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.