International Education Studies; Vol. 10, No. 2; 2017
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Factor Structure and Reliability of Test Items for Saudi Teacher Licence Assessment

Abdullah Saleh Alsadaawi[1]

[1] King Saud University, National Centre for Assessment, Saudi Arabia

Correspondence: Abdullah Saleh Alsadaawi, King Saud University, National Centre for Assessment, Saudi Arabia. E-mail: alsadaawi@gmail.com

Received: August 15, 2016 Accepted: September 20, 2016 Online Published: January 30, 2017

doi:10.5539/ies.v10n2p26 URL: http://dx.doi.org/10.5539/ies.v10n2p26

Abstract

The Saudi National Assessment Centre administers the Computer Science Teacher Test for teacher certification. The aim of this study is to explore gender differences in candidates' scores, and investigate dimensionality, reliability, and differential item functioning using confirmatory factor analysis and item response theory. The confirmatory factor analysis results for 6 371 examinees' scores of 66 multiple-choice items when grouped into three content domains showed that the test data were unidimensional (ability, trait). The domains were highly correlated (0.883 to 0.949) within this dimension. Data reliability estimated through latent variable modelling was acceptable at 0.848. Gender results for DIF signalled 13 items, five cases against males and eight cases against females; a finding of some balance in DIF direction against males and females. The study results confirm the validity of the Computer Science Teacher Test and support further refinement of multiple forms of the test.

Keywords: teacher assessment, Saudi Arabia, confirmatory factor analysis, differential item functioning, gender differences

1. Introduction

Teacher assessment is used for measuring and supporting pre-teacher education outcomes and teachers' professional development. In a review, DeLuca and Bellara (2013) found a multitude of teacher assessment standards used by national educational authorities with numerous assessment literacy measures. Further, the authors noted shifts in teacher education curricular concepts together with evolution in the national measures for student outcomes. In another review, Blömeke and Delaney (2014) noted that whilst teacher assessment studies from North America and other English-speaking countries focussed on internal assessment systems and practices and contained some cross-country comparisons, a trend to cultural comparison of teacher assessment systems had not yet emerged.

This paper first introduces the Saudi educational environment, and this is followed by a short literature review. The methodology and results are presented and discussed, and conclusions drawn.

2. Saudi Education System

The Ministry of Education is the sole authority for education in Saudi Arabia, providing a free education for all Saudi students through to higher education. The Ministry also oversees a small educational private sector, generally for expatriates. Saudi schools are gender segregated, thus there is a significant number of men in the profession. The World Bank (2016) reported that in 2014 there were 761 737 trainees and teachers, pre-school to secondary school, of whom 52 per cent were women.

Teachers in Saudi Arabia are viewed in two ways. Because of their early association with mosques they are admired, although the secular education system set up after the 1932 declaration of the Kingdom foundered for decades due to its concepts of over-worked and underpaid teachers (Al-Rasheed, 2010). From the 1950s the corporation Saudi Aramco (2016) assisted the government in establishing schools to alleviate issues with illiteracy, eventually building 139 schools. Initially, boys only were permitted an education due to traditional beliefs in the conservative society, but by 1960 the first primary school for girls was opened with one student (Bowen, 2015).

As oil revenues became available, Saudi Arabia was in a better position to plan for the future, and in 1970 set in place the first of its five-year economic plans. Education was a priority, both for literacy for the population and

to provide the nascent public sector with Saudis to replace the largely foreign workforce (Alshahrani & Alsadiq, 2014). However, the population growth in the late 20th century surpassed the Ministry of Education's ability to provide all Saudis with a quality education, and by the 6th economic development plan (1995-1999) a concentrated effort was made to improve the 'Saudisation' of the country's workforce, that is, replacing skilled expatriates with skilled Saudis. This emphasis on education continues today (Ahmed, 2016).

3. Teacher Education

Teaching in 20th century Saudi schools was criticised as being conservative and didactic (Norton & Syed, 2003). Teachers' education was expected to be at bachelor degree standard, but due to the pressure of population growth, this was not enforced and diplomates were accepted. Pedagogical practices were didactic, teachers explained principles of the curriculum, but did not engage the students who were thus passive learners, recording their lessons and memorising for examinations. A report for the Ministry of Education recommended, inter alia, improved teacher education, and in 2004 the Ministry embarked on a decade-long plan (Tatweer) to improve the quality of education in the Kingdom (International Bureau of Education, 2011).

As part of Tatweer's emphasis on teacher education, competencies were prepared for pedagogical, numeracy and literacy skills; however, these were not adequately administered and did not achieve the standards expected (Alzaydi, 2011; Alsharif, 2011; Al Shannag, Tairab, Dodeen, & Abdel-Fattah, 2013). Elyas and Pickard (2013) stated that teacher outcomes were challenged by variables in students' backgrounds, the rise of educational technology, and universities' hierarchies. Shortfalls in teacher competencies had external effects for the Ministry of Education. Comparing Saudi and Singapore results for grade 8 students from a 2007 international study (Trends in Mathematics and Science Study), Al Shannag et al. (2013) found that the Saudi teachers retained their teacher-centric style, whilst the more successful Singaporean teachers practised a student-centric educational system.

The Ministry of Education responded to these reports by implementing a change in focus from teacher to student. The National Centre for Assessment in 2010 developed a new teacher assessment framework, the National Professional Teacher Standards. The framework comprises 12 standards in two groups, the first of which was pedagogical: professional knowledge, promoting learning, supporting learning, and professional responsibility (Al-Saud & Al-Sadaawi, 2014). The second type is the subject-specific teaching standards for 25 curricular courses. The standards guide teacher licensing examinations, identify training needs for new teachers, and set the quality of teaching programs.

As an example, one of the courses is the Computer Science Teacher Test (CSTT) for secondary school. It consists of three domains: computer and math, engineering and science, computer applications, and computer and education. Based on the 2010 standards, the test has been administered to 20 028 candidates, of whom 37 per cent were female (Ministry of Education, 2016).

This study investigates the validity of the test data by examining their dimensionality and key features such as reliability, and differential item functioning on gender in the framework of item response theory.

4. Literature Review

Confirmatory factor analysis seeks relationships between measurement data, which is, test results or indicators, and is used to identify latent variables (factors) (Brown, 2015). Unlike exploratory factor analysis, confirmatory factor analysis is hypothesis-based, thus all aspects of the model are pre-specified. This form of analysis is used to 'verify the number of underlying dimensions of the instrument (factors) and the pattern of item-factor relationships (factor loadings)' (Brown, 2015, p.1). Netemeyer et al. (2013) stated that confirmatory factor analysis can be used to assess dimensionality (fit, correlated measurement errors, degree of cross-loading).

In designing tests and measures which produce large data such as the Computer Science Teacher Test, dimensionality refers to the homogeneity of items and sub-items. Netemeyer, Bearden, and Sharma (2003) explained that a unidimensional measure indicates a single latent variable that accounts for item data (responses), whereas a multidimensional measure has more than one latent variable among the data. In designing such tests, a unidimensional internal structure is a step towards establishing reliability (consistency between items) and validity (consistency between the measure's constructs). Whilst unidimensionality is used in confirmatory factor analysis, it is also a fundamental assumption in item response theory (Deng, Wells, & Hamilton, 2008).

In longitudinal research, the analysis of measurement invariance of latent constructs is important as scores may vary over time. For example, in education, repetitive examination of cohorts of students determines the progress of individuals over the course of their education or is used to compare group scores. Measurement invariance was predated by Jöreskog's (1971, p.409) observation of 'similarities and differences in factor structures

between different groups'. Jöreskog posited that parameters in factor analysis models (factor variances, factor loadings, factor covariance and unique variances) may be constrained, or assigned an arbitrary value. Measurement invariance was introduced by Byrne, Shavelson, and Muthén (1989) using sensitivity analyses for stability in baseline models, 'determining partially invariant measurement parameters, and . . . testing for the invariance of factor covariance and mean structures, given partial measurement invariance' (Byrne et al. 1989, p. 456). Measurement invariance, or measurement equivalence, thus establishes that each iteration measures the same construct (latent variable).

Reliability concerns the permanent effect that is being investigated does persist from one sample to another. Raykov (2004) and Raykov, Dimitrov, and Asparouhov (2010) used latent variable modelling for measurement invariance and reliability. Raykov (2004, 2012) argued that coefficient alpha does not estimate scale reliability at population levels, and proposed another reliability coefficient model based on scale reliability rather than the restrictions of Cronbach's α (Cronbach, 1951). Cronbach's α requires that the factor loadings of all items are equal. More recently, Raykov, Gabler, and Dimitrov (2016, p.1) established a latent variable modelling procedure 'for point and interval estimation of the difference between the maximal and scale criterion validity coefficients'. This overcomes issues regarding the use of unidimensional multicomponent measures.

Criterion-related validity is one aspect of validating an instrument, referring to an item on a questionnaire actually measuring the intended outcome (Lodico, Spaulding, & Voegtle, 2010). The others include face validity (relevance of items to intent), content validity (items relevant to the content being measured). Criterion-related validity reflects the relationship between two scores on two different measures, and tests whether the outcome from the measure, its performance, can be predicted (Lodico et al. 2010). Raykov's (2007) latent variable modeling approach is used in this research for reliability and criterion validity.

Item response theory, a paradigm for the measurement of items in relation to the latent variable, is used extensively in education tests, including test construction, estimating ability and score reporting (Deng et al., 2008). Item response models take into consideration the degree of difficulty of each item in scaling items. Item response theory has, as noted, an assumption of unidimensionality (Deng et al. 2008).

Differential item functioning refers to the potential for bias in the test items which could skew data (be unfair) to sub-groups based on gender, race or age (Strobl, Kopf, & Zeileis, 2010). The bias may exist in a single item, or goodness-of-fit tests may show a trend, or a likelihood of bias among the variables.

There is a wide variety of statistical techniques for evaluating difference in both dichotomous and polytomous items (Gómez-Benito, Hidalgo, & Zumbo, 2013; Hambleton & Swaminathan, 2013; Sireci & Rios, 2013). Among these, that of Mantel-Haenszel (1959) remains a reference technique (Guilera, Gómez-Benito, Hidalgo, & Sánchez-Meca, 2013). Strobl et al. (2010) explained that testing for difference can be based on the specific sub-group supporting interpretation but leaving open the possibility of unexplained bias. At an extreme, all item parameter differences can be tested for bias among all possible sub-groups, leading to interpretation difficulty. Strobl et al. proposed a semi-parametric model using recursive partitioning to address this.

5. Methodology

The data were the scores of 6 371 examinees on 66 multiple-choice items on the Saudi Computer Science Teachers Test. The test had four response options per item, one only of which was correct, so the item scoring is 1 for correct response and 0 otherwise. The test items were classified as follows:

Domain 1: Computer and math, engineering and science (35 items).

Domain 2: Computer applications (12 items).

Domain 3: Computer and education (19 items).

Confirmatory factor analysis was used to test the validity of hypothesised models of the test and its three content-specific domains. The first question concerned the dimensionality of the data. Three different confirmatory factor analysis models were tested and compared on data fit with the teacher test scores: model A: a one-factor model; model B: a three-factor model with the three content-specific domains as correlated latent factors; and model C: a three-factor model with the three content-specific domains as uncorrelated latent factors. The models were tested for data fit using the program Mplus (Muthén, 2016). In the Mplus syntax for the three models, the factor indicators (test items) were declared as categorical variables because the item scores are dichotomous (0/1). Thus the factor analysis was based on the tetrachoric correlations (i.e., observed values are dichotomous) for the scores of the test items. This avoided issues using Pearson correlations for factor analysis of categorical variables. The analysis of test data for item response theory used the program Xcalibre 4 (Assessment Systems, 2016).

The score reliability was estimated through the use of a latent variable modelling (LVM) approach taking into account the binary nature of the item scores (Dimitrov, 2012; Raykov, 2007; Raykov et al., 2010). The congeneric model for latent normal variables Y_1^*, Y_2^*, …., Y_p^*, assumed to underlie a set of binary items Y_1, Y_2, …, Y_p according to Jöreskog (1971) is:

$$Y_i^* = \lambda_i \eta + \varepsilon_i \quad (1)$$

where η is a common latent factor with a variance set equal to 1, λ_i are factor loadings, ε_i are latent disturbances, and the probability of correct response on Y_i is given by the area under the standard normal curve to the right of a pertinent threshold κ_i ($i = 1, 2, \ldots, p$). Under this model, the score reliability, ρ, is estimated through the following equation (e.g., Bollen, 1989):

$$\rho = \frac{(\lambda_1 + \lambda_2 + \cdots + \lambda_p)^2}{(\lambda_1 + \lambda_2 + \cdots + \lambda_p)^2 + \mathrm{VAR}(\varepsilon_1) + \mathrm{VAR}(\varepsilon_2) + \cdots + \mathrm{VAR}(\varepsilon_p)} \quad (2)$$

where the numerator represents the true-score variance and the denominator represents the total variance (i.e., the sum of true variance and error variance).

Cronbach's α for internal consistency (reliability) was also used, however, the results underestimated the reliability obtained from the latent variable modelling approach, confirming the literature review discussion. Further, under the congeneric measurement model in equation 1, the assumption of tau-equivalency is met when the factor loadings are equal, $\lambda_1 = \lambda_2 = \ldots = \lambda_p$ (e.g., Jöreskog, 1971).

In differential item functioning analyses, groups are compared on item performance after adjusting for overall performance on the measured trait (Hambleton & Swaminathan, 2013). The Mantel-Haenszel techniques under the null hypothesis are distributed as a chi-square distribution with one degree of freedom. Under this procedure, an effect size estimate based on the common odds ratio α is expressed as

$$\alpha_{MH} = \frac{\sum_{j=1}^{K} A_j D_j / N_{..j}}{\sum_{j=1}^{K} B_j C_j / N_{..j}} \quad (3)$$

Holland and Thayer (1988) proposed a logarithmic transformation of α for interpretive purposes, with the aim of obtaining a symmetrical scale in which a zero value indicates an absence of DIF, a negative value indicates that the item favours the reference group over the focal group, and a positive value indicates DIF in the opposite direction. This transformation, delta metric, is expressed as

$$\Delta\alpha_{MH} = -2.35 \ln (\alpha_{MH}) \quad (4)$$

6. Results

The test results for data fit of the three models (A, B, and C) are summarised in Table 1.

Table 1. Data fit of three CFA models from Teacher Test Data

CFA Model	χ^2	*df*	CFI	TLI	WRMR	RMSEA	90% CI for RMSEA *Lower limit*	90% CI for RMSEA *Upper limit*
A: one factor	4851.829	2079	.923	.920	1.471	.014	.014	.015
B: 3 correlated factors	4752.691	2076	.926	.923	1.455	.014	.014	.015
C: 3 uncorrelated factors	No convergence							

The assessment of model fit is based on the evaluation of the following goodness-of-fit indices, with cutting scores for an excellent fit as follows:

- Comparative fit index: CFI > 0.95; Incremental Fit Index: IFI > 0.95;
- Standardised root mean square residual: SRMR = 0.00 (SRMR < 1.00 for an adequate fit);
- Root mean square error of approximation: RMSEA = 0.00 (RMSEA ≤ 0.05 for an adequate data fit (Hu & Bentler, 1999; Marsh, Wen, & Hau, 2004).

The results in Table 1 indicate that the one-factor model (model A) provides an adequate data fit. A very slight improvement in data fit is obtained with model B, where the correlations between the three domains of the test are taken into account. These correlations were found to be very high, ranging from 0.883 to 0.949 (see Table 2).

Table 2. Correlations among Teacher Test Domains

Domain	Domain 1	Domain 2	Domain 3
1: Computer & math, engineering and science	1.000		
2: Computer applications	0.909	1.000	
3: Computer & education	0.883	0.949	1.000

Data fit results in table 1 showed high correlations among the domains in models A and particularly B, therefore the teacher test data are essentially unidimensional. Model C, where the three test domains are assumed uncorrelated, does not converge with the test data.

The standardised item factor loadings and thresholds of the 66 items of the test under the one-factor CFA model (model A) are provided in the appendix. The analysis of the sample showed 60.5% were females and 39.5% males, which differed from the overall population. All factor loadings were statistically significant ($p < .001$), with the exception of the loading for item 45 ($p = .428$) and item 65 ($p = .340$).

The reliability of the data was estimated by a latent variable modelling (LVM) (equations 1 and 2). The reliability estimate was found to be 0.848 at a 95% confidence level = (0.842; 0.854). Cronbach's α was 0.749 and thus underestimated the LVM reliability (α < 0.848), as discussed above.

The data were tested for DIF across gender using the two Mantel-Haenszel statistics: α_{MH} and $\Delta\alpha_{MH}$ (equations 3 and 4), the results of which are provided in the appendix. For interpretation of these results, α_{MH} is reported with a z-statistic and its *p*-value, where DIF is signalled by statistically significant z-value ($p < .05$); with DIF against males if z >0 and DIF against females if z <0. The absolute values of the statistic $\Delta\alpha_{MH}$ are used to classify DIF into three categories: category A ─ negligible DIF when $|\Delta\alpha_{MH}| < 1.0$; category B ─ moderate DFI when $1 \leq |\Delta\alpha_{MH}| \leq 1.5$; and category C ─ large DIF, when $|\Delta\alpha_{MH}| > 1.5$ (Holland & Thayer, 1988).

Based on these criteria, the results in the appendix indicated that DIF is signalled for 13 items, of which 9 items fall in the category B for moderate DIF (6 against females and 3 against males) and 4 items in the category C for large DIF (2 against females and 2 against males). The remaining 43 items are either not signalled for DIF or were categorised as A, negligible DIF, acceptable for the purposes of this study (see Zwick & Ercikan, 1989).

7. Conclusion

This study examined the factor structure of the Computer Science Teacher Test and its psychometric characteristics to validate interpretations and decisions about certification of teachers in Saudi Arabia. The results showed that the test items are essentially unidimensional, confirming the use of item response modelling.

The results in this study support the validity of interpretations and decisions related to certification of teachers in Saudi Arabia based on their computer test scores. This outcome should guide test developers and researchers at the National Assessment Centre in further the evolution of the Computer Science Teacher Test.

References

Ahmed, M. (2016). The effects of Saudization on the universities: Localization in Saudi Arabia. *International Higher Education, 86*, 25-27.

Al Saud, F., & Alsadaawi, A. (2014, 25-30 May). *Raising the quality of education: Developing professional standards for Saudi teachers*. Paper presented to 40th Annual Conference of the International Association for Educational Assessment, Singapore.

Al Shannag, Q., Tairab, H., Dodeen, H., & Abdel-Fattah, F. (2013). Linking teachers' quality and student achievement in the Kingdom of Saudi Arabia and Singapore: The impact of teachers' background variables on student achievement. *Journal of Baltic Science Education, 12*(5), 652-655.

Al-Rasheed, M. (2010). *A history of Saudi Arabia*. Cambridge, Eng: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511993510

Alshahrani, M., & Alsadiq, M. (2014). *Economic growth and government spending in Saudi Arabia: An*

empirical investigation. Washington, DC: International Monetary Fund.

Alsharif, K. (2011). *Towards quality teacher education: Productive pedagogies as a framework for Saudi pre-service teachers' training in mathematics education* (PhD thesis, Curtin University, Perth, WA.).

Alzaydi, D. (2010). *Activity theory as a lens to explore participant perspectives of the administrative and academic activity systems in a university–school partnership in initial teacher education in Saudi Arabia* (Doctorate in Education thesis, University of Exeter, Exeter, Eng.).

Assessment Systems. (2016). *Xcalibre 4*. Retrieved July 6, 2016, from http://www.assess.com/xcalibre/

Blömeke, S., & Delaney, S. (2014). Assessment of teacher knowledge across countries: A review of the state of research. In S. Blömeke, F-J. Hsieh, G. Kaiser, & W. Schmidt (Eds.), *International perspectives on teacher knowledge, beliefs and opportunities to learn*. Dordrecht, Netherlands: Springer. http://dx.doi.org/10.1007/978-94-007-6437-8_25

Bollen, K. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*(3), 303-316. http://dx.doi.org/10.1177/0049124189017003004

Bowen, W. (2015). The history of Saudi Arabia (2nd ed.). Santa Barbara CA: Greenwood.

Brown, T. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford.

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. http://dx.doi.org/10.1037/0033-2909.105.3.456

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. http://dx.doi.org/10.1007/BF02310555

DeLuca, C., & Bellara, A. (2013). The current state of assessment education aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education, 64*(4) 356-372. http://dx.doi.org/10.1177/0022487113488144

Deng, N., Wells, C., & Hambleton, R. (2008, 23 August). *A confirmatory factor analytic study examining the dimensionality of educational achievement tests*. Paper delivered at the annual North-eastern Educational Research Conference, Rocky Hill, CT.

Dimitrov, D. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

Dimitrov, D., Al-Saud, F., & Alsadaawi, A. (2015). Investigating population heterogeneity and interaction effects of covariates The case of a large-scale assessment for teacher licensure in Saudi Arabia. Journal of Psychoeducational Assessment. http://dx.doi.org/10.1177/0734282914562121

Elyas, T., & Picard, M. (2013). Critiquing of higher education policy in Saudi Arabia: Towards a new neoliberalism. *Education, Business and Society: Contemporary Middle Eastern Issues, 6*(1), 31-41. http://dx.doi.org/10.1108/17537981311314709

Fleiss, J., Levin, B., & Paik, M. (2013). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley.

General Authority for Statistics. (2016). *Statistical year book 2015*. Retrieved July 2, 2016, from http://www.stats.gov.sa/en/413-0

Gómez-Benito, J., Hidalgo, M., & Zumbo, B. (2013). Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous items. *Educational and Psychological Measurement, 73*(5), 875-897. http://dx.doi.org/10.1177/0013164413492419

Guilera, G., Gómez-Benito, J., Hidalgo, M., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods, 18*(4), 553-571. http://dx.doi.org/10.1037/a0034306

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

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlebaum.

Holland, P., & Wainer, H. (1993). Preface. In P. Holland, & H. Wainer (Eds.), *Differential item functioning*. New York, NY: Routledge http://dx.doi.org/10.1017/cbo9780511622687.001

Hu, L. T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. http://dx.doi.org/10.1080/10705519909540118

International Bureau of Education. (2011). World data on education: Saudi Arabia. Retrieved July 3, 2016, from http://www.ibe.unesco.org/fileadmin/user_upload/Publications/WDE/2010/pdf-versions/Saudi_Arabia.pdf

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426 http://dx.doi.org/10.1007/BF02291366

Lodico, M., Spaulding, D., & Voegtle, K, (2010). Methods in educational research: From theory to practice (2nd ed.). San Francisco, CA: John Wiley/Jossey-Bass.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute, 22*(4), 719-748.

Marsh, H., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*(3), 275-300. http://dx.doi.org/10.1037/1082-989X.9.3.275

Ministry of Education. (2016). *Internal documentation*. Contact author.

Muthén, B. (2016). *Mplus information*. Retrieved July 6, 2016 from https://www.statmodel.com/index.shtml

Netemeyer, R., Bearden, W., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985772

Norton, B., & Syed, Z. (2003). TESOL in the Gulf: The sociocultural context of English language teaching in the Gulf. *TESOL Quarterly, 37*(2), 337-341. http://dx.doi.org/10.2307/3588508

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35*(2), 299-331. http://dx.doi.org/10.1016/S0005-7894(04)80041-8

Raykov, T. (2007). Evaluation of weighted scale reliability and criterion validity: A latent variable modeling approach. *Measurement and Evaluation in Counseling and Development, 40*(1), 42-52.

Raykov, T. (2012). *Scale reliability evaluation with LISREL 8.50*. Retrieved July 6, 2016 from http://www.ssicentral.com/lisrel/techdocs/reliabil.pdf

Raykov, T., Dimitrov, D., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling, 17*(2), 265-279. http://dx.doi.org/10.1080/10705511003659417

Raykov, T., Gabler, S., & Dimitrov, D. (2016). *Maximal criterion validity and scale criterion validity: A latent variable modeling approach for examining their difference*. Structural Equation Modeling: A Multidisciplinary Journal, In print. http://dx.doi.org/10.1080/10705511.2016.1155414

Saudi Aramco. (2016). *History*. Retrieved July 2, 2016 from http://www.saudiaramco.com/en/home/about/history/1950s.html

Sireci, S., & Rios, J. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187. http://dx.doi.org/10.1080/13803611.2013.767621

Strobl, C., Kopf, J., & Zeileis, A. (2010). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289-316. http://dx.doi.org/10.1007/s11336-013-9388-3

World Bank. (2016). Dataset for Saudi teacher characteristics, 2014. Accessed 2 July 2016 from http://databank.worldbank.org/data/reports.aspx?source=education-statistics-~-all-indicators

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*(1), 55-66. http://dx.doi.org/10.1111/j.1745-3984.1989.tb00318.x

Appendix

Testing for gender on teacher test items (Differential Item Functioning)

Item	α_{MH}	z-statistic	p-value	$\Delta\alpha_{MH}$	DIF against	DIF category
1	0.5911	6.9970	0.0000	1.2355	males	**B** (moderate)
2	1.1523	-1.3610	0.1740	-0.3332		No DIF
3	1.2330	-2.7190	0.0065	-0.4922		A (negligible)
4	0.9865	0.1230	0.9024	0.0321		No DIF
5	0.9349	0.7359	0.4618	0.1581		No DIF
6	0.7298	3.6550	0.0003	0.7401		A (negligible)
7	0.7343	3.3002	0.0010	0.7258		A (negligible)
8	1.1017	-1.2138	0.2248	-0.2275		A (negligible)
9	1.846	-6.4345	0.0000	-1.4406	females	**B** (moderate)
10	0.8479	2.2229	0.0262	0.3879		A (negligible)
11	0.9725	0.3688	0.7123	0.0655		No DIF
12	0.7039	4.7878	0.0000	0.8250		A (negligible)
13	1.2737	-2.6286	0.0086	-0.5685		A (negligible)
14	1.3073	-3.3718	0.0007	-0.6296		A (negligible)
15	0.6606	5.1835	0.0000	0.9742		A (negligible)
16	0.6828	5.0853	0.0000	0.8968		A (negligible)
17	0.8416	2.0918	0.03650	0.4051		A (negligible)
18	0.9132	1.0142	0.3105	0.2134		No DIF
19	0.7786	2.6756	0.0075	0.5880		A (negligible)
20	0.7304	3.6901	0.0002	0.7383		A (negligible)
21	0.5252	7.1830	0.0000	1.5133	males	**C** (large)
22	0.9829	0.2220	0.8243	0.0406		No DIF
23	1.0071	-0.0882	0.9297	-0.0165		No DIF
24	0.7974	2.6956	0.0070	0.5322		A (negligible)
25	1.1605	-1.8146	0.0696	-0.3499		No DIF
26	0.6043	6.5963	0.0000	1.1836	males	**B** (moderate)
27	0.7716	3.4417	0.0006	0.6093		A (negligible)
28	0.6076	5.6195	0.0000	1.1708	males	**B** (moderate)
29	0.8163	2.6602	0.0078	0.4769		A (negligible)
30	0.7745	3.1275	0.0018	0.6004		A (negligible)
31	0.8153	2.6957	0.007	0.4799		A (negligible)
32	1.3394	-2.2987	0.0215	-0.6867		A (negligible)
33	1.6394	-5.4650	0.0000	-1.1617	females	**B** (moderate)
34	1.0716	-0.7715	0.4404	-0.1624		No DIF
35	1.0406	-0.4418	0.6586	-0.0936		No DIF
36	1.6061	-4.6009	0.0000	-1.1135	females	**B** (moderate)
37	2.0133	-8.3547	0.0000	-1.6445	females	**C** (large)
38	1.5842	-5.2056	0.0000	-1.0811	females	**B** (moderate)
39	0.8059	1.9446	0.0518	0.5070		No DIF
40	0.6878	4.3058	0.0000	0.8796		A (negligible)
41	0.8392	2.2340	0.0255	0.4121		A (negligible)
42	1.2712	-3.0046	0.0027	-0.5639		A (negligible)
43	0.7265	4.3195	0.0000	0.7508		A (negligible)
44	1.5249	-5.2729	0.0000	-0.9916		A (negligible)
45	0.6903	3.2308	0.0012	0.8709		A (negligible)
46	0.7761	3.0779	0.0021	0.5958		A (negligible)
47	0.7401	3.8934	0.0001	0.7072		A (negligible)
48	0.9653	0.4565	0.6480	0.0830		No DIF
49	0.8060	2.9540	0.0031	0.5068		A (negligible)
50	1.4009	-4.3295	0.0000	-0.7922		A (negligible)
51	1.0747	-0.6942	0.4875	-0.1692		No DIF
52	1.6379	-4.2836	0.0000	-1.1596	females	**B** (moderate)

53	1.1705	-1.8934	0.0583	-0.3699		No DIF
54	1.0602	-0.7830	0.4336	-0.1373		No DIF
55	0.7934	3.1834	0.0015	0.5438		A (negligible)
56	0.8003	2.9867	0.0028	0.5234		A (negligible)
57	1.0537	-0.6966	0.4860	-0.1228		No DIF
58	1.0834	-1.0354	0.3005	-0.1883		No DIF
59	0.8656	1.9848	0.0472	0.3392		A (negligible)
60	0.9149	0.4148	0.6783	0.2090		No DIF
61	0.7785	3.4656	0.0005	0.5883		A (negligible)
62	0.5101	9.0260	0.0000	1.5820	males	**C** (large)
63	1.1264	-1.4038	0.1604	-0.2797		No DIF
64	1.6543	-6.0123	0.0000	-1.1830	females	**B** (moderate)
65	2.1570	-7.7439	0.0000	-1.8065	females	**C** (large)
66	1.0573	-0.5265	0.5986	-0.1308		No DIF

Note. The items in category A (negligible DIF) are considered to function for the purposes of this study (e.g., Zwick & Ercikan, 1989)

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).