Psychological Test and Assessment Modeling, Volume 58, 2016 (3), 471-483

Differential Item Functioning on mathematics items using multilevel SIBTEST

Brian F. French[1], *W. Holmes Finch*[2] &
Juan Antonio Valdivia Vazquez[3]

Abstract

In many testing contexts, data are collected using a multilevel sampling design, in which clusters of individuals are sampled, and individuals within clusters are administered the assessment. Multilevel structured data lead to correlated item responses for individuals within the same cluster, in turn leading to model parameter estimation bias. Specifically, standard errors are biased which leads to increased Type I error rates. This has been shown with DIF detection analysis as well. In this study, a new multilevel version of SIBTEST (MSIBTEST) is demonstrated for DIF assessment in a multilevel data context. This study investigates DIF in the Brigance Comprehensive Inventory of Basic Skills-II (CIBS-II) mathematics assessment between girls and boys. We focus on sex differences given previous sex DIF findings in mathematics achievement relating to item components such as format. More importantly, we demonstrate a recently developed method that accounts for multilevel data structures with dichotomous items for DIF. As hypothesized, adjusting DIF statistics for clustered data resulted in fewer items flagged for DIF compared to no adjustment.

Keywords: SIBTEST, MSIBTEST, gender differences, mathematics

[1] *Correspondence concerning this article should be addressed to:* Brian F. French, PhD, Department of Educational Leadership and Counseling Psychology, Cleveland Hall, Washington State University, Pullman, Washington, 99164, United States of America; email: frenchb@wsu.edu

[2] Ball State University

[3] Washington State University

Interest in the examination of performance differences on mathematics assessments related to sex is strong given the current assessment climate. Sex differences in mathematics assessments are being examined at national (e.g., Scholastic Aptitude Test; Rinn, McQueen, Clark & Rumsey, 2008) and at international (e.g., PISA-Mathematics; Liu, Wilson & Paek, 2008) levels. These differences are examined regardless of a student's grade level. Evaluations start as early as preschool (e.g. number exposure; Chang, Sandhofer, & Brown, 2011) with an aim to determining strengths and weaknesses of students and understanding why differences exist between groups, including differences between boys and girls.

Sex differences in mathematics scores have been associated with social (e.g. stereotypes; Cvencek & Meltzoff, 2012), cognitive (e.g., spatial reasoning; Klein, Adi-Japha & Hakak-Benizri, 2010), and affective (e.g., mathematics ability attribution; Dickhauser & Wulf-Uwe, 2006) aspects of mathematical learning. However, although documented, it is not clear what causes these sex differences in scores. Possible reasons include adult's expectations towards a child's mathematics performance (Ambady, Shih, Kim & Pittinsky, 2001; Penner & Paret, 2008). Such expectations can influence girls' and boys' mathematics achievement starting in the early stages of development (Keller, 2012). Parents, for instance, work on mathematical concepts with boys more often than with girls (Watt et al., 2012), and teacher beliefs reflect student's math-related attitudes (Gunderson, Ramirez, Levine & Beilock, 2012).

When examining sex differences in mathematical ability measured by standardized tests, outcomes can become ambiguous. Systematic reviews evaluating information from large datasets including international assessments created to measure mathematical performance (e.g., TIMMS, PISA) have shown sex differences in mathematics scores can be small (i.e., Cohen's *d* ranging from .05 to .15; Else-Quest, Hyde & Linn, 2010; Lindberg, Hyde, Petersen & Linn; 2010). In contrast, evidence also suggests the use of standardized assessments can perpetuate gender differences given the impact an item's format may have on gender responses. Item format (e.g., constructed-response), content (e.g. algebra), or objectives (e.g. evaluating "real life" mathematical problems) can favor boys and girls differently (Harris & Carlton, 1993; Taylor & Lee, 2012). Such results are supported when systematic reviews consider a test's format at an item level when examining outcomes (Zhang & French, 2010). Thus, sex differences in mathematics scores may, in part, be attributed to a lack of measurement invariance at the item level driven by such item components as item format.

At the item level, measurement invariance in mathematic tasks can be evaluated via differential item functioning (DIF) analysis. DIF refers to the situation where examinees from two groups (e.g., boys and girls) with equal levels on the measured trait (e.g., mathematics ability) have different probabilities of endorsing a specific level or response to an item (e.g., correct response). DIF can be found in the item difficulty and the item discrimination parameters, which are labeled as uniform DIF and non-uniform DIF, respectively. The former occurs when the probability of endorsing an item, conditioned on underlying ability, differs for members of the reference and focal groups, whereas the latter is present when group differences in the probability of endorsement conditioned on ability are not constant across the entire spectrum of abilities.

DIF methodology is frequently used to: evaluate sex differences across cultures (Lyons-Thomas, Sandilands & Ercikan, 2014); examine how these differences are developed through student's educational advancement (e.g. elementary and secondary schooling; Robinson & Lubieski, 2010); or understand how mathematics achievement gaps increase longitudinally (Carmichael, 2013). DIF methodology should be applied to the data in such a fashion that the statistical model matches the underlying data. One area of interest is with examining sex differences in mathematics assessments when data are collected from students nested in classrooms (i.e., multilevel data structures). This data collection structure is common yet often ignored in statistical analysis of differences.

Multilevel structured data can lead to correlated item responses for individuals within the same cluster (e.g., classroom; school), which can in turn lead to model parameter estimation bias, underestimation of standard errors for these estimates, and incorrect statistical results. These problems, particularly the underestimation of standard errors, can lead to increased Type I errors, which has been shown to be the case in DIF detection (e.g., French & Finch, 2013; French & Finch, 2015). Thus, in presence of clustered data, it becomes necessary to examine analytic models that can investigate sources of DIF, both at the person level (e.g. student characteristics) and at additional levels of nesting (e.g., classrooms, regional locations) (Albano & Rodriguez, 2013). At the very least, the statistical test should be adjusted to account for nested data. For this reason, to account for such data structures in DIF analysis, DIF methods are being adjusted to match the data structure and lead to more accurate results. One such method is a multilevel version (French & Finch; 2015) of SIBTEST (Shealy & Stout, 1993), a popular DIF procedure. There are additional methods that can be used to account for such data structures. Chen, Chen, and Shih (2013), for example, like others, demonstrate how hierarchical generalized models improve DIF detection, especially when mean differences are present. We focus on SIBTEST because this method was designed specifically for DIF and handles small samples and mean differences well.

SIBTEST was developed for DIF identification, and the version adjusted for MLM data in this study was designed for uniform DIF (Shealy & Stout, 1993). The SIBTEST statistic for uniform DIF is a weighted difference in the proportion of individuals in group 1 and group 2 answering an item correctly, conditioning on the ability that is measured by the assessment. The statistic is expressed as:

$$\hat{\beta} = \sum_{k=0}^{n} \hat{p}_x \left(\hat{P}_R \left[\hat{T}_x \right] - \hat{P}_F \left[\hat{T}_x \right] \right)$$

where

$\hat{p}_x$ = proportion of subjects with a matching subtest score of $X=x$

$\hat{P}_R \left[\hat{T}_x \right]$ = proportion of subjects in reference group with a matching subtest score of x answering the item correctly (1)

$\hat{P}_F \left[\hat{T}_x \right]$ = proportion of subjects in focal group with a matching subtest score of x answering the item correctly

As mentioned above, uniform DIF is present when the probability of a correct item response differs between these groups of examinees, controlling for or matching on the measured ability. For purposes of this matching, the subtest score, *X*, was originally proposed. Specifically, examinees would be matched on the observed scale score minus the item that is being assessed for DIF. For example, if item 5 is being examined for DIF, the matching subtest would include the remaining items on the test, minus item 5. The $\hat{\beta}$ statistic in (1) is the average difference in the probability of a correct item response between group 1 and group 2, conditioning on ability. A negative value of $\hat{\beta}$ connotes DIF in favor of group 2, while a positive value indicates DIF in favor of the group 1. The null hypothesis of no DIF (H_0: β=0) is tested using the ratio

$$\frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \tag{2}$$

where $\hat{\sigma}_{\hat{\beta}}$ is the estimated standard error of $\hat{\beta}$ (Shealy & Stout, 1993). If the null is rejected, then uniform DIF exists.

This study combines both examination of DIF in the mathematics items on large, standardized measures in samples of young children, and demonstrates a recently developed version of SIBTEST that accounts for multilevel data. See French & Finch, 2015 for details of the extension to MLM. To do so, the study employs data from the Brigance Comprehensive Inventory of Basic Skills-II (CIBS-II, French & Glascoe, 2010) mathematics assessment. Both content and methodological hypothesis were examined. Given there is a lack of evidence supporting sex differences in mathematics scores when kindergarten examinees are evaluated through standardized assessments (Robinson, & Lubienski, 2011), we hypothesized items from examined subtests would not exhibit DIF. On the other hand, if DIF items are found, DIF interpretation could be explained according to item content rather than format as documented with older students. Such a priori hypotheses follow recommendations for DIF assessment (Roussos & Stout, 2007). We hypothesized that the number of DIF items identified with SIBTEST would be higher compared to DIF items identified with multilevel SIBTEST (MSIBTEST). Support of this hypothesis would reflect the inflated Type I error rate due to downward biased standard errors resulting from ignoring the clustered or nested data. The analyses were conducted twice to be able to compare the performance of SIBTEST with and without accounting for the multilevel data. Thus, results are presented for standard SIBTEST as well as a multilevel version of SIBTEST (i.e., MSIBTEST).

Method

Sample

Item response data from 381 participants (49.86% male) from the CIBS II national standardization sample were employed. The standardization sample available to us had complete data on all items. Thus, all participants were included in the analysis. The sample was collected across 40 schools (clusters) in 27 states and matched closely the U.S.

population on a number of important demographic variables (e.g., age, race/ethnicity, geographic region), and is based on the U.S. Bureau of the Census projections for 2007 and the U.S. Department of Education's National Center for Education Statistics (Hussar & Bailey, 2006). Items available for study were only provided for children in kindergarten to be able to evaluate the items for DIF on the measure that is used to assess school readiness at this age. Students ranged in age from 5 to 7 years of age. Detailed information is available about the measure at http://www.curriculumassociates.com.

Instrument

Three mathematics subtests (Understands Quantitative Concepts, Count Objects, and Reads Numerals) of the CIBS II Readiness Assessments were utilized in this study (French & Glascoe, 2010). Understands Quantitative Concepts (15 items) assesses students' ability to correctly use in response to pictures and verbal prompts, such terms as big/little, or thick/thin; Count Objects (6 items) samples students' ability to count 3 to 24 pictures of objects; Reads Numerals (7 items) assesses students' ability to name numerals between 2 and 100 when presented out of order. All items were administered individually and examinees responded orally. Items are scored dichotomously as correct or incorrect. Internal consistency reliability estimates for the three subscale scores ranged from 0.65 to 0.87. Reliability and validity evidence supports the scores on the CIBS II assessment at the domain and total score levels. Please refer to the technical manual for details regarding reliability and validity information.

Analysis

Both standard SIBTEST and MSIBTEST were applied to the items within each subscale separately. We also only focused on uniform DIF. When examining each subtest, all items initially were included in the matching score or anchor set. Given that additional DIF items, beyond the studied item, can influence matching and DIF detection accuracy, the matching subtest score was purified, as recommended in the DIF literature (e.g., French & Maller, 2007). If an item displayed DIF on the initial identification step, it was removed from this score; the process was repeated until a subset of items was identified as not presenting DIF results. Then, observed DIF items were examined using the obtained purified matching subset. Standard SIBTEST was estimated using DIFPACK v.1.7 (William Stout Institute for Measurement, 2005) whereas MSIBTEST was estimated using SAS 9.3 (SAS Institute, 2010) once estimates were obtained from SIBTEST. DIF items are interpreted in terms of beta estimates, where categories A ($|\beta| < .059$), B ($.059 \leq |\beta| < .088$), and C ($|\beta| \geq .088$) represent small, median and large DIF, respectively ($\beta > .088$; Roussos & Stout, 1996), significance test results (p values) are reported as well. As we used both effect size criteria and effect size levels, no adjustment to the alpha level was used to try to control for inflation of Type I error due to multiple comparisons. We report exact p values if the reader wishes to draw a different conclusion. Given the applied nature of the article, we do not offer an in-depth explanation of

SIBTEST and the adjusted version. This description is presented elsewhere in the literature (French & Finch, 2015).

Results

The intraclass correlation coefficients (ICC) were computed for the three subtests and ranged from .09 to .21. This value reflects the amount of clustering that is present in the data. The larger the value, or closer to 1.0, indicates groups are more similar within themselves compared to other groups. A value of 0.0 would be the same as if individual students were sampled. Results presented in Table 1 show that standard SIBTEST identified 6 of 28 items (21.4%) exhibiting DIF. DIF items were distributed as follows: Count Objects (item 2 favoring boys, item 6 favoring girls), Read Numerals (item 2 favoring boys, 4 favoring girls), and Understands Quantitative Concepts (items 6 and 8, both items favoring girls). The results based on MSIBTEST revealed DIF items that were different in total number. In fact, only 4 items (14%) of the 28 items were identified. Thus, as expected, more items were identified when not adjusting for nested data by just over 7%. Our methodological hypothesis was supported.

As shown in Figure 1, DIF items found after controlling for multilevel data were distributed as Count Objects (items 2 favoring boys, and item 6 favoring girls), Read Numerals (item 2 favoring boys), and Understands Quantitative Concepts (item 8 favoring girls). Only item 6 in CO was a large DIF item favoring girls at the lower end of the ability distribution and then favoring boys at the higher end of the distribution, as seen in Figure 1, Panel B. Only item 8 in UQC was a large DIF item, where boys were favored at the low end of ability and there appears to be little difference at the high end of ability, as seen in Figure 1, Panel D. These differences are discussed below.

Discussion

The results from this study support the hypothesis focused on content, that in general, subtests do not contain a large amount of sex non-invariance at the item level at this young age in mathematics items. Although four DIF items were found, presence of DIF was distributed across subtests, minimizing influence these items would have on a score. And in fact, only two items had a large effect size worth attention for content and format review. These also appeared in separate subtests. This small number of DIF items would suggest it is unlikely to influence mean differences or result in differential prediction among boys and girls. This is likely, as well, given the ceiling rules that are in action in assessments such as this one. Thus, although it is possible to conclude that the item format does not promote differences in item performance between boys and girls in kindergarten, there is evidence that large magnitude of DIF can occur at this young age, even if for one or two items.

Table 1:
Results for Uniform DIF Analysis Using SIBTEST and MSIBTEST Methods

		SIBTEST			MSIBTEST		
Sub-test	Item	Beta estimate	Standard error	*p*-value	Beta estimate	Standard error	*p*-value
CO	1	0.000	0.000	1.000	0.000	0.000	1.000
	2¹	**0.085**	**0.032**	**0.007**	**0.073**	**0.027**	**0.006**
	3	0.069	0.043	0.112	0.070	0.049	0.149
	4	-0.011	0.070	0.876	0.004	0.072	0.959
	5	-0.049	0.095	0.602	0.054	0.094	0.566
	6²	**-0.208**	**0.096**	**0.030**	**-0.225**	**0.099**	**0.023**
RN	1	-0.011	0.010	0.295	-0.011	0.010	0.296
	2¹	**0.094**	**0.039**	**0.015**	**0.087**	**0.035**	**0.014**
	3	-0.048	0.061	0.430	-0.040	0.059	0.495
	4²	**-0.093**	**0.047**	**0.049**	-0.094	0.049	0.055
	5	0.006	0.047	0.902	0.004	0.049	0.940
	6	0.027	0.048	0.580	0.030	0.051	0.558
	7	0.087	0.049	0.074	0.093	0.051	0.069
UQC	1	0.004	0.019	0.847	0.009	0.020	0.639
	2	0.002	0.058	0.967	-0.014	0.059	0.816
	3	0.042	0.042	0.314	0.024	0.040	0.550
	4	0.000	0.000	0.000	0.000	0.000	1.000
	5	0.019	0.044	0.671	0.021	0.042	0.610
	6²	**-0.086**	**0.042**	**0.041**	-0.078	0.046	0.088
	7	-0.037	0.045	0.419	-0.028	0.044	0.530
	8²	**-0.169**	**0.050**	**0.001**	**-0.176**	**0.049**	**0.000**
	9	0.032	0.046	0.488	0.042	0.045	0.358
	10	0.076	0.057	0.182	0.064	0.059	0.277
	11	0.084	0.065	0.195	0.073	0.064	0.253
	12	-0.057	0.057	0.315	-0.040	0.056	0.473
	13	0.039	0.050	0.437	0.040	0.047	0.398
	14	0.001	0.043	0.985	0.013	0.040	0.751
	15	0.055	0.052	0.286	0.054	0.052	0.293

Notes:
1) CO=Counts Objects; RN=Reads Numerals, UQC=Understands Quantitative Concepts.
2) Bold= DIF item significant at p<.05
3) ¹= item favors boys, ²= item favors girls.
4) DIF categories: Small: | β | <.059; Medium: 059 ≤ | β | < .088; Large: | β | ≥ .088.

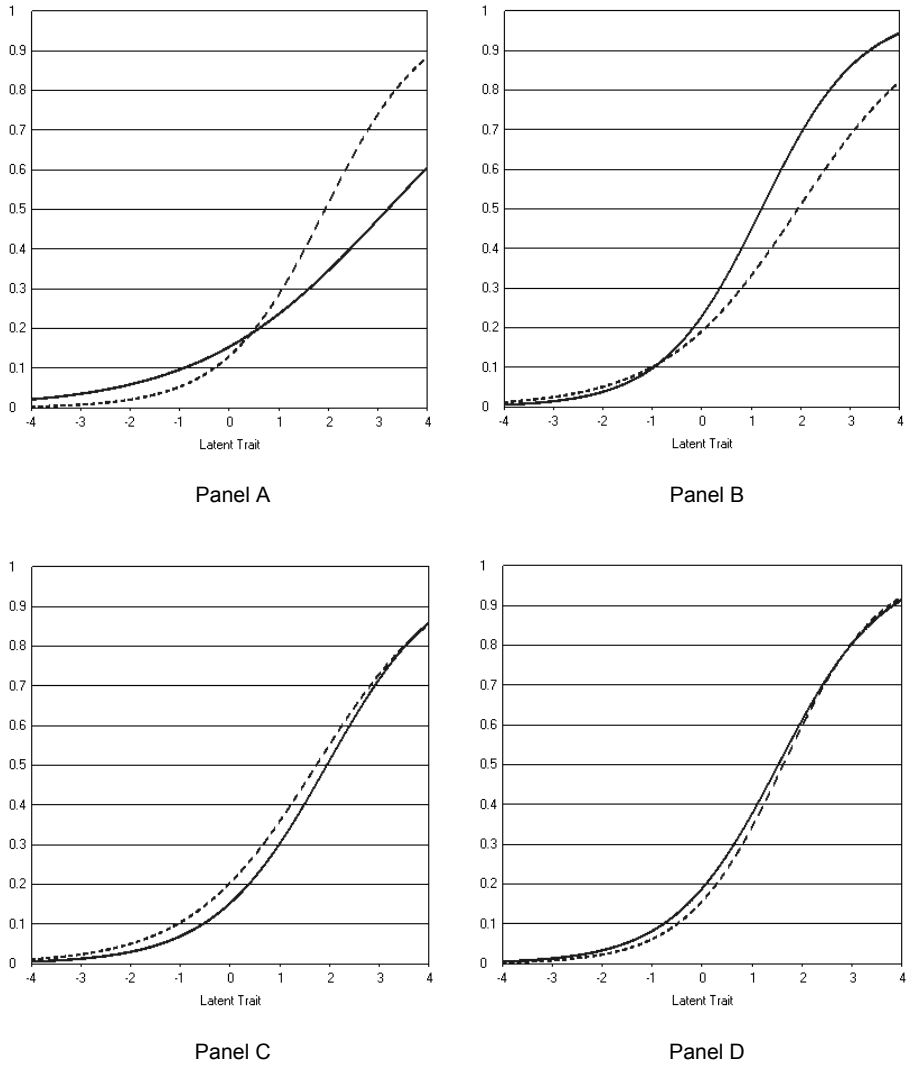

Figure 1:
Item Characteristic Curves for the Four DIF Items Across Subdomains

The dotted line represents girls and the solid line represents boys. Panel A: Counts objects = “Counts 6 faces”, Panel B: Counts objects = “Counts 24 stars”, Panel C = Reads numerals = “Reads 8, 7, 6, & 10”, Panel D: Understands Quantitative Concepts = "Understands long/short via responses to two questions: 1:I need a haircut because it is too ____; I can’t buckle the belt because it is too___”.

Our methodological hypothesis was also supported. That is, when the multilevel structure of the data was accounted for, there were fewer DIF items identified. That is, MSIBTEST flagged fewer items compared to SIBTEST. More importantly, the reduction that was most prevalent was with the subtest that had the highest ICC, as expected given the greater need for adjustment. This reduction in identification is important given that not accounting for clustered data would have led to a higher Type I error rate, a need for additional item review by a panel of experts, and cost more in resources in test development than necessary. The assessment design and evaluation processes are already expensive. If appropriate analytic techniques can be developed to match the data structure, as was accomplished here, artificial inflation of costs (time, money) can be controlled. These outcomes support the utility of MSIBTEST methodology to control Type I error. Thus, these results underscore the importance of the data structure and the statistical analysis being congruent. This same trend of good Type I error control with applied DIF analyses using different methods supports this conclusion (Finch & French, 2010) as does recent simulation work with MSIBTEST (French & Finch, 2015) across a variety of conditions. This method could be expanded from the educational environment where data are nested (patients in wards; workers in organization units) where there is a need to examine health, personality, or employment assessments for item level invariance.

In terms of content focus in mathematics, our findings are in accord with other work showing there can be a lack of sex differences in mathematics scores when standardized assessments are used to evaluate kindergarten children's mathematical ability (Robinson & Lubienski, 2011). Moreover, panel graphs in Figure 1 belonging to the same subtests (A and B / *Counts Objects*) show item behavior across boys and girls can be interpreted as affecting both boys and girls potentially in a non-uniform manner. Information from panels C and D present more typical uniform DIF findings whereas panels A and B, present findings that appear to be non-uniform DIF. Although we did not test for this type of DIF, graphically it appears this may be occurring and requires future investigation. However, when combining graphs from all panels and the content of the items, findings are also in agreement with cognitive research showing how mathematical skills are acquired at early developmental stages. Differences may be attributed to different cognitive strategies used to solve mathematical problems between boys and girls. Girls, for instance, use verbal skills more than boys (Klein, Adi-Japha & Hakak-Benizri, 2010), and boys are more familiar with mathematical concepts (Chang, Sandhofer & Brown, 2011) compared to girls not requiring verbal mediation.

Such cognitive processing issues certainly deserve further examination in controlled item development scenarios to inform future assessment development for mathematics. Extending DIF research with the use of think-aloud protocols or cognitive interviewing could assist in gaining a better understanding of the sources of DIF in such items. In turn, such in-depth information could be used to adjust or develop new mathematics items that control for identified components that may be causing DIF. This would be a step forward in not only building better assessments for more accurate comparisons across groups, but also in understanding the causes of DIF in such content areas.

On a methodological note, we only investigated uniform DIF. However, when Figure 1 is studied closely, it is clear that non-uniform DIF appears to be present. As has been

suggested before (Bolt & Gierl, 2006; Finch & French, 2008), researchers should combine the use of graphical tools (i.e., comparing graphs of ICCs) with DIF hypothesis testing to more accurately identify the nature of DIF. This combination would allow for a more thorough understanding of the nature or shape of DIF in a given item. Such a combination of statistical tests, effect sizes, and graphical analysis would not only lead to a better understanding of DIF but hopefully better test equity. Continued methodological development and work on combining the various approaches (i.e., context, cognitive processes, statistical, effect sizes, graphical considerations), for understanding and identifying DIF should aid the test development process and, in the end, aid the inference and action related to the decision about an individual.

In conclusion, developmental changes may better explain mathematical achievement at the early stages of learning by examining relationship between counting and estimation skills (Simms, Muldoon & Towse, 2013). Small DIF effects found in a few of the items in these early years of mathematics assessment may be attributable to the incipient symbolic numerical information children start to learn when entering the school system (Kolkman, Kroesbergen, & Leseman, 2013). This could be the starting point of the differences seen at later ages with assessment's item format (Zhang & French, 2010). Although incipient, results from this study give important information necessary to screening early mathematics achievement which is important to predict later achievement in mathematics (Gunderson, Ramirez, Levine & Beilock, 2012) and school outcomes (Cowan et al., 2011). Moreover, continuously improving the assessment development methodology and analysis used to enhance standard assessments can help overcome gender differences at early and formative school years by providing accurate information to training programs in mathematical thinking (Tzuriel & Egozi, 2010) as well as accurate assessments used in evaluation processes.

Author note

The multilevel DIF example was a component of the research supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110014 to Washington State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Albano, A.D., & Rodriguez, M.C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement, 73*(5), 836-856. doi: 10.1177/0013164413487375

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science,* 12 (5), 385-390.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement, 43,* 313-334.

Carmichael, C. (2013). Gender differences in children's mathematics achievement: Perspectives from the longitudinal study of Australian children. In V. Steinle, L. Ball & C. Bardini (Eds.) *Mathematics education: Yesterday, today and tomorrow (Proceedings of the 36th annual conference on the Mathematics Education Research Group of Australia*). Melbourne, VIC:MERGA.

Chang, A., Sandhofer, C.M. & Brown, C.S. (2011). Gender biases in early number exposure to preschool-aged children. *Journal of Language and Social Psychology, 30* (4), 440–450. doi 10.1177/0261927X11416207

Cowan, R., Donlan, C., Shepherd, D.-L., Cole-Fletcher, R., Saxton, M., & Hurry, J. (2011). Basic Calculation Proficiency and Mathematics Achievement in Elementary School Children. *Journal of Educational Psychology, 103* (4), 786-803. doi: http://dx.doi.org/10.1037/a0024556

Cvencek, D., & Meltzoff, A.N. (2012). Math-gender stereotypes in elementary school children. In J.A. Banks (Ed.), *Encyclopedia of diversity in education* (Vol. 3, pp. 1455-1460). Thousand Oaks, CA:Sage.

Devine, A., Fawcett, K., Szucs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions, 8* (33), 1-9. doi:10.1186/1744-9081-8-33

Dickhauser, O., & Wulf-Uwe, M. (2006). Gender differences in young children's math ability attributions. *Psychology Science, 48* (1), 3-16.

Else-Quest, N.M., Hyde, J.S., & Linn, M.C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103-127. doi: 10.1037/a0018053

Finch, W. H., & French, B. F. (2008) Anomalous Type I error rates for identifying one type of DIF in the presence of another, *Educational and Psychological Measurement, 68,* 742-759.

French, B.F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47,* 299-317.

French, B.F., & Finch, W.H. (2013). Extensions of Mantel–Haenszel for multilevel DIF detection. *Educational and Psychological Measurement, 73* (4), 648-671. doi: 10.1177/0013164412472341.

French, B. F., & Finch, W. H. (2015). SIBTEST in a multilevel data environment. *Journal of Educational Measurement*, *52,* 159-180.

French, B.F. & Glascoe F.P. (2010). *Comprehensive Inventory of Basic Skills II: Standardization and validation manual*. North Billerica, MA: Curriculum Associates.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for DIF detection. *Educational and Psychological Measurement, 67, 373-393.*

Gunderson, E.A., Ramirez, G., Levine, S.C., & Beilock, S.L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles,* 66, 153-166. doi 10.1007/s11199-011-9996-2

Harris, A.M. & Carlton, S.T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6* (2), 137-151. 3

Hussar, W. J., & Bailey, T. M. (2006). *Projections of education statistics to 2015 (NCES 2006-084*). U. S. Department of Education, National Center for Education Statistics. Washington, DC: U. S. Government Printing Office. Retrieved on November 7, 2006 from http://nces.ed.gov/pubs2006/2006084.pdf

Keller, J. (2012). Differential gender and ethnic differences in math performance a self-regulatory perspective. *Zeitschrift für Psychologie, 220* (3), 164-171. doi: 10.1027/2151-2604/a000109

Klein, P.S., Adi-Japha, E., & Hakak-Benizri, S. (2010). Mathematical thinking of kindergarten boys and girls: similar achievement, different contributing processes. *Educ Stud Math*, 73, 233–246. doi: 10.1007/s10649-009-9216-y

Kolkman, M.E., Kroesbergen, E.H., & Leseman, P.P.M. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction*, 25, 95-103. http://dx.doi.org/10.1016/j.learninstruc.2012.12.001

Lindberg, S.M., Hyde, J.S., Petersen, J.L., & Linn, M.C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135.doi: 10.1037/a0021276

Liu, O.L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch Analysis of gender differences in PISA Mathematics. *Journal of Applied Measurement, 9*(1), 18-35.

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in Mathematics in four international jurisdictions. *Educational and Science, 39*(172), 20-32.

Penner, A.M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research,* 37, 239–253. doi:10.1016/j.ssresearch.2007.06.012

Rinn, A.N., McQueen, K.S., Clark, G.L., & Rumsey, J.L. (2008). Gender differences in gifted adolescents' math/verbal self-concepts and math/verbal achievement: Implications for the STEM fields. *Journal for the Education of the Gifted, 32* (1), 34–53.

Robinson, J.P., & Lubienski, S.L. (2011). The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings. *American Educational Research Journal, 48* (2), 268–302. doi: 10.3102/0002831210372249

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33, 215-230.

SAS Institute Inc. (2010). *SAS 9.3 for Windows*. SAS Institute Inc., Cary, NC.

Sahin, H., French, B.F., Hand, B. & Gunel, M. (2015) Detection of differential item functioning in the Cornell Critical Thinking Test between Turkish and United States students. *European Journal of Psychological Assessment.*

Schmitt, N., Golubovich, J., & Leong, F. T. (2010). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using Big Five and RIASEC measures. *Assessment*, *18*, 412-427.

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

Taylor, C.S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25* (3), 246-280. doi: 10.1080/08957347.2012.687650

Tzuriel, D., & Egozi, G. (2010). Gender Differences in Spatial Ability of Young Children: The Effects of Training and Processing Strategies. *Child Development, 81* (5) , 1417–1430.

Watt, H. M. G., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., & Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: A comparison of samples from Australia, Canada, and the United States. *Developmental Psychology, 48*(6), 1594-1611. doi: 10.1037/a0027838

William Stout Institute for Measurement. (2005). *DIFPACK*. Assessment Systems Corporation, St. Paul, MN.

Zhang, M., & French, B. F. (2010, May). Gender related differential item functioning in mathematics tests: A meta-analysis. *Paper presented at the National Council on Measurement in Education conference,* Denver, CO.