Assessing Differential Item Functioning of the PISA 2018 Academic Resilience Scale.

Abstract

Differential Item Functioning is a common measurement issue that plagues measurement scales, rendering results from such scales biased as it yields an unfair advantage to a particular group. This study conducts DIF and DTF analyses on the PISA 2018 Academic Resilience Scale. Results indicate consistency between gender groups.

Introduction

Wang defines academic resilience as “the heightened likelihood of success in school and other life accomplishments, despite environmental adversities brought about by early traits, conditions and experiences”. Academic resilience and self-efficacy have also been found to be positively related. Again, academic resilience has been found to be positively related to higher level of on-task behavior.

One major component of an accurate measurement instrument is fairness. An accurate measurement instrument should not yield unfair advantage to one subpopulation. These instruments render the any conclusion drawn from the scores invalid. DIF is present when an instrument functions differently across subpopulations. Given the relevance of DIF-free scales, this study seeks to examine psychometric properties of the 5-item academic resilience scale used in PISA 2018. with a specific focus on both differential item functioning (DIF) and differential test functioning (DTF). With evidence on gender differences in academic resilience, the focus of the study will be on gender.

Method

Participants and Procedures.

The study utilized data from PISA 2018, with the focus on United states. A total of 4838 students took part in the exams. After deleting the missing observations, we ended up with 4548 observations. Among the participants, 49.9% were females and 50.1% were males. Table 1 shows the 5-item resilience scale measured on a four-point Likert scale (ranging from strongly disagree to strongly agree.

Data analysis

This research employed Item Response Theory as its analytical framework. Initially, a Graded Response Model was applied to both the reference and focal groups, aligning with the suggestion of Desjardins & Bulut. The female category was used as the reference group. As indicated by Tay, the selection of the reference group and focal group is arbitrary and does not impact DIF analyses. The DIF analysis was conducted by following the two-stage approach outlined by Meade. This method employs Likelihood Ratio Tests, comparing a more constrained (baseline) model with a less constrained (comparison) model. To quantify the effect sizes, various indices recommended by Meade were computed. At the item level, calculations included Signed Item Differences (SIDS), Unsigned Item Differences (UIDS), and Expected Score Standardized Differences (ESSD). On the scale level, the Signed Test Difference in the sample (STDS) and the Expected Test Score Standardized Difference (ETSSD) were computed.

Results

Graded Response Model

Table 2 displays the outcomes of the Graded Response Model. Although the chi-square test for both groups was statistically significant, it does not necessarily indicate a poor fit because the test is highly sensitive to sample size. Hence, we rely on the overall goodness of fit measures. For the reference group, the values are RMSEA = 0.08, SRMR = 0.04, and CFI = 0.98. Similarly, for the focal group, the values are RMSEA = 0.09, SRMR = 0.05, and CFI = 0.98. According to the recommended threshold by Hu and Bentler, these fit statistics align with the recommended threshold values.

DIF/DTF results

The results of the DIF analyses are presented in Table 3. The p-values were adjusted using the Benjamini-Hochberg's Procedure to reduce the chances of committing type-1 error. Items 1, 2, 4, and 5 displayed DIF in the first rounds of analyses. After, using item 3 as the anchor, only item 1, 2, and 5 displayed DIF. Figure 2 shows that the nature of DIF present for all the three DIF items is uniform since the item characteristic curves for the two groups never crossed. At the scale level, there is still evidence that the nature of DIF is uniform as can be observed from Figure 1 where there two curves do not cross.

Table 4 shows the item-level effect sizes. Items 1, 2, and 5 had both UIDS and SIDS .06, .13, and -.06 respectively. This indicates that, for item 1 and 2, females scored .064 and .128 points lower than males at any given theta level. However, for item 5, males scored .06 higher than females. Similarly, the ESSD for the DIF items (i.e., items 1, 2 and 5) were all less than .21, .34, and -.14 respectively, indicating a medium effect size according to Cohen's guidelines. Table 5 shows estimates of effect sizes at the scale-level. The STDS value of .012 indicates that on average females are expected to score .13 points higher on the summed scale than males at a given theta level. The ETSSD value was .07, indicating a small effect size. This value shows that females score .07 standard deviation units higher than males at any given ability level.

Conclusion

The study was conducted to find out whether the 5-item resilience scale utilized in PISA 2018 performs similarly between males and females. Although more than half of the items (i.e., items 1, 2, and 5) showed DIF, the associated effect sizes were medium. Again, the scale level effect sizes for both groups were very small indicating that the DIF did not accumulate to DTF. Hence, users of the resilience scale can confidently rely on it and make gender comparisons without worrying about the results getting confounded by DIF or DTF. In other words, any gender differences found using this scale can be seen as a real difference and not an artifact of DIF.

Reference

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Carlson, D. J. (2001). Development and validation of a college resilience questionnaire (Publication No. 3016308) [Doctoral dissertation, University of Nebraska]. ProQuest Dissertations and Theses Global.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Karami, H. (2012). An introduction to differential item functioning. The International Journal of Educational and Psychological Assessment.

Martin, A. J., & Marsh, H. W. (2006). Academic resilience and its psychological and educational correlates: A construct validity approach. Psychology in the Schools, 46(1), 53–83.

Meade, A.W. (2010). A taxonomy of effect size measures for differential function of items and scale. *Journal of Applied Psychology, 95*(4), 728-743.

Morales, E. (2008). Exceptional female students of color: Academic resilience and gender in higher education. Journal of Higher Education 33: 197-213.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. Harvard Educational Review, 80(1), 106–134.

Shehu, J., & Mokgwathi, M. (2008). Health locus of control and internal resilience factors among adolescents in Botswana: A case-control study with implications for physical education. South African Journal for Research in Sports, Physical Education and Recreation 30(2): 95-105.

Tay, L., Meade, A.W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3-46.

Wang, Haertal, G., & Walgberg, H. (1994). Educational resilience in inner-city. In M. Wang & G. E (Eds.), Educational resilience in inner-city America: Challenges and prospects (pp. 45–72). Hillsdale: Lawrence Erlbaunm Associates.

Waxman, H. C., Rivera, H., & Powers, R. (2012). English learners' educational resilience and classroom learning environment. Educational Research Quarterly, 35(4), 53–72.

Table 1

GRM Model fit indexes (one-factor model)

Numbers	Items
1.	I usually manage one way or another.
2.	I feel proud that I have accomplished things.
3.	I feel that I can handle many things at a time.
4.	My belief in myself gets me through hard times.
5.	When I'm in a difficult situation, I can usually find my way out of it.

Table 2

GRM Model fit indexes (one-factor model)

Group	*N*	*RMSEA*	*RMSEA* 90% CI	*SRMSR*	*CFI*
Females	2268	.08	[.065, .096]	.04	.98
Males	2280	.09	[.078, .110]	.05	.98

Table 3

Two-Stage Likelihood Ratio Test results for the groups

	All-others-as anchors model		Anchor-item model (2^{nd} round)	
Items	G^2	BH_P	G^2	BH_P
1. I usually manage one way or another.	24.91	< .001	29.23	< .001
2. I feel proud that I have accomplished things.	72.60	< .001	60.61	< .001
3. I feel that I can handle many things at a time.	7.84	.098	-	-
4. My belief in myself gets me through hard times.	15.95	.004	-	-
5.When I'm in a difficult situation, I can usually find my way out of it.	59.05	< .001	23.00	< .001

Table 4

Item-level effect sizes

Items	*SIDS*	*UIDS*	*ESSD*
1. I usually manage one way or another.	.064	.064	.209
2. I feel proud that I have accomplished things.	.128	.128	.342
3. I feel that I can handle many things at a time.	-	-	-
4. My belief in myself gets me through hard times.	-	-	-
5.When I'm in a difficult situation, I can usually find my way out of it.	-.058	-.059	-.138

Table 7

Scale-level effect sizes for all groups

	STDS	*ETSSD*
Gender	.134	.066

Figure 1

Scale-level expected scores

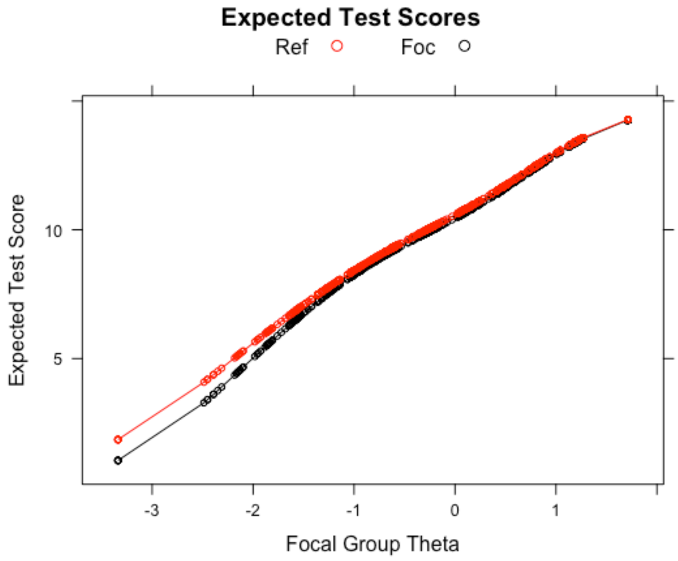
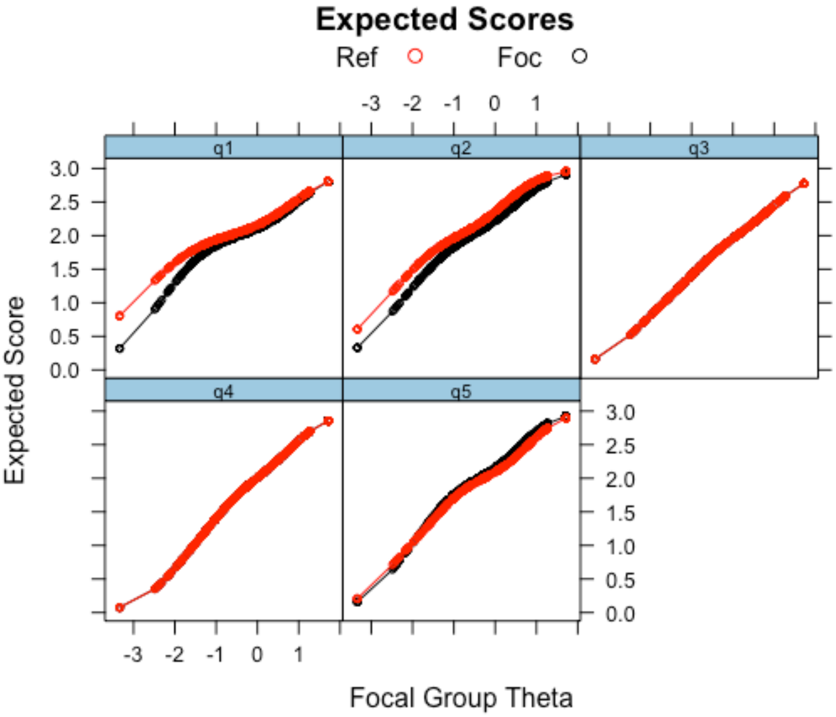


Figure 2

Item-level expected scores


Expected Scores
Ref ○
Foc ○
q1
q2
q3
q4
q5
Expected Score
Focal Group Theta