

A key assumption of any latent measure (any questionnaire trying to assess an unobservable construct) is that it functions equally across all different groups. In psychometrics this is called measurement invariance and means that members of different groups understand and respond to the scales similarly and that items have the same relationship with the latent measure across groups (Embretson and Reise, 2000). Having ascertained this, data users can confidently assert that differences between groups are actual differences unrelated to any measurement error.

Traditional methods of assessing these possible forms of bias might be through focus groups or cognitive interviews with different populations to assess their understanding of the question (Ouimet, Bunnage, Carini, Kuh, and Kennedy, 2004), by assessing differences in group data (e.g. missingness, means, relative group item-test correlations, or group reliability estimates) (see Glaser, Van Horn, Arthur, Hawkins, and Catalano, 2005; Ware, Kosinski, Gandek, Aaronson, Apolone, et al., 1998; Embretson and Reise, 2000), or by investigating factor structure across groups (Vandenberg and Lance, 2000). Analyses of these types have been conducted on NSSE and FSSE data (see the Psychometric Portfolios for NSSE and FSSE).

While these kinds of approaches are effective in identifying group differences, they do not effectively account for intra-subject disparities in the construct(s) of interest. Differential item functioning (DIF) occurs when *individuals from different groups with the same level of the latent measure have different propensities for responding to an item*. If individuals with the same level of the latent measure differ, this indicates that the item is not equivalent across groups which means the use of the item invites measurement bias. DIF can be either uniform, where an item is biased on behalf of one group across all levels of the latent trait, or non-uniform, where an item is biased on behalf of one group at certain levels of the latent trait but on the other group's behalf at different levels of the latent trait.

Methods

We follow the work of Choi, Gibbons, and Crane (2011) on looking at DIF in polytomous items. We employ an ordinal logistic framework for identifying DIF using the graded response model generated matching score. This approach is valuable because of its ability to identify both uniform and non-uniform DIF, as well as its relative superiority to other methods of identifying DIF (Narayanan & Swaminathan, 1994; Narayanan & Swaminathan, 1996; Rogers and Swaminathan, 1993). Detection of DIF is based on the following models where u_i is the ordinal response to the item, $P(u_i \geq k)$ is the cumulative probabilities that item response falls in category k or beyond, and a_k are the varying categorical intercepts (Choi, Gibbons, Crane, 2011):

Model 1: $\text{logit } P(u_i \geq k) = a_k + \beta_1 * \text{latent trait}$

Model 2: $\text{logit } P(u_i \geq k) = a_k + \beta_1 * \text{latent trait} + \beta_2 * \text{group}$

Model 3: $\text{logit } P(u_i \geq k) = a_k + \beta_1 * \text{latent trait} + \beta_2 * \text{group} + \beta_3 * \text{latent trait} * \text{group}$

There are a few different possible approaches for considering the magnitude of DIF in the ordinal logistic framework:

1. The first compares changes in log likelihood values between models to X^2 distribution with the appropriate degree of freedom (i.e. $df = 1$ for comparing models 1 and 2 or 2 and 3; $df = 2$ if comparing models 1 and 3).
2. The second computes a pseudo R^2 change between the models.
3. The final option considers the degree of change in the β_1 coefficient.

An examination of these different criteria found that the first and the last were particularly sensitive while the second was not. Accounting for these flagged items did result in significant differences for a handful of individuals, but did not change estimated means for the groups. This suggests that they may be overly sensitive, identifying DIF where it is irrelevant (see Crane, Gibbons, Ocepek-Welikson, Cook, Cella et al., 2007). Given that the primary use of FSSE results is at the group and institution level, we used the pseudo R^2 measure to identify DIF items that substantively impacted measurement. Previous research suggests that pseudo $R^2 < 0.035$ is negligible and pseudo $R^2 \geq 0.07$ is large DIF (Gelin & Zumbo, 2003). We conducted an iterative process to identify the level of pseudo R^2 change at which the analysis flagged items in the scale for DIF. Thus, the tables below show items flagged for DIF at much lower, non-significant levels of pseudo R^2 .

Data and Measures

For the purpose of this analysis, we took a random sample of 3,000 faculty respondents from the 2017 FSSE administration. Large sample sizes can lead to spurious identification of DIF, thus we used only a small portion of the dataset. We examined DIF based in 8 of the scales identified by previous, theory-driven confirmatory factor analysis.¹ Thus, the matching is based on the graded response model estimate of the latent trait for each of the scales. DIF is examined over the following characteristics: STEM v. Non-STEM, Upper v. Lower Division, Face to Face v. Online, Adjunct v. Non-Adjunct, Full-time v. Part-time, Man v. Woman, and White v. Non-white.

Results

The tables below are assessed by identifying pseudo R^2 values greater than or equal to 0.035. Items with DIF at least that large are items that significantly affect the equivalence of the scale between the groups. Additionally, charts with test characteristic curves (TCC) are provided to illustrate the impact of DIF on the scale between the groups of interest. The horizontal axis represents the latent trait distribution for each scale, and the vertical axis shows the expected total score on the scale. Scales with substantive DIF will show significant differences between the different groups' TCC. For each scale, we present the scale chart most subject to DIF. As a reminder, 95% of a distribution fall between -2 and 2, while 99% fall between -3 and 3.

Inspection of the tables and charts suggest that the analysis did not identify an item that substantively suffers from DIF. None of the items exceed the pseudo R^2 value of 0.035 threshold for moderate DIF. None of the charts indicate that the use of all of the items would make a substantive difference in comparing the groups. We thus conclude that the FSSE items are not threatened by the validity threat of measurement invariance.

¹ The 3-item Learning Strategies and Quantitative Reasoning scales could not be analyzed because of too few items.

References

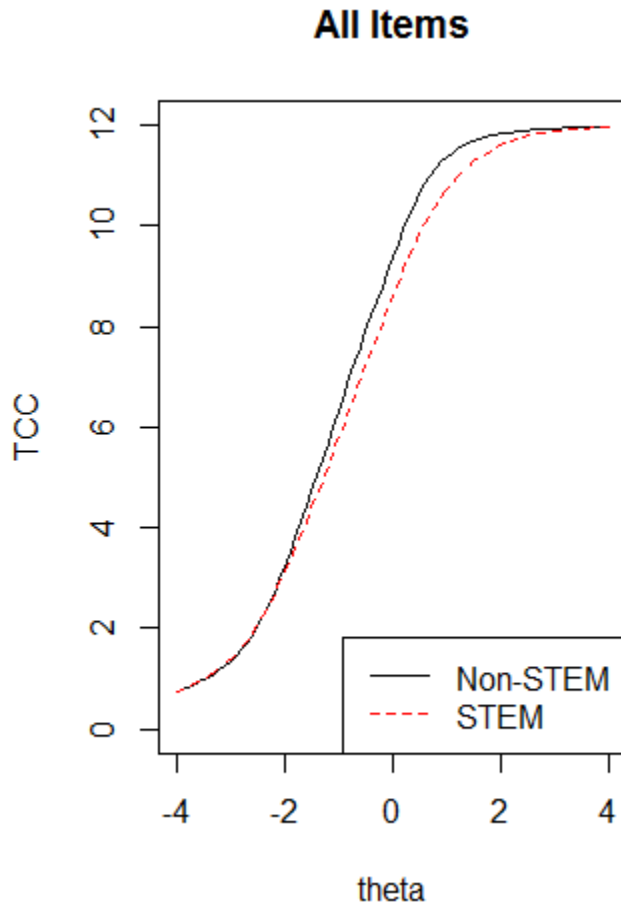
- Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of statistical software*, 39(8), 1.
- Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., ... & Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, 16(1), 69.
- Embretson, S. E., & Reise, S. P. Item response theory for psychologists. 2000. *Lawrence Erlbaum Associates, Mahwah, NJ*.
- Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement*, 63(1), 65-74.
- Glaser, R. R., Horn, M. L. V., Arthur, M. W., Hawkins, J. D., & Catalano, R. F. (2005). Measurement properties of the Communities That Care® Youth Survey across demographic groups. *Journal of Quantitative Criminology*, 21(1), 73-102.
- Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, 18(4), 315-328.
- Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20(3), 257-274.
- Quimet, J. A., Bunnage, J. C., Carini, R. M., Kuh, G. D., & Kennedy, J. (2004). Using focus groups, expert advice, and cognitive interviews to establish the validity of a college student survey. *Research in Higher Education*, 45(3), 233-250.
- Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17(2), 105-116.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1), 4-70.
- Ware, J. E., Kosinski, M., Gandek, B., Aaronson, N. K., Apolone, G., Bech, P., ... & Prieto, L. (1998). The factor structure of the SF-36 Health Survey in 10 countries: Results from the IQOLA Project. *Journal of clinical epidemiology*, 51(11), 1159-1165.

Higher-Order Learning

Table 1. DIF in the Higher-Order Learning Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fHOapply		0.001	0.001	0.001	0.001		0.001
fHOanalyze		0.001				0.001	
fHOevaluate	0.02		0.001	0.001		0.001	
fHOform		0.001					

Figure 1: Impact of Higher-Order Learning STEM Group DIF on Expected Total Score

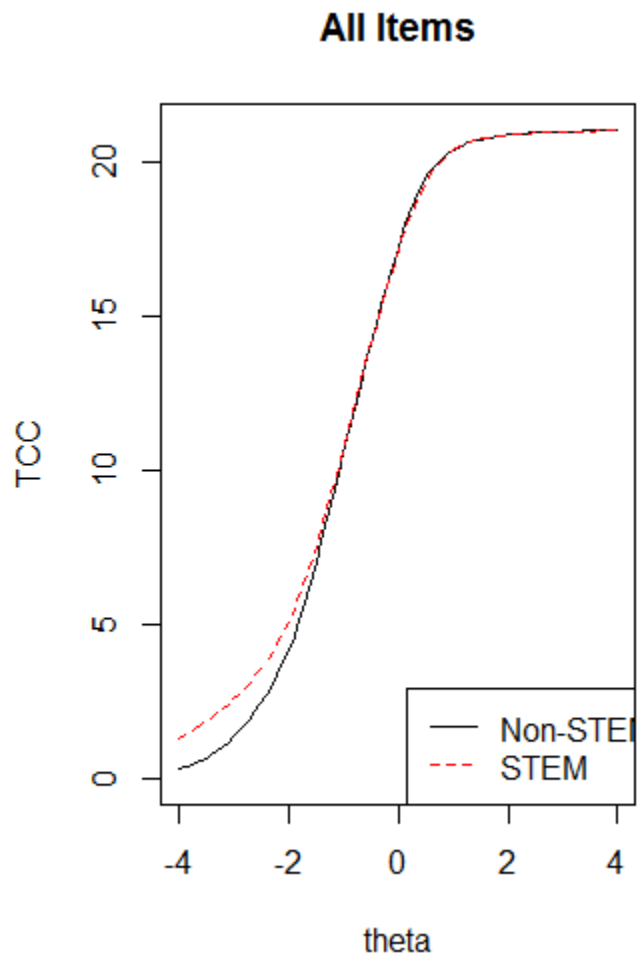


Reflective & Integrative Learning

Table 2. DIF in the Reflective & Integrative Learning Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fRIintegrate	0.01	0.01		0.001	0.001		0.001
fRIsocietal			0.001	0.001	0.001	0.001	0.001
fRIdiverse	0.01		0.001			0.001	0.001
fRIlowview			0.001				
fRIspect			0.001	0.001	0.001		
fRInewview					0.001		0.001
fRIconnect	0.01			0.001	0.001	0.001	0.001

Figure 2: Impact of Reflective & Integrative Learning STEM Group DIF on Expected Total Score

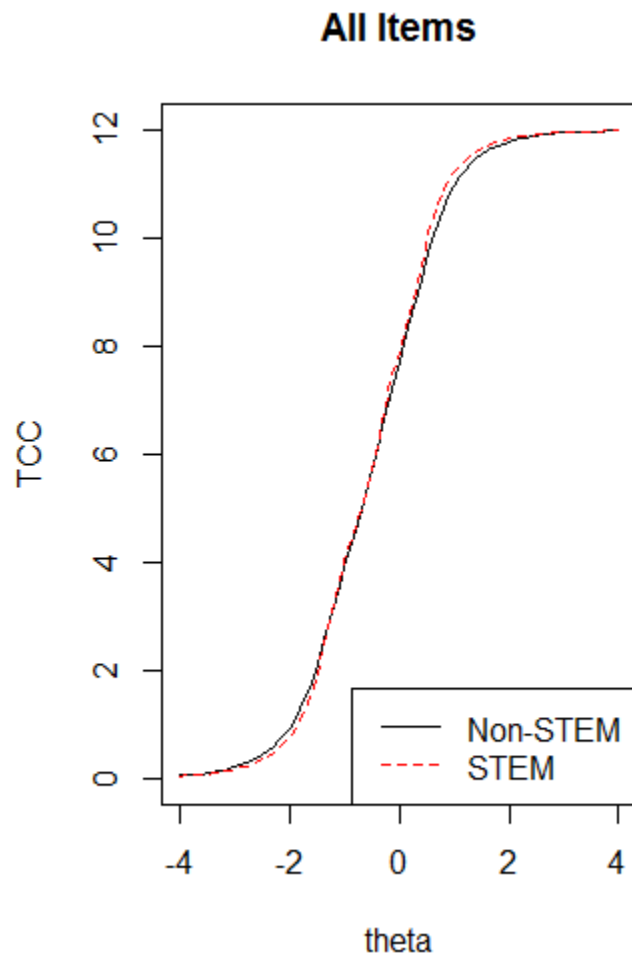


Collaborative Learning

Table 3. DIF in the Collaborative Learning Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fCLaskhelp		0.001					
fCLexplain	0.0007		0.003	0.001			0.0005
fCLstudy	0.0007				0.0007		
fCLproject		0.001	0.003			0.001	

Figure 3: Impact of Collaborative Learning STEM Group DIF on Expected Total Score

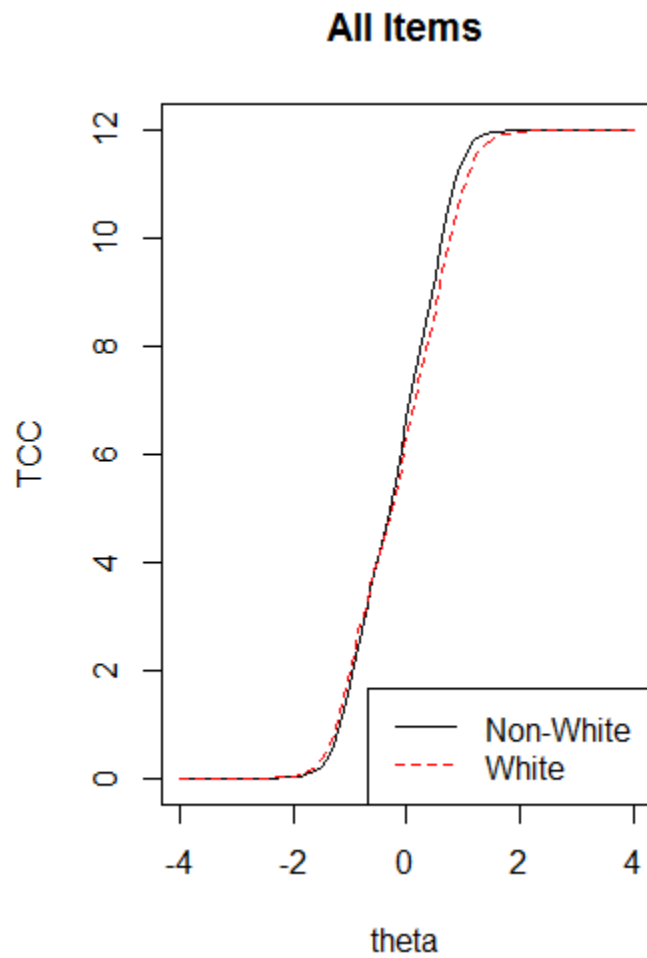


Discussions with Diverse Others

Table 4. DIF in the Discussions with Diverse Others Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fDDrace	.001	.0005		.0005	.0007		.0005
fDDeconomic							
fDDreligion				.0005	.0007		.0005
fDDpolitical			.0003	.0005	.0007	.001	.0005

Figure 4: Impact of Discussions with Diverse Others Race Group DIF on Expected Total Score

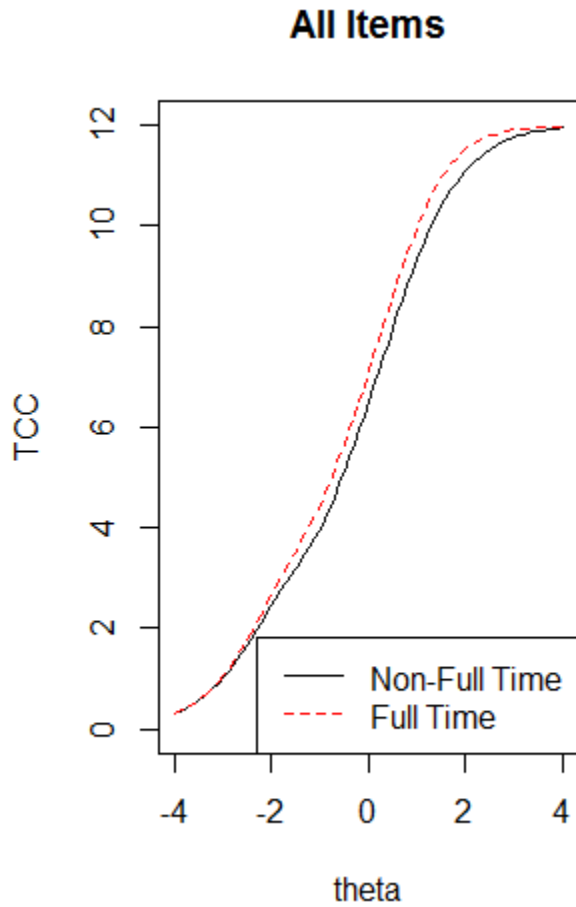


Student-Faculty Interaction

Table 5. DIF in the Student-Faculty Interaction Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fSFcareer	.001	.005					
fSFotherwork				.01	.02	.001	
fSFdiscuss	.001						
fSFperform	.001		.005	.01		.001	.001

Figure 5: Impact of Student-Faculty Interaction Full-Time Group DIF on Expected Total Score

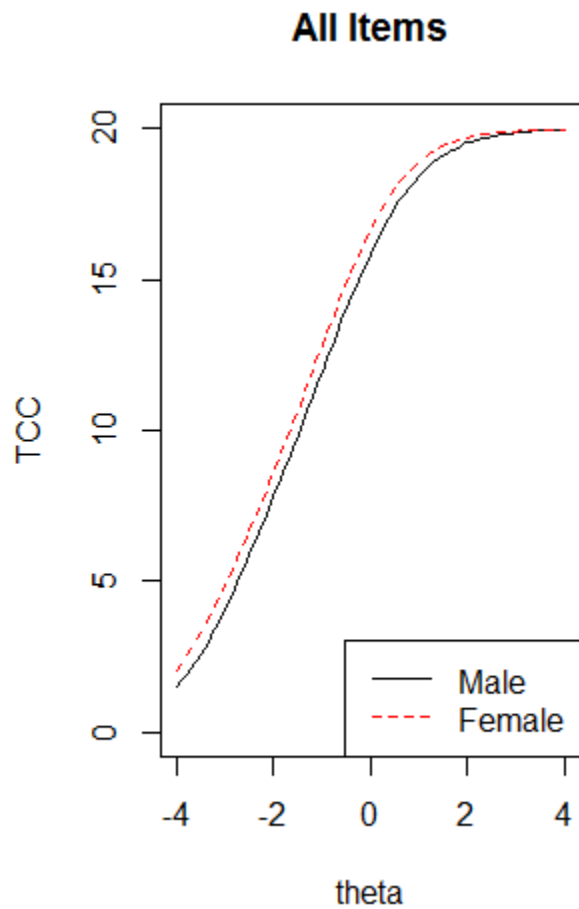


Effective Teaching Practices

Table 6. DIF in the Effective Teaching Practices Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fETgoals							
fETorganize	.001	.0001					
fETexample	.001	.0001	.01	.01	.01		
fETvariety	.001	.0001				.005	.001
fETreview		.0001					
fETstandards	.001	.0001				.005	
fETdraftfb	.001					.005	
fETfeedback		.0001					.001

Figure 6: Impact of Effective Teaching Practice Gender Group DIF on Expected Total Score

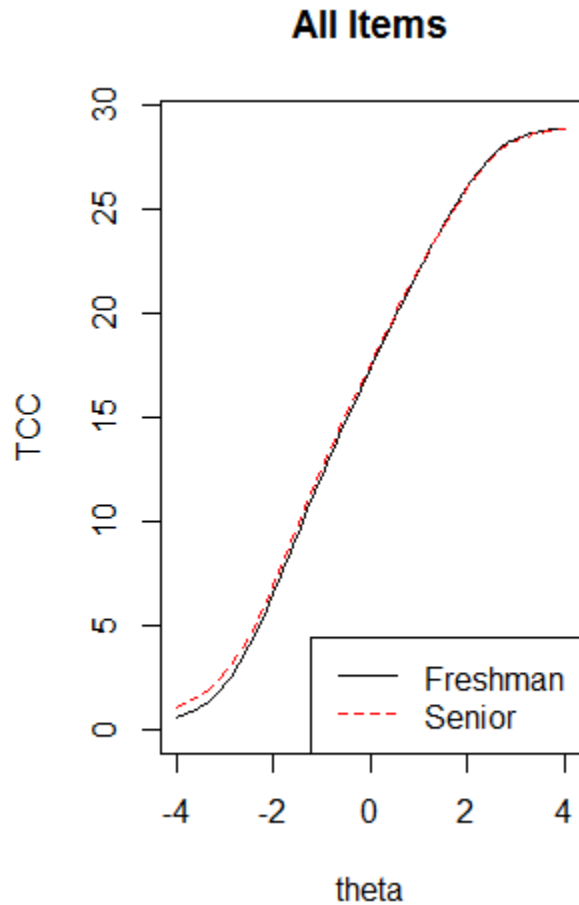


Quality of Interactions

Table 7. DIF in the Quality of Interaction Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fQIstudent	.0001	.001	.001		.001	.001	.001
fQIadvisor					.001		
fQIfaculty	.0001	.001		.001			
fQIstaff	.0001	.001					
fQIadmin			.001	.001	.001		

Figure 7: Impact Quality of Interaction Division Group DIF on Expected Total Score



Supportive Environment

Table 8. DIF in the Supportive Environment Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fSEacademic	.001		.001	.001		.001	.001
fSElearnsup		.001	.001	.001	.001	.001	.001
fSEdiverse	.001					.001	
fSEsocial	.001		.001	.001	.001	.001	.001
fSEwellness	.001	.001	.001	.001	.001	.001	
fSEnonacad		.001		.001		.001	.001
fSEactivities	.001			.001	.001		.001
fSEevents	.001				.001	.001	

Figure 8: Impact of Supportive Environment Gender Group DIF on Expected Total Score

