Employing differential item function analysis in survey research

Justin Paulsen
Ryan T Merckle
Allison BrckaLorenz

Indiana University Bloomington

Paper presented at the 2019 American Educational Research Association Annual Meeting, Toronto, Canada

Abstract

One of the key assumptions involved in using any survey or measurement is that the instrument works consistently across groups. This is particularly important in survey research where group comparisons are ubiquitous. Differential item functioning (DIF) analysis examines whether the instrument systematically biases in favor of one group. The findings from such an analysis are unattainable in traditional approaches to examining instrument validity, and yet, it is rare to find DIF analysis in surveys. This process illustrates DIF analysis with logistic regression using the Faculty Survey of Student Engagement. We find FSSE items did not show the presence of DIF. This provides confidence to users of this instrument that it measures the same constructs in the same way across different groups.

Employing differential item function analysis in survey research

A key assumption of any questionnaire or instrument trying to assess an unobservable construct is that it functions equally across all different groups. In psychometrics, this is called measurement invariance. Measurement invariance 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.

Assessing differential item functioning (DIF) is an important approach for determining whether an instrument is biased against particular groups (AERA, APA, NCME, 2014). Despite this, researchers outside of the testing industry rarely use DIF. A review of the literature found few instances of survey instruments that assess DIF as part of its validity evidence. This study illustrates the potential use of DIF in the context of a large scale, multidimensional survey instrument, the Faculty Survey of Student Engagement (FSSE). In this paper, we define what DIF is, how to assess it, and how to interpret the findings from it.

Perspectives

Traditionally, researchers assess surveys for forms of bias in the following ways:

- Conducting focus groups or cognitive interviews with different populations to assess their understanding of the question (Ouimet, Bunnage, Carini, Kuh, and Kennedy, 2004),
- 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
- Investigating factor structure across groups (Vandenberg and Lance, 2000).

Researchers have conducted analyses of these types FSSE data (FSSE, n.d.) 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 from different groups with the same level of the latent construct differ in their probability to answer the item correctly, this indicates that the item is not equivalent across groups. DIF means the use of the item invites measurement bias between those groups. 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 estimate an individual's score on the latent variable to be used for matching across groups by using the graded response model (GRM). The GRM effectively deals with ordered response data while offering the virtues of item response theory modeling (e.g., sample invariant estimates). We employ an ordinal logistic framework for identifying DIF in items. 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: $logit\ P(u_i \geq k) = a_k + \beta_1 * latent\ trait$

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

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

The logistic regression approach identifies DIF by comparing the models. The addition of a group variable to a model including the latent trait would only result in a statistically significant coefficient if the group variable predicted different probabilities of a correct response after controlling for the latent trait. Similarly, if β_3 in Model 3 is significant, then the impact of the grouping variable differs at different levels of the latent trait, indicating non-uniform DIF.

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

1. The first examines changes in log likelihood values between models relative 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 potentially too sensitive to DIF. Items flagged for DIF by these methods resulted in significant differences for a handful of individuals, but did not change estimated means for the groups. This suggests that these methods 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 at these levels. 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 (see Appendix) show items flagged for DIF at much lower, non-significant levels of pseudo R^2.

Data

For the purpose of this analysis, we took a random sample of 3,000 faculty respondents from the 2017 FSSE administration. FSSE measures faculty members' expectations of student engagement in educational practices that are empirically linked with high levels of undergraduate learning and development at four-year colleges and universities. We selected the FSSE dataset because of its large sample, the number of separate scales it includes, and the number of covariates along which DIF can be assessed. 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 eight of the scales identified by previous, theory-driven confirmatory factor analysis.[1]

- Higher-Order Learning (HOL) – Coursework emphasis on higher-order learning activities
- Reflective & Integrative Learning (RIL) – Importance of students participating in reflection and integrative learning
- Learning Strategies (LS) – How much faculty encourage the use of learning strategies
- Quantitative Reasoning (QR) – Importance of students participating in quantitative reasoning
- Collaborative Learning (CL) – How much faculty encourage students to work together
- Diverse Discussions with Others (DD) – How much opportunity students have to engage with people who are different from them
- Student-Faculty Interaction (SFI) – How frequently faculty interact with students outside of class
- Effective Teaching Practices (ET) – Faculty perceptions of their clarity and use of good pedagogical practices
- Quality of Interactions (QI) – Faculty perceptions of students' interactions with others
- Supportive Environment (SE) – Importance of increasing institutional support for students

The FSSE uses Likert-type scales with four to five response options. This requires that matching is based on the graded response model estimate of the latent trait for each of the scales. A full description of the

[1] We could not analyze the 3-item Learning Strategies and Quantitative Reasoning scales because of too few items.

model is outside the context of this paper, but Samejima (2016) provides valuable instruction.

We then examined DIF over the following characteristics: STEM v. Non-STEM disciplinary fields, upper- v. lower-division courses taught, face-to-face v. online course instruction, adjunct v. non-adjunct faculty rank, full-time v. part-time faculty employment, man v. woman, and White v. non-White. We chose these for the perception that these categories or characteristics may differentiate faculty.

Results

The tables (see Appendix) present pseudo R^2 values for each item. Items with DIF at least as large as $R^2 \geq 0.035$ are items that may affect the equivalence of the scale between the groups. Additionally, we provide charts with test characteristic curves (TCC) 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 exceeds the pseudo R^2 value of 0.035 threshold for moderate DIF. The figures in the appendix uniformly indicate that the use of all of the items in a given scale would not make a substantive difference in comparing the groups. We thus conclude that the FSSE items are not in danger from the validity threat of measurement invariance.

Scholarly Significance

One of the key assumptions involved in using any survey or measurement is that the instrument works consistently across all groups. This is particularly important in survey research where comparing across groups is a common analytical approach. DIF analysis examines whether the instrument systematically biases in favor of one group relative to another. The findings from such a DIF analysis are unattainable in traditional approaches to examining instrument validity, and yet, it is rare to find surveys that have undergone such analysis.

This paper presents one approach to DIF analysis using logistic regression. Logistic regression's general familiarity among researchers makes it a more accessible approach to identifying DIF. In addition, using the logistic regression method allows for readily available graphing functions for assessment of DIF in the item and its impact on the scale. As demonstrated in the illustrated example with FSSE, researchers can conduct this process for any instrument where items are grouped into scales. Previous literature has identified thresholds at which point DIF significantly impacts the comparison between groups. As noted above, across a variety of different groupings, FSSE items did not show the presence of DIF. This provides confidence to users of this instrument that it measures the same constructs in the same way across different groups. Researchers using surveys and other measurement tools should more broadly adopt this approach.

References

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

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. (2000). *Item response theory for psychologists*. Lawrence Earlbaum Associates, Mahwah, NJ.

Faculty Survey of Student Engagement. (n.d.) *FSSE Psychometric Portfolio*. Retrieved from fsse.indiana.edu.

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.

Ouimet, 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.

Samejima, F. (2016). Graded response models. In *Handbook of Item Response Theory, Volume One* (pp. 123-136). Chapman and Hall/CRC.

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.

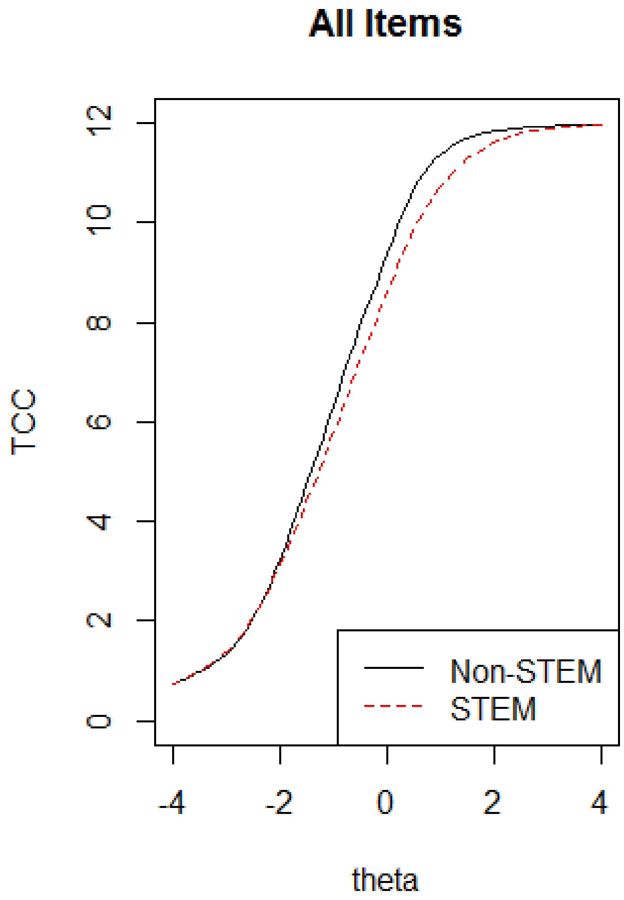
Appendix

Higher-Order Learning

Table 1. DIF in the Higher-Order Learning Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fHOapply		.001	.001	.001	.001		.001
fHOanalyze		.001				.001	
fHOevaluate	.02		.001	.001		.001	
fHOform		.001					

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

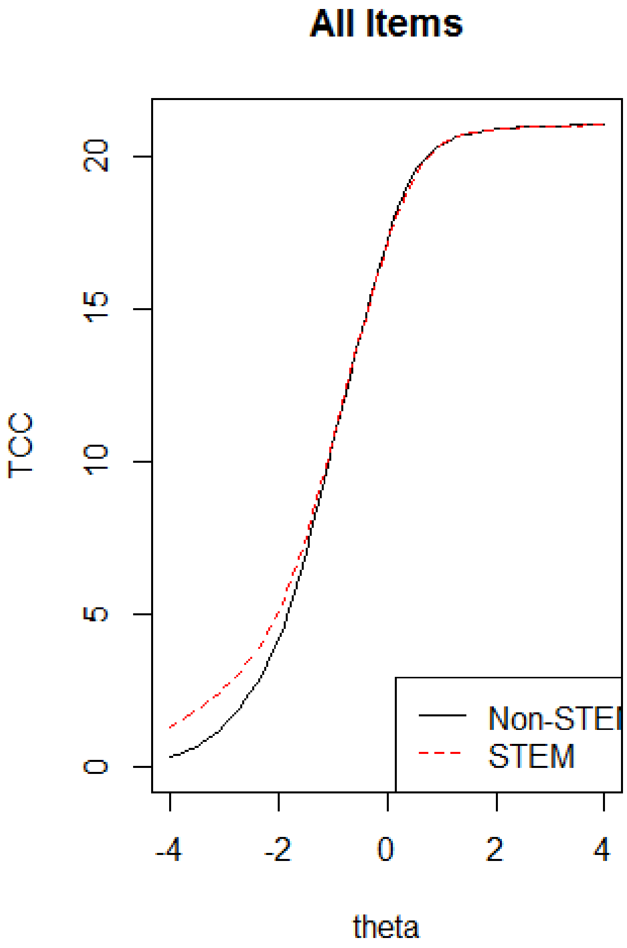


Reflective & Integrative Learning

Table 2. DIF in the Reflective & Integrative Learning Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fRIintegrate	.01	.01		.001	.001		.001
fRIsocietal			.001	.001	.001	.001	.001
fRIdiverse	.01		.001			.001	.001
fRIownview			.001				
fRIperspect			.001	.001	.001		
fRInewview					.001		.001
fRIconnect	.01			.001	.001	.001	.001

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

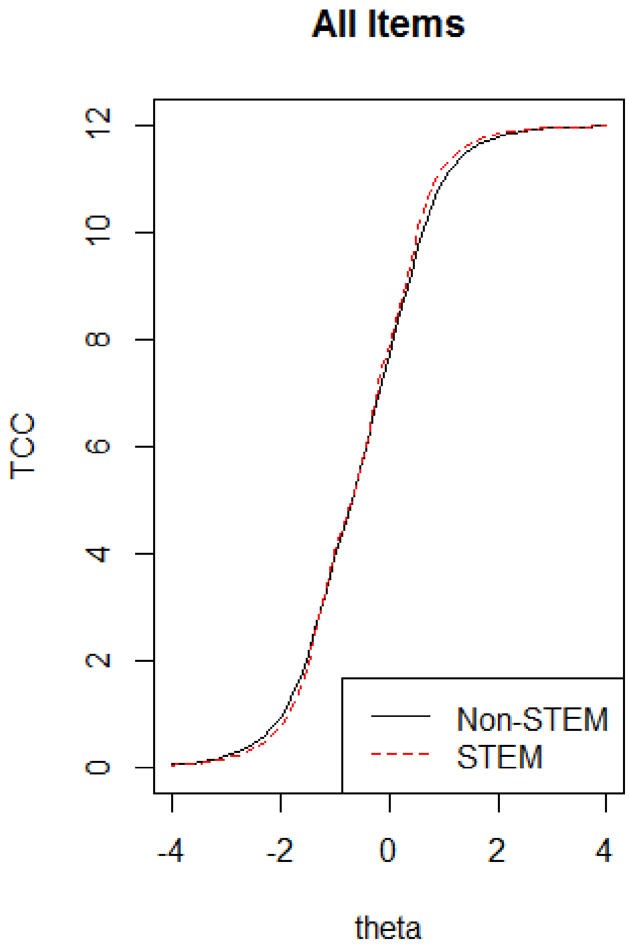


Collaborative Learning

Table 3. DIF in the Collaborative Learning Scale

	STEM	Division	Online	Adjunct	Full-Time	Gender	White
fCLaskhelp		.001					
fCLexplain	.0007		.003	.001			.0005
fCLstudy	.0007				.0007		
fCLproject		.001	.003			.001	

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

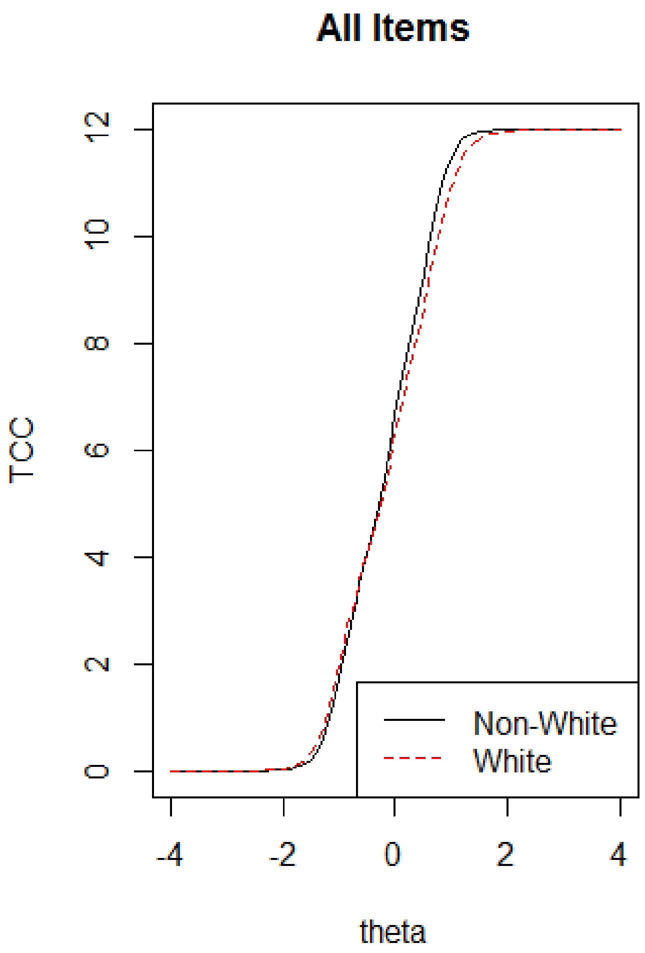


Discussions with Diverse Others

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

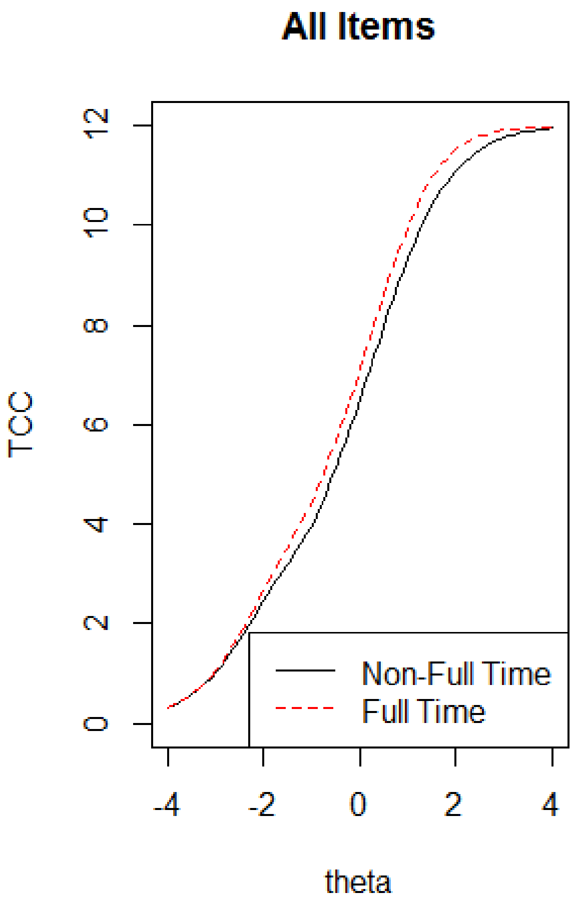


Student-Faculty Interaction

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

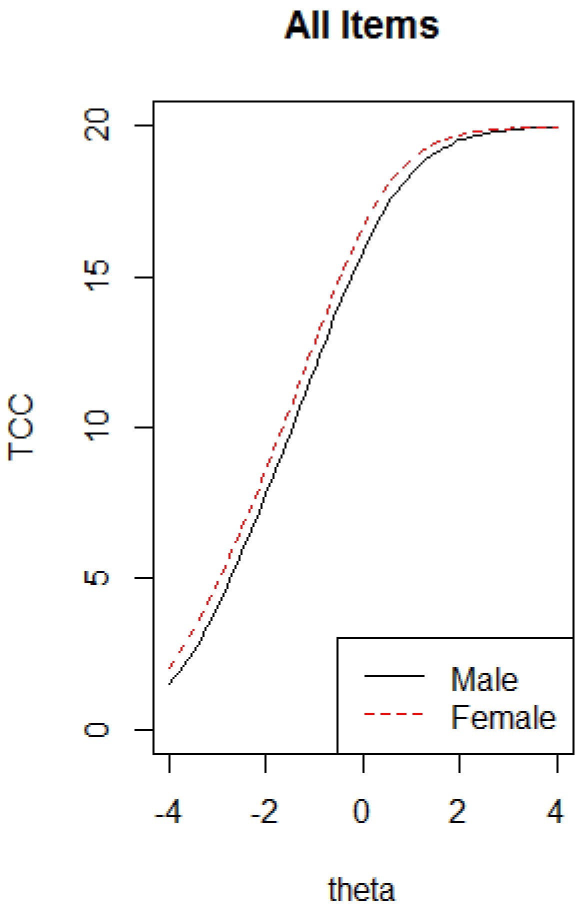


Effective Teaching Practices

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

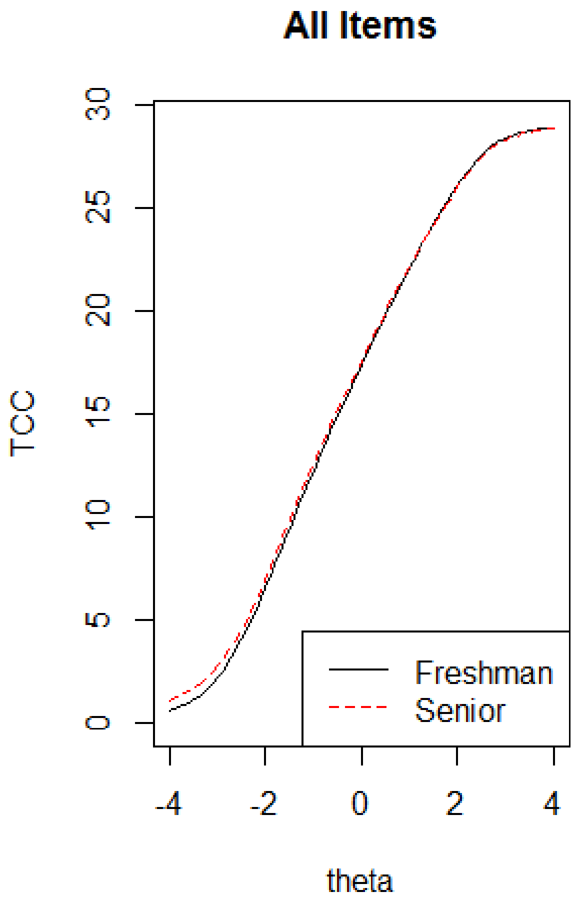


Quality of Interaction

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

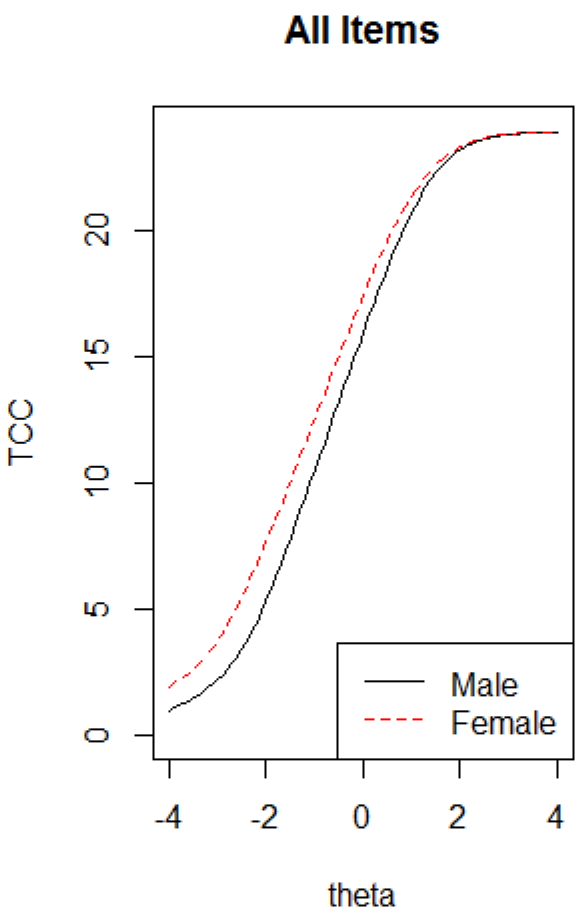


Supportive Environment

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


All Items
TCC
theta
Male
Female