2019 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
SCIENCE, TECHNOLOGY & ENGINEERING, ARTS, MATHEMATICS & EDUCATION JUNE 5 - 7, 2019
HAWAII PRINCE HOTEL WAIKIKI, HONOLULU, HAWAII

DETECTING DIFFERENTIAL ITEM FUNCTIONING IN A LEARNING ANALYTIC FEEDBACK SATISFACTION SURVEY

SCHMIDT, MATTHEW ET AL
COLLEGE OF EDUCATION
UNIVERSITY OF SASKATCHEWAN
SASKATCHEWAN, CANADA

Mr. Matthew Schmidt
Dr. Amin Mousavi
Dr. Vicki Squires
Dr. Kenneth Wilson
College of Education
University of Saskatchewan
Saskatchewan, Canada

Detecting Differential Item Functioning in a Learning Analytic Feedback Satisfaction Survey

Synopsis:

This paper assessed a feedback satisfaction survey for the presence of differential item functioning (DIF) in the evaluation of a learning analytic (LA) feedback intervention. A select few items were flagged for DIF with respect to student type and gender; however, these effects were of negligible size. Overall, it was found that the feedback satisfaction survey was measuring pre-identified demographic groups equally well.

Detecting Differential Item Functioning in a Feedback Satisfaction Survey

Matthew Schmidt, Amin Mousavi, Vicki Squires, and Kenneth Wilson

University of Saskatchewan

Abstract

The present paper assessed a feedback satisfaction survey for the presence of differential item functioning (DIF) in the evaluation of a learning analytic (LA) feedback intervention. A select few items were flagged for DIF with respect to student type and gender; however, these effects were of negligible size. Overall, it was found that the feedback satisfaction survey was measuring pre-identified demographic groups equally well. Directions for future research and limitations are discussed.

Introduction

In the space of educational assessment, few could argue that valid and reliable measurement is not of the utmost importance. This expectation is especially true in the context of higher education, where decisions based upon educational data have the greatest reach and potential impact in terms of both educational advancement and vocational success (Zhao, Ferguson, Dryburgh, Rodriguez, & Gibson, 2017). Coupled with this emergent understanding, the past decade has seen a rise in the use of learning analytic (LA) systems at post-secondary institutions (Viberg, Hatakka, Bälter, & Mavroudi, 2018). Leveraging the previously untapped resource of 'big data', these systems aim to improve upon a host of student learning outcomes through data-driven activities ranging from predicting drop-out risk (Arnold & Pistilli, 2012) to providing ongoing course feedback interventions (Wright, McKay, Hershock, Miller, & Tritz, 2014). Of paramount importance to this process is the valid and reliable measurement of key variables in LA research. Despite this need, there remains a noticeable absence in the field's application of certain measurement-related techniques, namely, those used to detect differential item functioning (DIF). These statistical procedures permit the researcher to assess whether participants from different demographic groups exhibit the same response patterns to items on a psychometric instrument (Wiberg, 2007). In the following paper, techniques for assessing DIF were applied to a feedback satisfaction survey aimed at assessing student feedback satisfaction with a LA-based intervention. To ensure the survey was not demonstrating bias in assessing students, DIF analyses were conducted with respect to the demographics of student type (new first-time vs non-new-first time), student background (urban vs rural), and gender.

Literature Review

The following review refers to the background literature on the three grouping variables of student type, student background, and gender that will be subjected to DIF analysis. Each variable will be reviewed to the exclusion of the others as each poses a distinct rationale for its inclusion with respect to the DIF analysis. The first variable of student type refers to whether or not a student is a new first-time university student. It is suspected that new-first time students have different expectations regarding in-course feedback compared to non-first-time students, and thus may respond differentially to our feedback satisfaction survey. Student background refers to whether the student attended an urban or rural high school. It is suspected that divergent high school experiences between these groups – possibly related to smaller class sizes and reduced course offerings – may result in varying expectations around feedback, thus warranting its inclusion for analysis. Lastly, gender will be addressed with respect to participant membership as either a male or female. There exists a wealth of literature citing the presence of DIF with respect to gender when measuring similar constructs, all of which will be discussed.

Student Type: The transition from high school into post-secondary education is a key period of transition. New first-year university students come into higher education with expectations largely informed by their prior experience in the school system, many of which are not applicable to the new situation. Looking at these misconceptions more specifically, Rowley, Hartley, and Larkin (2008) examined the beliefs undergraduate psychology students had about university life at the start of their first term and at the end of their first year. Results showed that students had unreal expectations regarding coursework, with participants expressing their desire to have had received more assignment specific feedback throughout the year. In a similar vein, a study by Crisp et al. (2009) noted that many new university students believed that they would have ready access to lecturers outside of class time for assistance. It was also found that upwards of ninety percent of respondents agreed with the statement that "obtaining feedback on drafts of their work would be important for their learning" (Crisp et al., 2009). This information was later presented to teaching staff, who expressed their worry that this desire for widespread and timely delivery of feedback simply could not be accomplished in most situations (Crisp et al., 2009). More recent reviews have echoed the same sentiments, noting that students often have mistaken beliefs about many domains of the university experience ranging from modes of assessment and feedback expectations to teaching style that they must quickly adapt to throughout their early university education (Hassel & Ridout, 2018; Winstone & Bretton, 2013). The departure between first-year student expectations and reality regarding feedback in university coursework makes a strong case for the investigation of DIF when assessing an instrument used to measure feedback satisfaction.

Student Background: The use of DIF techniques to assess differences in the score patterns of students from either urban or rural presents another seldom studied subject. With such little information to act upon, the rationale for the present inquiry comes from the often observed differences in the lived experience between students that attend either rural or urban primary and secondary education. McCracken and Barcinas (1991) highlighted several differences between the urban and rural school experience of students in Ohio. Among these, are the substantial differences in class and staff sizes, course offerings, and extra curricular opportunities, with rural students experiencing the lesser of all categories mentioned. The authors also note that rural Ohio students typically experienced far less ethnically diverse student populations. With respect to a Canadian context, attendance at university and one's likelihood of obtaining a degree correlate positively with the size of community a student comes (Newbold & Brown, 2015).

Perhaps most interesting is the difference of educational experience between students of urban and rural backgrounds. Preston (2012) undertook a thematic analysis of the experiences of urban and rural educators. Findings revealed that the role of developing personal relationships as a teacher was of the utmost importance for those teaching in rural communities – with these relationships ranging from students, to parents, to other important/influential community members. By contrast, the experience of urban teachers found relationships to be far less personal. These teachers were more likely to take a numbers-based approach to evaluating performance – where success is determined by reaching the greatest number of students possible. They noted however, that this approach was more likely to result in students falling through the cracks (Preston, 2012). These differences in feedback expectations combined with a dissimilar school experience establish the logical basis for examining DIF with respect to student background.

Gender: In our recent paper that examined student feedback satisfaction with respect to an LA system, it was discovered that female students in the personalized feedback condition reported significantly higher levels of feedback satisfaction than all other groups (even males within the same experimental condition)(Schmidt, Mousavi, Squires, & Wilson, 2018). Though this finding serves as the primary launch point for a DIF investigation, it does not exist in isolation. A selection of measurable gender differences have been well documented in the psychological sciences (Hyde, 1990). Further, evidence of DIF with respect to gender is commonplace, with no shortage of examples spanning the subjects of science and mathematics (Gierl, Khaliq, & Boughton, 1999; Gierl, Rogers, & Klinger, 1999) to evaluating depressive symptoms (Broekman et al., 2008; Lange, Thalbourne, Houran, & Lester, 2002). Regarding scales assessing various domains of satisfaction specifically, the results are mixed. Gendered DIF has been observed in assessments of job satisfaction (Collins, Raju, & Edwards, 2000), while others examining participant satisfaction with life have found no evidence of DIF with respect to gender (Salsman et al., 2014). These findings suggest that, while gendered DIF in a satisfaction scale is no guarantee, the precedent to check for its existence has been established.

Methods

Participants/Data

Participants for the study are University of Saskatchewan students who enrolled in four sections of an introductory biology course spanning from September 2016 to January 2018. In total there were 3068 participants (1049 Male, 2019 Female, average age = 19.9 years, SD = 3.27, age range = 15-72 years). Participants consented to participate in the study shortly after registering in the course through the university's learning management system — Blackboard. Data was collected from the university's data warehouse, entrance and exit surveys, and from student activity on Blackboard.

Instruments

The feedback satisfaction survey is a short single-factor questionnaire designed to assess a student's level of feedback satisfaction with the University of Saskatchewan's learning analytic platform Student Advice Recommender Agent (SARA) (Greer et al., 2015; Schmidt et al., 2018). The questionnaire utilized two dichotomous (yes/no) items and three Likert-type polytomous items (Strongly Disagree to Strongly Agree) for its first three installations running from September 2016 to September 2017. In its January 2018 installation the questionnaire featured the same five questions; however, the response scales for the first and second items had been updated to the five-point Likert-type items of the final three questions (See Appendix A). The feedback satisfaction survey was contained within a larger exit survey administered to the entire course upon completion. Across years, the same items were administered in the same order and were situated in the same position within the larger exit survey. Higher scores on the survey imply greater levels of feedback satisfaction, with the maximum possible score of 17.

To ensure results across years could be aggregated, the first and second items in the January 2018 version of the survey were dichotomized. This process meant that responses on the disagree side of the scale were coded as not endorsed while responses on the agree side of the scale were coded as endorsed. In accordance with Hanisch's (1992) observation that neutral responses to positively keyed items resembled those of negative responses, neutral responses in the present survey were coded as having not endorsed the item. This coding procedure was done in a similar fashion with the dichotomization of polytomous item responses from Huang, Church, and Katigbak's (1997) study of the NEO-Personality Inventory (Costa & McCrae, 1985). Despite the potential loss of information from this process, the resulting dichotomized scale reported more than adequate reliability with an internal consistency estimate of $\alpha = .85$ (DeVellis, 2016).

Analytical Framework

The demographic variables of interest are student type, student background, and gender. Student status refers to whether a student is registered as a new first-time student, or as a returning student. Student background refers to whether the student completed their high school education at either an urban or rural high school. Lastly, gender refers to the gender that the student self-identified as on the entrance survey.

The presence of DIF was assessed using the R package lordif (Choi, 2015), using a hybrid of logistic ordinal regression and item response theory (IRT) with a graded response model. For each item three hierarchical models were generated with an increasing number of explanatory variables. Significance testing for DIF was accomplished using the likelihood ratio χ^2 test. Items flagged for DIF were evaluated with Jodoin and Gierl's (2001) guidelines for categorizing effects using each flagged comparison's corresponding pseudo r-squared values. Uniform DIF was assessed by comparing the first and second models, while non-uniform DIF was assessed by comparing the second and third. Finally, an overall effect of DIF was assessed by comparing the first and third models.

Results

Table 1 shows the mean scores and standard deviations reported for each demographic pair. Descriptive statistics were highly similar across groups – the only exception being the fifth of a standard deviation difference between female and males.

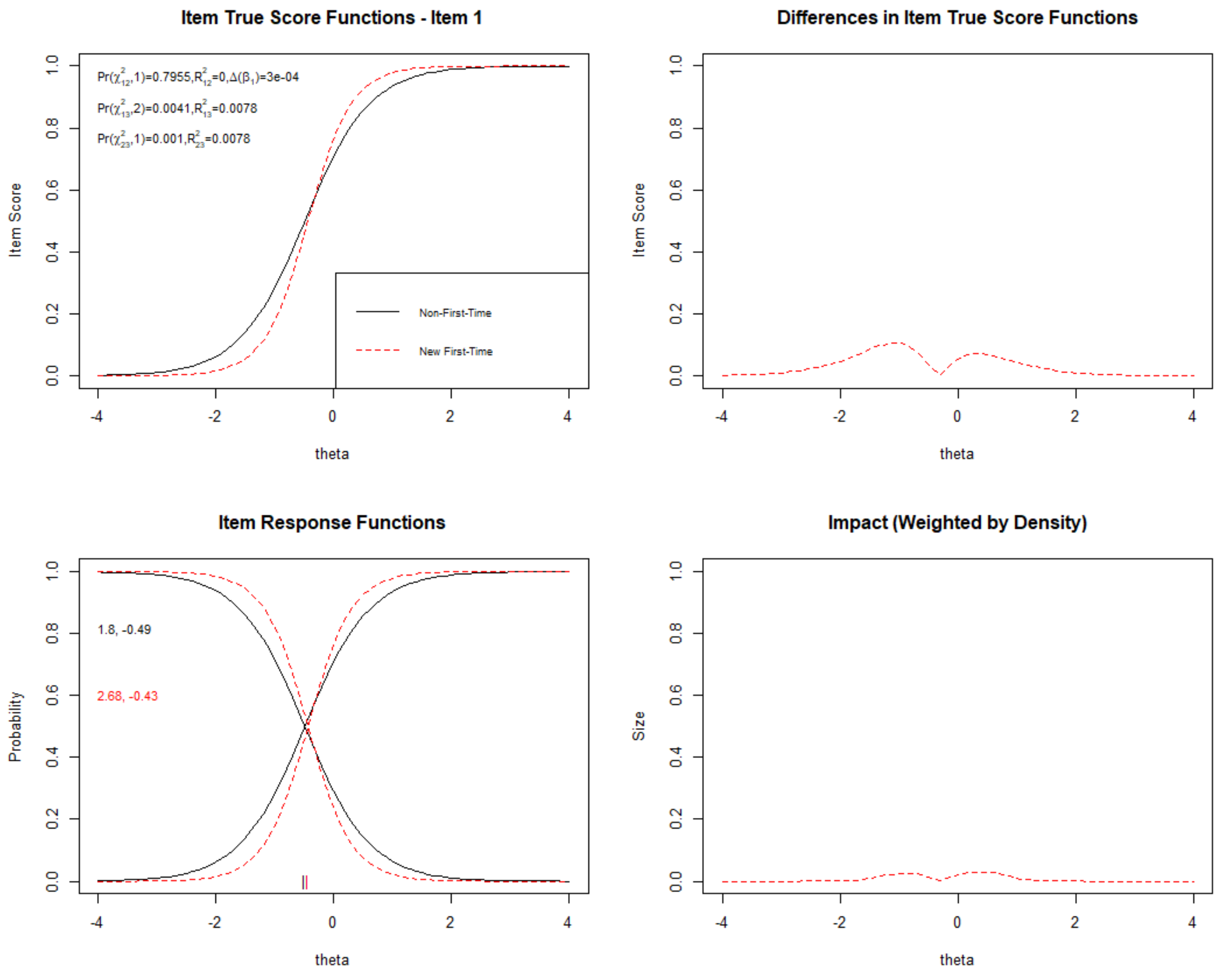
Table 1 Means and Standard Deviations of Subgroups on the Feedback Satisfaction Survey

	M	*SD*
Student Status		
New First-Time	11.38	2.99
Non-First-Time	11.41	3.49
Student Background		
Urban School	11.11	3.18
Rural School	11.47	2.91
Gender		
Female	11.55	3.05
Male	10.93	3.22

Two questions exhibited DIF using a significance threshold of α = .01. Items 1 and 3 were flagged for student type, while item 3 was flagged for gender. It was found however, that all significant effects were of negligible size. These results imply that DIF has only a minimal impact on the observed group differences. No items were flagged for DIF regarding student background. Figures 1 and 2 depict a detailed breakdown of the items flagged for DIF regarding Student Type. Referring to both the true score function and item response function of Figure 1, one can observe an example of non-uniform DIF. Compared to non-first-time students, new first-time students were less likely to endorse the item between feedback appreciation levels (θ) from approximately -3 to -1 and were more likely to endorse the item between appreciation levels ranging from 0 to 2. Looking at the difference in true score function and impact graphs one can see that the greatest absolute differences occurred within these same theta estimates, and when weighted by the score distribution of the new students, the results amounted to minimal impact. Differences were deemed to be of negligible effect size, reporting pseudo r-squared values of $R^2 = .0078$ for both significant model comparisons (model 1 vs model 3 and model 2 vs model 3).

Figure 1 DIF diagnostic for item 1 with reference to student type

Q1 Did you appreciate receiving your weekly note from SARA?



Item True Score Functions - Item 1

Differences in Item True Score Functions

Item Response Functions

Impact (Weighted by Density)

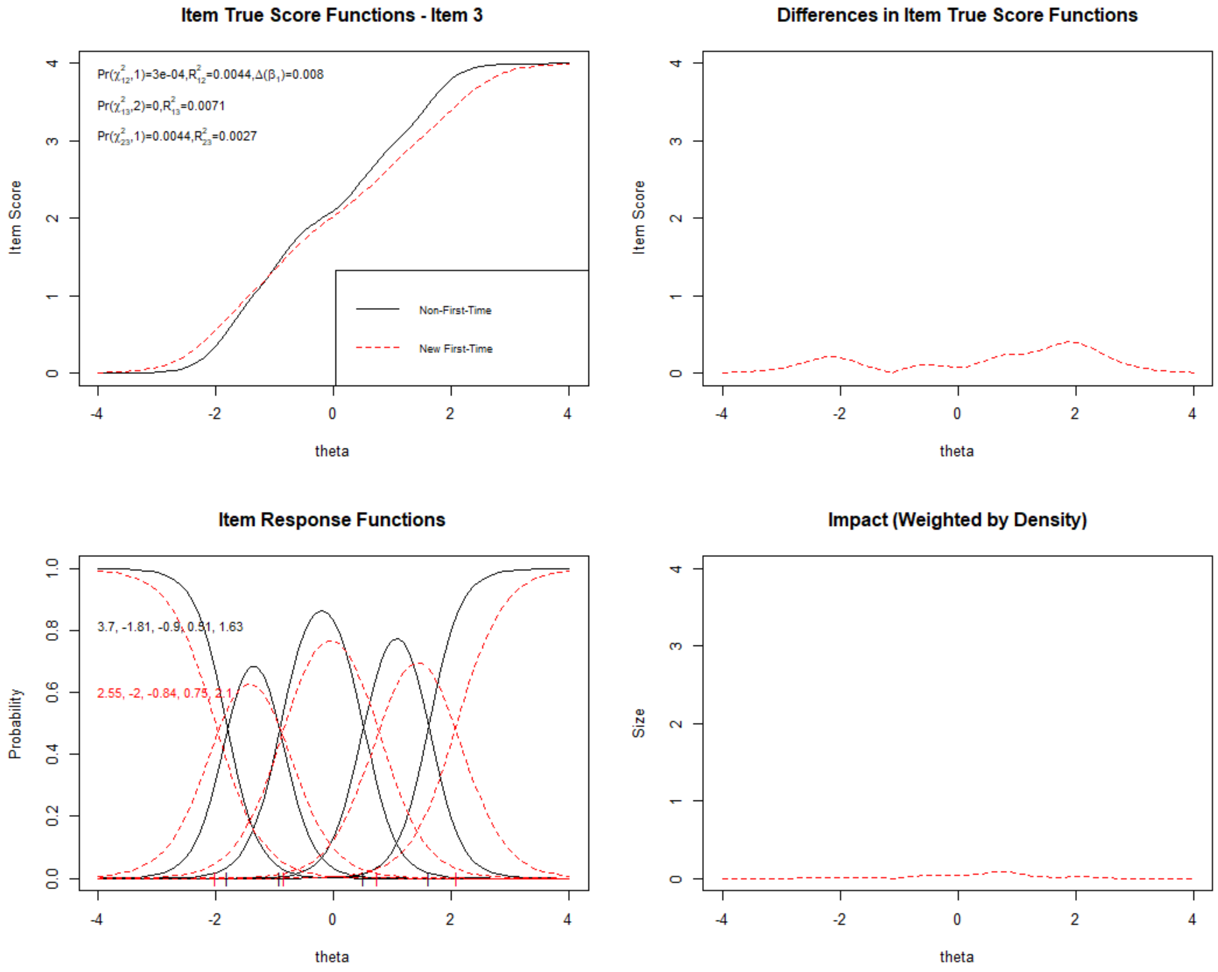
Note: “Item True Score Functions” shows the item characteristic curves (ICCs) for the identified groups. “Differences in Item True Score Functions” shows the absolute difference between the ICCs for the identifed groups. “Item Response Functions” shows the item response functions and item parameters for both demographic groups. “Impact” shows the absolute difference between the ICCs weighted by the score distribution for the focal group (New First-Time).

Referring to Figure 2, item 3 is showing signs of both uniform and non-uniform DIF with respect to student type. The item true score and response functions show that at lower levels of feedback satisfaction (θ < -1), new first-time students are more likely than non-first-time students to respond to the item favourably. This finding is reversed as satisfaction levels are estimated to be above θ = 1. Two distinct peaks in absolute differences between item true scores occurred at θ

= -2 and θ = 2, though when weighted by the new student's score distribution revealed little to no impact. Pseudo r-squared values for significant model comparisons range from R^2 = .0044 (model 1 vs model 2), to R^2 = .0071 (model 1 vs model 3), to R^2 = .0027 (model 2 vs model 3).

Figure 2 DIF diagnostic for item 3 with reference to student type

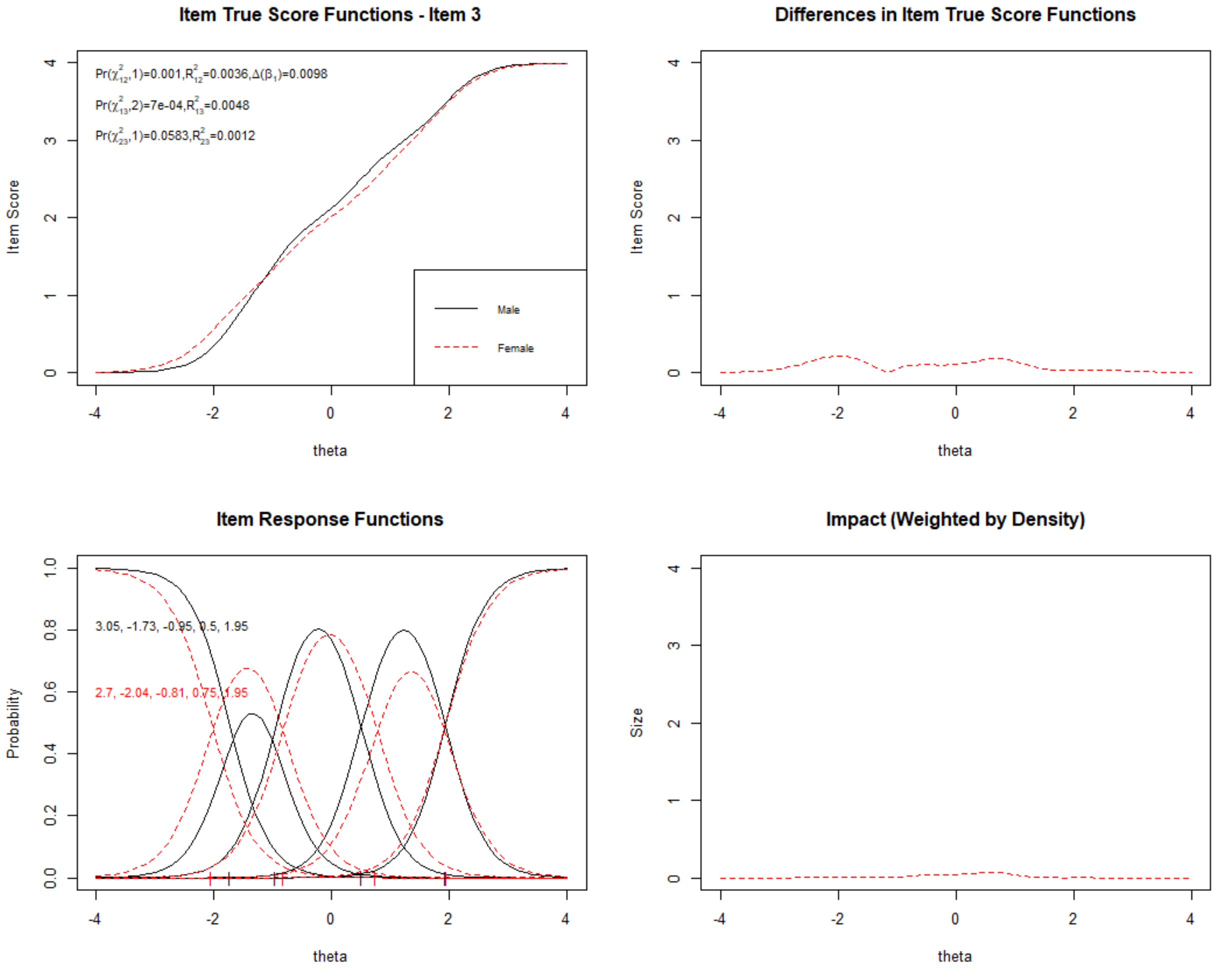
Q3 My weekly note increased my sense of belonging in the University community.



Note: "Item True Score Functions" shows the item characteristic curves (ICCs) for the identified groups. "Differences in Item True Score Functions" shows the absolute difference between the ICCs for the identifed groups. "Item Response Functions" shows the item response functions and item parameters for both demographic groups. "Impact" shows the absolute difference between the ICCs weighted by the score distribution for the focal group (New First-Time).

Figure 3 DIF diagnostic for item 3 with reference to gender

Q3 My weekly note increased my sense of belonging in the University community.



Note: "Item True Score Functions" shows the item characteristic curves (ICCs) for the identified groups. "Differences in Item True Score Functions" shows the absolute difference between the ICCs for the identifed groups. "Item Response Functions" shows the item response functions and item parameters for both demographic groups. "Impact" shows the absolute difference between the ICCs weighted by the score distribution for the focal group (Female).

Only one item was flagged for DIF when comparing participants across gender. Referring to Figure 3 one can observe the incredibly small effects of both uniform and non-uniform DIF. Looking at both the item true score and response function, at approximately $\theta < -1$ female participants are more likely to respond favourably to the item, while from approximately $-1 < \theta < 2$, male participants are more likely to respond favourably to the item. Peak differences in true scores were observed at $\theta = -2$ and $\theta = 1$ though this amounted to no real item impact. Pseudo r-squared values for these flagged effects ranged from R^2 = .0098 (mode 1 vs model 2) to R^2 = .0048 (model 1 vs model 3).

Discussion

This paper represents one of the first times an instrument used to inform an LA feedback system has been subjected to techniques used to assess DIF. Despite the handful of items within the feedback satisfaction survey flagged for DIF, all effects were of negligible size, indicating that conclusions drawn from the survey with respect to the pre-identified demographic groups could be trusted.

It is of interest that item three was flagged for DIF for both student type and gender. Despite the negligible effect size, further investigation may reveal that the item requires some revision. With regards to student type, the greatest differences between response patterns were observed at higher levels of student feedback satisfaction. At this range, it was observed that new first-time students were less likely to endorse higher response values. Though future research is needed to make sense of this finding, it is suspected that expectations around what constitutes "university belonging" between first-time and non-first-time students contributed to this result. The DIF observed for gender on item three is of particular interest, as multiple separate factors may be the cause. Within the context of STEM education, differences in a sense of belonging have long been observed with respect to men and women (Good, Rattan, & Dweck, 2012; Smith, Lewis, Hawthorne, & Hodges, 2013). On the other hand, this difference in response pattern may be to due to gender differences at a more fundamental level. Baumeister and Sommer (1997) suggested that men and women seek different types of social connection and interactions to acquire a sense of belonging. They suggested that women generally seek relationships where they can develop intimate connections with others, while men prefer to have a sense of separateness and independence.

DIF aside, an analysis of item three suggests that the item is performing well. It was found however, that item three possesses the lowest discrimination parameter of the instrument both with respect to a classical test and IRT analysis, suggesting that it has both the lowest correlation with total score on the instrument and provides the least amount of peak information. Though the item is functioning adequately, future research conducted on subsequent iterations of this course may consider revising the item.

The DIF effect flagged for item one may represent the previously discussed differences in expectations between new first-time and non-first-time students in course feedback. The smaller class sizes often observed in high school more easily allow an instructor to deliver the one-on-one feedback that is often impossible to reproduce in large introductory post-secondary courses. This disparity may have contributed to the slight difference observed in response patterns. Further, the post-secondary experience that non-first-time students already possessed with regards to feedback expectations at university may have contributed as well. It is important however, to bear in mind that these flagged effects were of negligible size.

No items were flagged for DIF with respect to student background. Given that the rationale for exploring student type and background was largely overlapping, this result is somewhat surprising. One might have expected that the differences in educational experience – particularly with respect to class sizes, availability of instructor relationships, and teaching styles – might have more thoroughly changed the ways in which urban and rural students responded to a survey addressing satisfaction with feedback. Though surprising, the results also serve as a relief. Knowing conclusively that the feedback satisfaction survey is measuring those of urban and rural backgrounds equally well is reassuring and gives us confidence in our results moving forward.

Limitations

The present paper used IRT-based methods for detecting DIF, which is only one of many available methods. IRT methods were selected for the advantage of providing specific item characteristics for both the items and the test as a whole. It is possible however, that the use of another method may have yielded different results.

A number of limitations relating to design and data manipulation must be addressed. One key limitation is the degree to which results can be generalized across the years the course was implemented. Though each iteration was highly similar, it cannot be considered the same in the way a tightly controlled experimental condition might be. Despite this potential limitation, the aggregation of data across years was carried out for reasons relating to sample size. The use of IRT-based methods and DIF detection methods more specifically, requires substantial sample sizes – especially as the number of item parameters to be estimated increases (Edelen & Reeve, 2007). Lastly, it is possible that some meaningful information was lost through the dichotomization of the first two items in the January 2018 administration of the course, and that this may have adversely impacted the results.

Another potential limitation of the research is the length of the feedback satisfaction survey. Despite achieving a respectable reliability estimate, the survey is relatively short, and may be lacking content tapping the full spectrum of the construct of satisfaction with SARA's feedback. In other words, there may exist equally applicable items not considered in the original creation of the survey that more explicitly tap the differences between our identified demographics.

As it pertains to future research directions, ongoing research opportunities into the current project regarding DIF are limited. However, this paper should serve as an example to other developers interested in using psychometric instruments to inform the functions of an LA system. Fortunately for the system under study in the present paper, no substantial effects of DIF were found. However, the presence of significant differences in response patterns in the present paper illustrates this key warning to developers. As such, future research should bear in mind the potential for such issues in measurement and ensure that their instruments are invariant across groups.

Conclusion

To make informed decisions about one's results in the realm of the social sciences, the validity of one's measurements must be ensured. It was suggested at the outset of this paper that given the demographics of student type, gender, and student background, students may very well display evidence of DIF when responding to a survey used to assess the level of feedback satisfaction with a computer-generated LA feedback intervention. No items were observed to exhibit DIF with respect to student background, and though a select number of items were flagged for DIF concerning both student type and gender; these were all of negligible effect. These findings suggest that the information acquired from the feedback satisfaction survey was both reliable and valid, but that improvements might still be made to the survey, such that it may better serve its purpose.

References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, (April 2012), 267. https://doi.org/10.1145/2330601.2330666

Baumeister, R. F., & Sommer, K. L. (1997). What do men want? Gender differences and two spheres of belongingness: Comment on Cross and Madson (1997).

Broekman, B. F. P., Nyunt, S. Z., Niti, M., Jin, A. Z., Ko, S. M., Kumar, R., … Ng, T. P. (2008). Differential item functioning of the Geriatric Depression Scale in an Asian population. *Journal of Affective Disorders, 108*(3), 285–290.

Choi, S. W. (2015). Lordif: logistic ordinal regression differential item functioning using IRT.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85*(3), 451.

Costa, P. T., & McCrae, R. R. (1985). The NEO personality inventory.

Crisp, G., Palmer, E., Turnbull, D., Nettelbeck, T., Ward, L., LeCouteur, A., … Schneider, L. (2009). First year student expectations: Results from a university-wide student survey. *Journal of University Teaching and Learning Practice*, *6*(1), 11–26.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and judgmental reviews to identify and interpret translation DIF. In *annual meeting of the National Council on Measurement in Education, Montreal, QC*.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999). Gender differential item functioning in mathematics and science: Prevalence and policy implications. In *Improving Large-Scale Assessment in Education Symposium at the Annual Meeting of the Canadian Society for the Study of Education, Canada. Retrieved February* (Vol. 25, p. 2008).

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, *102*(4), 700.

Greer, J. E., Frost, S., Banow, R., Thompson, C., Kuleza, S., Wilson, K., & Koehn, G. (2015). The Student Advice Recommender Agent: SARA. In *UMAP Workshops*.

Hanisch, K. A. (1992). The Job Descriptive Index revisited: Questions about the question mark. *Journal of Applied Psychology*, *77*(3), 377.

Hassel, S., & Ridout, N. (2018). An Investigation of First-Year Students' and Lecturers' Expectations of University Education. *Frontiers in Psychology*, *8*, 2218.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*(2), 192–218.

Hyde, J. S. (1990). Meta-analysis and the psychology of gender differences. *Signs: Journal of Women in Culture and Society*, *16*(1), 55–73.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349.

Lange, R., Thalbourne, M. A., Houran, J., & Lester, D. (2002). Depressive response sets due to gender and culture-based differential item functioning. *Personality and Individual Differences*, *33*(6), 937–954.

McCracken, J. D., & Barcinas, J. D. T. (1991). High School and Student Characteristics in Rural and Urban Areas of Ohio.

Newbold, K. B., & Brown, W. M. (2015). The Urban--Rural Gap in University Attendance: Determinants of University Participation Among Canadian Youth. *Journal of Regional Science*, *55*(4), 585–608.

Preston, J. P. (2012). Rural and urban teaching experiences: Narrative expressions. *Alberta Journal of Educational Research*, *58*(1), 41–57.

Rowley, M., Hartley, J., & Larkin, D. (2008). Learning from experience: the expectations and experiences of first-year undergraduate psychology students. *Journal of Further and Higher Education*, *32*(4), 399–413.

Salsman, J. M., Lai, J.-S., Hendrie, H. C., Butt, Z., Zill, N., Pilkonis, P. A., … Cella, D. (2014). Assessing psychological well-being: self-report instruments for the NIH Toolbox. *Quality of Life Research*, *23*(1), 205–215.

Schmidt, M., Mousavi, A., Squires, V., Wilson, K. (2018). Assessing the effectiveness of automated personalized feedback in an undergraduate biology course. In *Proceedings of Hawaii International Conference Science, Technology & Engineering, Arts, Mathematics &*

Education (pp. 1–17).

Smith, J. L., Lewis, K. L., Hawthorne, L., & Hodges, S. D. (2013). When trying hard isn't natural: Women's belonging with and motivation for male-dominated STEM fields as a function of effort expenditure concerns. *Personality and Social Psychology Bulletin*, *39*(2), 131–143.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods. Institutionen för beteendevetenskapliga mätningar, Umeå.

Winstone, N., & Bretton, H. (2013). Strengthening the Transition to University by Confronting the Expectation-Reality Gap in Psychology Undergraduates. *Psychology Teaching Review*, *19*(2), 2–14.

Wright, M. C., McKay, T., Hershock, C., Miller, K., & Tritz, J. (2014). Better Than Expected: Using Learning Analytics to Promote Student Success in Gateway Science. *Change: The Magazine of Higher Learning*, *46*(1), 28–34. https://doi.org/10.1080/00091383.2014.867209

Zhao, J., Ferguson, S. J., Dryburgh, H., Rodriguez, C., & Gibson, L. (2017). Does Education Pay? A Comparison of Earnings by Level of Education in Canada and Its Provinces and Territories. Census of Population, 2016. Census in Brief. *Statistics Canada*.

Appendix A

Feedback Satisfaction Survey (Schmidt et al., 2018)

Feedback Satisfaction Survey for September 2016, January 2017, and September 2017 course offerings.

Your NSID and survey feedback will remain confidential. Your instructors will never see individual data.

		Yes	No	N/A
1.	Did you appreciate receiving your weekly note from SARA?	○	○	○
2.	Did you find that your weekly note from SARA was personalized to your academic situation?	○	○	○

To what extend do you agree or disagree with the following statements:

		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	N/A
3.	My weekly note increased my sense of belonging in the University community.	○	○	○	○	○	○
4.	My weekly note was encouraging with respect to my academic situation.	○	○	○	○	○	○
5.	My weekly note was a good reminder of my performance in Biology 120.	○	○	○	○	○	○

Note: Survey was administered electronically following completion of the course and was nested within the larger Biology 120 Student Resource Exit Survey.

Feedback Satisfaction Survey for January 2018 course offering.

Your NSID and survey feedback will remain confidential. Your instructors will never see individual data.

To what extend do you agree or disagree with the following statements:

		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	N/A
1.	I appreciated receiving my weekly note from SARA.	○	○	○	○	○	○
2.	My weekly note was personalized to my own academic situation.	○	○	○	○	○	○
3.	My weekly note increased my sense of belonging in the University community.	○	○	○	○	○	○
4.	My weekly note was encouraging with respect to my academic situation.	○	○	○	○	○	○
5.	My weekly note was a good reminder of my performance in Biology 120.	○	○	○	○	○	○

Note: Survey was administered electronically following completion of the course and was nested within the larger Biology 120 Student Resource Exit Survey