Measuring the Power of Learning.®

Research Memorandum
ETS RM–17-07

Exploratory Analysis of Differential Item Functioning and Its Possible Sources in the National Survey of Student Engagement

Stephanie Barclay McKeown

María Elena Oliveri

December 2017

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Exploratory Analysis of Differential Item Functioning and Its Possible Sources in the National Survey of Student Engagement

Stephanie Barclay McKeown
University of British Columbia, Vancouver, Canada

María Elena Oliveri
Educational Testing Service, Princeton, New Jersey

December 2017

Corresponding author: S. Barclay McKeown, E-mail: stephanie.mckeown@ubc.ca

Suggested citation: Barclay McKeown, S., & Oliveri, M. E. (2017). *Exploratory analysis of differential item functioning and its possible sources in the National Survey of Student Engagement* (Research Memorandum No. RM-17-07). Princeton, NJ: Educational Testing Service.

Action Editor: Matthias von Davier and Rebecca Zwick

Reviewers: Shelby Haberman and John R. Lockwood

Abstract

In this study, we examined differential item functioning (DIF) of the Deep Approaches to Learning scale on the National Survey of Student Engagement (NSSE) for Asian international and Canadian domestic 1st-year university students. We also examined its potential sources of using focus-group interview results. Only 1 of the 12 items functioned differentially, possibly due to differences in how each group interpreted the term “faculty members” in the NSSE. Further research should be conducted beyond this initial exploratory study to examine DIF with other international student groups classified by country and/or language.

Key words: differential item functioning, validity, National Survey of Student Engagement, higher education, international, student experience

Currently, tertiary education students from outside North America are attending university in the United States and Canada in greater numbers. Surveys asking students about their learning engagement are administered to both international and domestic enrolling students to identify the extent to which they are benefitting from learning and teaching opportunities provided by the enrolling institutions. However, prior to drawing any conclusions based on survey responses across different groups, researchers should examine the extent to which the groups are interpreting the survey items in construct-relevant ways. Otherwise, construct-irrelevant conclusions may arise if one group interprets certain concepts or constructs differently than the other group. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and the International Test Commission (2017) guidelines highlight this issue and argue that the equivalent measurement of constructs across comparison groups is a basic requirement for making meaningful inferences based on test scores and survey responses.

When assessing culturally and linguistically diverse populations, construct nonequivalence may be introduced if items in a test or survey include language or terms that are interpreted differently by diverse subgroups or if the test’s items have a format that is not similarly familiar across groups (Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Sireci, Harter, Yang, & Bhola, 2003). In surveys, group differences also may arise in response styles such as differences in willingness to comply, tendency to guess, and social desirability. To illustrate, Thomson and Douglass (2009) compared self-reported responses of Asian and U.S. students from the University of California Undergraduate Experience Survey. The authors noted that the two (Asian versus U.S. domestic) student groups reported information about learning experiences differently. Specifically, they reported that Asian students rated their learning gains (e.g., quantitative, oral, writing, reading, and thinking) lower than non-Asian students did, which the researchers suggested might be indicative of differences in achievement-related expectations, differences in social desirability, or halo effects across groups.

Despite previous research suggesting that diverse cultural and linguistic groups may respond differentially to survey items due to possible construct-irrelevant reasons (Banks, 2006; Kristjansson, Desrochers, & Zumbo, 2003), to date most survey-based research is conducted without examining the comparability of constructs assessed by the survey across groups

(Grayson, 2008; Zhao, Kuh, & Carini, 2005). For instance, Dowd, Sawatzky, and Korn (2011) highlighted that often such analyses are not conducted with the National Survey of Student Engagement (NSSE; 2011), which collects information about domestic and international students' perceptions about their tertiary institutions' support for teaching, learning, and educational climate. The authors argued that reducing possible cultural biases in support of an accurate understanding of the diverse respondent groups' perceptions of the data collected by the NSSE is important given the NSSE's wide administration (over 1,500 post-secondary institutions across the United States and Canada) for international and domestic student groups (G. Kuh, 2009; Thomson & Douglass, 2009).

One way of examining the assumption of measurement comparability across groups at the item level is to conduct differential item functioning (DIF) analyses. DIF methods have primarily been used to examine comparability across subgroups using data from achievement tests and, less frequently, to analyze DIF using attitudinal survey data, as is the case of the NSSE. In surveys, DIF is conceptualized as occurring when survey respondents who have similar attitudes on a measured trait respond differentially due to construct-irrelevant factors such as translation or adaptations made to an original survey or differential interpretation of terms used in survey items. The flagging of DIF suggests that a test item may indicate bias if its sources point to particular construct-irrelevant features as a source of DIF across comparison groups (Penfield & Lam, 2000). When DIF is detected, further analyses may be conducted to examine why some items function differentially across respondents (Kristjansson et al., 2003; Sireci et al., 2003; Sireci, Yang, Harter, & Ehrlich, 2006). By identifying the sources of DIF, researchers can use these results to further inform test development practice and decide how to treat the item by either discarding or modifying the item or by changing particular terms or vocabulary that might be leading to DIF (Karami & Nodoushan, 2011).

Various approaches have been used to analyze sources of DIF and determine whether the DIF is construct relevant or irrelevant (Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2012; Oliveri & Ercikan, 2011). For instance, experts have often reviewed items after test administration. Ercikan et al. (2010) and Zumbo (2007) suggested that one challenge of this approach, wherein items are often reviewed post hoc, experts might (over-)fit their explanations to potentially idiosyncratic differences associated with an item rather than systematic differences. Moreover, Karami and Nodoushan (2011) suggested that solely relying on linguistic reviews of items using informants, or content experts, may not provide sufficient information to identify sources of DIF. Similarly, Ercikan et al. (2010) also suggested conducting additional analyses using think-aloud protocols to capture the thought-processes from examinees to confirm expert-based interpretations of DIF because they found that the panel of content experts participating in their study were able to identify only half of the DIF items (10 out of 20) on the assessment. Additionally, Roussos and Stout (1996) recommended a two-step confirmatory approach to examining sources of DIF. The first step involves hypothesizing where, why, and how DIF occurs based on substantive findings informed by previous research, and the next step involves testing out the hypotheses using multidimensional DIF detection approaches.

In this study, we first investigated whether NSSE items on the Deep Approaches to Learning scale functioned differentially across Asian international versus Canadian domestic first-year students. Second, we compared DIF findings with the results from a qualitative focus-group study conducted by Suderman (2015).

Method

Measure

We analyzed responses from the 2011 version of the NSSE, which is a large-scale 85-item survey administered to students to inform institutions about their students’ learning and engagement. Depending on the institution, it is administered annually, or every 2 years to only first-year and graduating senior year students. In 2011, more than one million college and university students at 761 four-year degree granting institutions in the United States and Canada were sent the survey (NSSE, 2011). There are two versions of the NSSE. One version is administered to first-year students and asks students to indicate whether they plan to participate in a learning community, internships, community service, study abroad, or work with a faculty member on research projects. The second is administered to graduating senior students and asks whether they have participated in the abovementioned activities. Students are asked to record their perceptions of their postsecondary environment, including the extent to which the institution offers the support students need to succeed academically and the quality of relations between various groups on campus such as faculty and students.

The NSSE is designed to tap into two notions of university quality: (a) the amount of time and effort students put into their studies and other educationally purposeful activities, and (b) how the institution deploys its resources and organizes the curriculum and other learning opportunities to get students to participate in activities intended to be linked to student learning (G. Kuh, 2009). NSSE results are used in facilitating benchmarking related to (a) level of academic challenge, (b) active and collaborative learning, (c) student interactions with faculty members, (d) richness of educational experience, and (e) level of support of the campus environment. Schools are considered to be effective if they score above average on these levels (G. D. Kuh et al., 2001). The results are embedded in institutional improvement initiatives, including accreditation reports and reviews, agendas for improvement, and faculty participation in improvement; they are also used to enhance student learning experiences by influencing institutional policy decisions and program development (Hayek & Kuh, 2004).

Data

Figure 1 lists the Deep Approaches to Learning items we used in the study, the definition of each item, and the subscales to which they belong. Because issues related to deep learning and its measurement in higher education have recently drawn much attention and may vary across cultures, we elected to analyze these items of the NSSE.

Sample

In this study, we compared responses of Asian international versus Canadian domestic students within one institution. C. M. Campbell and Cabrera (2011) discussed focusing NSSE research on a single institution to enable finer-grained, score-based interpretations. We selected students based on their visa status, which in Canada classifies students as “domestic” if they have Canadian citizenship or permanent residency permits or as “international” otherwise. We used 2011 NSSE responses from first-year students attending a large research-intensive university in western Canada.

Our sample consisted of 193 Asian international students and 1,369 Canadian domestic students. The Asian international students included students who self-identified as Chinese, Korean, Japanese, and Southeast Asian and were referred to collectively as “Asian international students.” Chinese students (59.3%) contributed the most to the sample of Asian international students. The remaining students included Korean (23.4%), Southeast Asian (12.0%), and Japanese (5.3%). We were interested in examining the experiences of Asian international students given the growth in the number of students who are coming from Asia to attend higher education institutions in Canada and the United States (cf. Altbach, Reisberg, & Rumbley, 2009; Kelly, Moores, & Moogan, 2012; Martel & D'Aoust, 2016).

Higher Order Learning Subscale
During the current school year, how much has your coursework emphasized the following mental activities? Response options: *very much, quite a bit, some, very little.*

1. Analyzing the basic elements of an idea, experience, or theory, such as examining a particular case or situation in depth and considering its components.
2. Synthesizing and organizing ideas, information, or experiences into new, more complex interpretations and relationships.
3. Making judgments about the value of information, arguments, or methods, such as examining how others gathered and interpreted data and assessing the soundness of their conclusions.
4. Applying theories or concepts to practical problems or in new situations.

Integrative Learning Subscale
In your experience at your institution during the current school year, about how often have you done each of the following? Response options: *very often, often, sometimes, never.*

5. Worked on a paper or project that required integrating ideas or information from various sources.
6. Included diverse perspectives (different races, religions, genders, political beliefs, etc.) in class discussions or writing assignments.
7. Put together ideas or concepts from different courses when completing assignments or during class discussions.
8. Discussed ideas from your readings or classes with faculty members outside of class.
9. Discussed ideas from your readings or classes with others outside of class (students, family members, co-workers, etc.)

Reflective Learning Subscale
During the current school year, about how often have you done each of the following? Response options: *very often, often, sometimes, never.*

10. Examined the strengths and weaknesses of your own views on a topic or issue.
11. Tried to better understand someone else's views by imagining how an issue looks from his or her perspective.
12. Learned something that changed the way you understand an issue or concept.

Figure 1. The Deep Approaches to Learning subscales and items.

To create our sample, we followed suggestions made by Fritz, Chin, and Demarinis (2008) who suggested further defining an international group of test takers into subcategories such as Asian international or European international where sample sizes permit. The authors discussed that finer-grained groupings can help better interpret differences across groups, suggesting that, in their study, students from diverse origins reported distinct differences in academic life stressors such as concerns with language versus family separation for Asian international and European international students, respectively. Clearly, such finer group distinctions are helpful because important differences are likely to remain within such groups in relation to language background, culture, and learning preferences that may not be detected by using a larger group such as Asian international. However, while making such finer distinctions (e.g., conducting analyses of international students per country of origin or language) would be ideal, such conditions also need to be considered against limitations of sample sizes needed for DIF analyses. In the discussion, we provide further suggestions for conducting DIF analyses in future research with finer group sizes. In this study, we were unable to use additional samples because we aimed to analyze the same group with which focus-group interviews had been conducted in the study by Suderman (2015) discussed later in the paper.

Analysis

Although ANCOVA is an unfamiliar approach used for identifying DIF, Sireci et al. (2003) first demonstrated the utility of ANCOVA to flag survey items that functioned differentially across groups. They applied the ANCOVA approach along with logistic regression to examine survey responses collected from different language versions of an employee survey. They reasoned that DIF analyses were variations of ANCOVA analyses, which examine group differences on an outcome measure after controlling for a partialled out variable. When they compared the results of the ANCOVA to the logistic regression results, they found that the ANCOVA and logistic regression approaches were almost identical in that they flagged similar items for DIF and that the effect size measures were highly correlated (.96). They concluded that ANCOVA could be useful to researchers wanting to evaluate DIF for survey data with Likert-type items, as is the case with our survey data (Sireci et al., 2003; Sireci et al., 2006).

We analyzed DIF in two steps. First, we investigated DIF using the ANCOVA (see Sireci et al., 2006) DIF approach. Next, if DIF was found, we investigated its possible sources by examining whether there was correspondence between our DIF results and findings from the focus groups conducted by Suderman (2015). We provide additional details of each of these steps next.

Differential Item Functioning Analysis

We started our ANCOVA analyses by testing the assumption of homogeneity-of-regression slopes, which evaluates whether there is an interaction between the covariate (the composite scores) and the group membership variable (Canadian domestic versus Asian international) in predicting the item score. We used group membership for the independent (fixed) variable; the dependent variable was the item score, and the covariate was the rest score on each of the subscales listed in Figure 1 (higher order, integrative, or reflective learning). The rest score is the total subscore minus the studied item score. In our analyses, because each of the Deep Approaches to Learning subscales has few items, we followed the suggestion made in Sireci et al. (2003) regarding using the rest score rather than the total score when a scale has few items in order to remove the influence of the studied item on the covariate. We note, however, that the use of the rest score is contrary to usual DIF practice in which traditionally the test score is used for analysis (Dorans & Holland, 1993).

We conducted separate ANCOVAs for each item in each subscale. We flagged items for DIF if there was a significant effect for the group membership variable ($p < .05$) and if they met the following effect size criteria. Effect size in ANCOVA uses the eta-squared index as an *R*-squared effect size measure. Based on the cut-off values used by Sireci et al. (2003, p. 138), we determined moderate DIF if the item had a partial eta squared value greater than .035 but less than .07. Large DIF is identified if the partial eta squared is equal to or greater than .07. Sireci et al. (2003, 2006) used these conservative cut-off values because they were equivalent to the approach used at Educational Testing Service for flagging large DIF items as described by Dorans and Holland (1993).

Analysis of Sources of Differential Item Functioning

To analyze sources of DIF, we examined the correspondence between the detected DIF items and the findings conducted in Suderman (2015). That study involved 18 focus groups with 77 randomly selected participants attending the same university as in our study. The students were Asian international students from five countries: Taiwan, mainland China, Hong Kong, Korea, and Japan. These focus groups were conducted to examine differences across Asian international and Canadian domestic students' perceptions of engagement with their institution and how their home culture may impact those perceptions.[1] Usefulness of the focus-group approach to analyze sources of DIF was discussed in a study by Linsday and Hubley (2006). In the study, the authors explored respondents' perceptions of their age based on age identity (participants' perceptions of age and the activities they were able to accomplish influencing their beliefs on age) rather than their chronological age. Linsday and Hubley suggested that conducting focus groups allowed participants to perceive their own views in a larger context of others' views and perceptions within a larger social context where language, meaning, and context are inextricable from each other. The authors noted that focus-group interviews yielded insights that went above and beyond providing them with an understanding of age identity that they could not obtain solely by conducting interviews with participants individually.

Results

Descriptive Statistics

Table 1 lists sample sizes and descriptive statistics for the entire sample as well as the Canadian domestic and Asian international groups separately. The means, standard deviations, and reliabilities are similar for both groups; however, the Canadian domestic group mean is slightly higher than the Asian international group mean. An independent samples *t*-test revealed no significant difference in total score means between the groups, *t* (1474) = -1.49, *p* > .05, as well as in the subscales higher order learning *t* (1530) = -1.40, *p* > .05, integrative learning *t* (1524) = -.68, *p* > .05, and reflective learning *t* (1534) = -1.42, *p* > .05.

Table 1. Descriptive Statistics of the Deep Approaches to Learning Scale and Subscales

Scale	Group	Mean (standard deviation)	Coefficient alpha
DAL	Canadian domestic	32.31 (5.66)	.80
DAL	Asian international	31.66 (5.34)	.77
DAL	Total group	32.22 (5.62)	.79
HOL	Canadian domestic	12.05 (2.47)	.74
HOL	Asian international	11.78 (2.55)	.79
HOL	Total group	12.01 (2.48)	.75
IL	Canadian domestic	12.39 (2.71)	.64
IL	Asian international	12.25 (2.67)	.62
IL	Total group	12.37 (2.71)	.63
RL	Canadian domestic	7.87 (2.18)	.79
RL	Asian international	7.63 (1.94)	.67
RL	Total group	7.84 (2.16)	.77

Note. DAL = Deep Approaches to Learning scale (12 items); HOL = Higher Order of Learning subscale (4 items); IL = Integrative Learning subscale (5 items); RL = Reflective Learning subscale (3 items). Canadian domestic: *n* = 1369. Asian international: *n* = 193. Total group: *n* = 1562.

In Table 2, we present the item means and standard deviations for the Canadian domestic students and the Asian international students. Means and standard deviations were similar for all items except Item 8. This item asks students, “How often did you discuss ideas from your readings or classes with faculty members outside of class?” with response options including “never,” “sometimes,” “often,” or “very often.” A nonparametric independent samples Mann-Whitney test revealed no significant differences between the two groups’ means on all items with the exception of Item 8, for which the group means were significantly different, with Asian international students more likely to report they discussed ideas with faculty members outside of class.

Table 2. Item Means (Standard Deviations) per Group

Study item number	Canadian domestic	Asian international
1	3.19 (.73)	3.10 (.72)
2	2.93 (.84)	2.92 (.85)
3	2.78 (.86)	2.78 (.82)
4	3.14 (.85)	2.99 (.85)
5	2.93 (.83)	2.87 (.86)
6	2.50 (.94)	2.50 (.90)
7	2.58 (.80)	2.49 (.75)
8	1.62 (.79)	1.97 (.89)*
9	2.75 (.88)	2.43 (.83)*
10	2.41 (.91)	2.36 (.83)
11	2.65 (.86)	2.54 (.91)
12	2.80 (.84)	2.74 (.75)

Note. Item 8 group means are significantly different.
**p* <. 05.

Differential Item Functioning Results

Table 3 presents the ANCOVA DIF results. We found statistically significant mean differences between the Canadian domestic and Asian international groups for Item 8, which we identified as having moderate DIF. This result suggests that the Asian international students endorsed this item more frequently than the Canadian domestic students. The remaining items do not exhibit DIF. None of the 12 Deep Approaches to Learning items had significant interactions at $p < .05$, which allowed us to run ANCOVAs for all items.

Table 3. ANCOVA Differential Item Functioning Results

Item number	F	Significance	Partial eta squared
1	0.17	.68	.00
2	0.42	.52	.00
3	0.07	.79	.00
4	0.04	.85	.00
5	0.00	.98	.00
6	0.09	.76	.00
7	0.05	.82	.00
8	17.56	.00	.04
9	2.94	.10	.01
10	1.04	.31	.00
11	.33	.56	.00
12	.04	.84	.00

Note. Bold indicates that Item 8 was flagged as exhibiting moderate DIF.

Comparison of Differential Item Analysis Findings and Focus Group Results

Our analysis of the correspondence between items detected as DIF and the focus-group interviews revealed that a possible source of DIF for Item 8 was the way in which the two (Asian international and Canadian domestic) student groups may have interpreted the term *faculty members*. The results of the focus-group interviews suggested that Asian international students had a much broader understanding of *faculty members,* which we label as "Faculty." We label the term with a capital "F" to convey the Asian international students' broader interpretation of this term as inclusive of anyone who would be a member of the faculty, including other students, staff, and administrators working within the faculty rather than professors only. Another plausible explanation is some ambiguity in the wording of the question. For instance, one might interpret the question as asking about whether students discuss ideas with their instructor, only they do it outside of the 50-minute class period, or whether students discuss their ideas with teachers outside of their own classroom instructor. Because the NSSE is administered to students from various countries and regions, item wording and additional explanations to key terms are important, as we elaborate further in the discussion section.

To further examine the possible differential use of the term *faculty members* by the two groups, we conducted a post hoc investigation of three additional items in the NSSE (outside the Deep Approaches to Learning scale) that also contained this term. Two of these items asked students how frequently they "talked about career plans with a faculty member or advisor" and "worked with faculty members on activities other than coursework (committees, orientation, student life activities, etc.)." These items used the same four-point response options that were used for the Deep Approaches to Learning items ("very often," "often," "sometimes," and "never"). The third item asked students about their relationship with faculty members using a seven-point response option, which ranged from "unavailable, unhelpful, unsympathetic" on one end to "available, helpful, sympathetic" on the other. Table 4 displays the means and standard deviations for these three items. Results of a nonparametric independent samples Mann-Whitney test indicate that there were differences in group means for two items (13 and 14). No significant differences between the two groups' means were found for Item 15, which had a seven-point response option.

Table 4. Item Means (Standard Deviations) per Group

Study item number	Canadian domestic	Asian international
13	1.56 (.74)	1.77 (.81)*
14	1.28 (.60)	1.76 (.86)*
15	4.74 (1.27)	4.71 (1.34)

Note. Bold indicates that Items 13 and 14 group means are significantly different.
**p* < .05.

Discussion

In this study, we analyzed DIF on the 12 items of the Deep Approaches to Learning scale among Asian international versus Canadian domestic students in Canada. We also examined correspondence between DIF and focus group results. One item was flagged as DIF. Results from the focus groups pointed to possible sources of DIF for this item related to differential use of the term faculty members by the two groups.

To further investigate whether the term faculty members led to DIF across all items of the NSSE scale, we tested NSSE items beyond the Deep Approaches to Learning subscale that contained this term. We found mean differences for two other items (13 and 14) in the NSSE scale containing this term. We conjectured that possible reasons for these differences may include differences between the two groups in terms of perceptions of relationships with faculty members. For instance, the wording of the items suggested a collaborative relationship with a faculty advisor or faculty member; this faculty–student relationship may not be similar across the two groups (Wong, 2004). For instance, Eaves (2011) pointed to cultural differences in the perception of teacher–student relations with students from an Eastern philosophy of education perceiving teachers as possessing all knowledge and students as passively absorbing the knowledge. In contrast, students from Western cultures expect teachers to be available to interact with students more frequently. This difference in philosophy of teachers' roles may lead to different expectations of teachers and instructors across groups, which may help explain the identified group mean differences noted in Table 4.

On the other hand, Item 15 did not reveal differences in group means. One reason may be that the item provides more information in the form of characteristics defining the relationship with faculty members, such as (un)available, (un)helpful, or (un)sympathetic, which might lead to a checklist type item with fewer differences across the selection of options across both groups. The analysis of the interaction between item response format, response style, and DIF is beyond the scope of our study and should be the subject of future studies.

In conclusion, one source of complexity we encountered in this study related to decisions of how finely to dissect the international population we analyzed. One approach, which has been used in previous studies (Grayson, 2008; Hu & Kuh, 2002; Zhao et al., 2005) is to include all international students and compare them with domestic students. Another approach is to group students by world region (e.g., European, Asian students; see J. Campbell & Li, 2008; Dowd et al., 2011; Marlina, 2009; Thomson & Douglass, 2009). In this study, we used the latter approach and included Asian students from five countries (Korea, Japan, Taiwan, mainland China, and Hong Kong) within one group of Asian international students. Although this may be a more tightly knit sample as compared to an all-international student group, the Asian international student is heterogeneous with respect to language and country of origin, as noted earlier. Students from each of the five countries may have different educational, philosophical, and socioeconomic backgrounds, which may in turn lead to different linguistic and cultural interpretations of survey items. Additional studies should be conducted to further examine within-country differences on the interpretation of survey items.

A second issue we encountered related to generalizing the results from the focus groups. Results from the Suderman (2015) study suggested that one potential source of DIF was different linguistic interpretations of terms, such as *faculty members*, which might have led international Asian students to endorse items with those terms more frequently. We suggest that further research be conducted to examine possible qualitative differences in the interpretation of test items in the NSSE with larger and more diverse groups across other participating institutions. One may also want to keep in mind that participant responses may vary depending on whether they are in a group situation rather than one-on-one for reasons such as possible reactiveness of respondents and their desire to go along with the groups' opinions. Therefore, although the focus-group approach may be helpful, it is important to use it alongside other approaches such as expert judges and think-aloud procedures.

In addition to these considerations, we also note the following limitations related to our DIF analysis. First, we note that we had small sample sizes for the Asian international group. As we sought to use the same sample as the one used in the Suderman (2015) study to examine whether focus-group interviews are a possible approach to examine sources of DIF, our sample size was limited. However, we suggest that future studies be conducted using larger sample sizes perhaps collected over test administration years, if needed. Second, our results should be interpreted with caution as our DIF analyses could be flawed because several item level analyses were conducted and there was a violation of the normality assumption of the ANCOVA approach. Finally, we note that the NSSE has short subscales. Future (simulation and applied) studies should examine the extent to which conditioning on the full scale helps reduce measurement error in the criterion or whether it is a viable option for the NSSE to increase the number of items in the NSSE subscales to encourage DIF and other research to be conducted. Despite these limitations, our study contributions are important because they serve as a model for future studies to be conducted with the NSSE as it becomes a more widely used measure of student engagement in tertiary education.

References

Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2009). Trends in global higher education: Tracking an academic revolution. *Executive report prepared for the UNESCO 2009 World Conference on Higher Education, Paris, France.* Retrieved from http://unesdoc.unesco.org/images/0018/001831/183168e.pdf

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

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education*, *19,* 115–132.

Campbell, C. M., & Cabrera, A. F. (2011). How sound is NSSE? Investigating the psychometric properties of NSSE at a public, research-extensive institution. *Review of Higher Education*, *35,* 77–103.

Campbell, J., & Li, M. (2008). Asian students' voices: An empirical study of Asian students' learning experiences at a New Zealand university. *Journal of Studies in International Education*, *12,* 375–396.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: MantelBHaenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Dowd, A. C., Sawatzky, M., & Korn, R. (2011). Theoretical foundations and a research agenda to validate measures of intercultural effort. *Review of Higher Education*, *35*, 17–44.

Eaves, M. (2011). The relevance of learning styles for international pedagogy in higher education. *Teachers and Teaching: Theory and Practice, 17*(6), 677–691.

Ercikan, K., Arim, R. G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Application of think-aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice, 29,* 24–35.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17,* 301–321.

Fritz, M., Chin, D., & Demarinis, V. (2008). Stressors, anxiety, acculturation among international and North American Students. *International Journal of Intercultural Relations, 32,* 244–259.

Grayson, J. P. (2008). The experiences and outcomes of domestic and international students at four Canadian universities. *Higher Education Research & Development, 27,* 215–230.

Hayek, J., & Kuh, G. (2004). Principles for assessing student engagement in the first year of college. *Assessment Update, 16,* 11–13.

Hu, S., & Kuh, G. D. (2002). Being (dis)engaged in educationally purposeful activities: The influences of student and institutional characteristics. *Research in Higher Education, 43,* 555–575.

International Test Commission. (2017). *ITC guidelines for the large-scale assessment of linguistically diverse populations.* Manuscript in preparation.

Karami, H., & Nodoushan, M. A. S. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies, 5,* 133–142.

Kelly, P., Moores, J., & Moogan, Y. (2012). Culture shock and higher education performance: implications for teaching. *Higher Education Quarterly, 66,* 24–46.

Kristjansson, E. A., Desrochers, A., & Zumbo, B. (2003). Translating and adapting measurement instruments for cross-linguistic and cross-cultural research: A guide for practitioners. *Canadian Journal of Nursing Research, 35,* 127–142.

Kuh, G. (2009). The National Survey of Student Engagement: Conceptual and empirical foundations. *New Directions for Institutional Research, 141,* 5–20.

Kuh, G. D., Hayek, J. C., Carini, R. M., Ouimet, J. A., Gonyea, R. M., & Kennedy, J. (2001). *NSSE technical and norms report.* Bloomington: Indiana University Center for Postsecondary Research and Planning.

Linsday, A., & Hubley, A. (2006). Conceptual reconstruction through a modified focus group methodology. *Social Indicators Research, 79,* 437–454.

Marlina, R. (2009). I don’t talk or I decide not to talk? Is it my culture? – International students’ experiences of tutorial participation. *International Journal of Educational Research, 48,* 235–244.

Martel, L. & D’Aoust, C. (2016). *Permanent and temporary immigration to Canada from 2012 to 2014*. Retrieved from http://www.statcan.gc.ca/pub/91-209-x/2016001/article/14615-eng.pdf

National Survey of Student Engagement. (2011). *Fostering student engagement campuswide—Annual results 2011*. Retrieved from http://nsse.iub.edu/NSSE_2011_Results/pdf/NSSE_2011_AnnualResults.pdf

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? *Applied Measurement in Education, 24,* 1–18.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19,* 5–15.

Roth, W. M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., & Ercikan, K. (2012, April). *Tracking sources of DIF using expert think-aloud protocols.* Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing, 3,* 129–150.

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations. *Journal of Cross-Cultural Psychology, 37,* 557–567.

Suderman, M. E. S. (2015). *Engagement for all? A study of international undergraduates at the University of British Columbia.* Retrieved from https://open.library.ubc.ca/cIRcle/collections/24/items/1.0166146

Thomson, G., & Douglass, J. (2009). *Decoding learning gains: Measuring outcomes and the pivotal role of the major and student backgrounds* (Research & Occasional Paper Series: CSHE.5.09). Berkley: University of California.

Wong, J. K. (2004). Are the learning styles of Asian international students culturally or contextually based? *International Education Journal, 4*(4), 154–166.

Zhao, C., Kuh, G., & Carini, R. (2005). A comparison of international student and American student engagement in effective educational practices. *The Journal of Higher Education, 76*, 209–231.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45–79). Netherlands: Elsevier Science B.V.

Note

[1] Note that by classifying Asian students into a single group, we may have lost unique-to-country variations within the heterogeneous Asian international student group. Similar classifications have been suggested in previous research (e.g., Fritz et al., 2008). We provide further suggestions for conducting finer-grain size classifications in the discussion section for future research.