Detecting Items That Function Differently for Two- and Four-Year College Students

Amy Thelk
James Madison University

Abstract

Differential Item Functioning (DIF) occurs when there is a greater probability of solving an item based on group membership after controlling for ability. Following administration of a 50-item scientific and quantitative reasoning exam to 286 two-year and 1174 four-year students, items were evaluated for DIF. Two-year students performed better-than-expected on 13 items and worse than expected on 10 items. Reasons for DIF are explored, along with the importance of conducting this type of study.

Introduction

As institutions commit to greater assessment activities on their campuses, the search for appropriate instrumentation ensues, especially in the measurement of student learning. Assessment professionals may opt to adopt or adapt an exam that was developed at another site to gauge student learning at their institutions. When using an exam developed at one location to assess students at a different school, the expectation is that any set of students with the same ability should perform about the same on a given test item. However, due to other factors, like on-campus culture, socioeconomic differences, and variations in exposure to material, student scores may diverge despite similar ability. Examination of differential item functioning (DIF; Hambleton, Swaminathan & Rogers, 1991) can inform consumers of tests about whether factors other than ability affect test scores.

For the community college and the 4-year institution that served as sites for this study, assessment has been incorporated into their academic schedules; students are aware of mandated testing at the time of application. Additionally, a professional partnership exists between the two schools: the four-year school serves as a transfer site for the community college, and some of the instruments developed at 4-year school are leased out to the community college.

Scientific and Quantitative Reasoning Assessment

The scientific and quantitative reasoning instrument (SR/QR) used for this study had been developed over the course of several years at the four-year institution. The items had been crafted by faculty experts in science and mathematical disciplines with the assistance of measurement experts.

At both institutions that served as data collection sites, dedicated "assessment days" were held during the spring semester; classes were cancelled for the day so that students participate in the required testing without potential schedule conflicts; the data used for this research were collected during such assessment days. For this study, both institutions administered the same version of the SR/QR.

Differential Item Functioning (DIF)

When students have the same ability level, the probability of solving a given item correctly should be the same for any student. However, sometimes factors other than ability are influential upon the score: access to information, language skills and testing conditions, for example. If different groups comprise the test-taking population, a DIF study can be designed and implemented. For this study, the data set was divided into two groups, 2-year-school students and 4-year-school students. Hambleton, et al (1991). provide a concise and useful definition of DIF: "An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting an item right" (p. 110). Figure 1 further illustrates DIF.

DIF is instrumental in alerting test users to the possible presence of bias at the item level. The presence of DIF is a necessary component of bias, although not sufficient in itself to deduce that bias is present. If DIF is found, further investigation must take place to determine whether the differences in performance on these items are due to unfairness. A somewhat less alarming situation would be the case of an item showing DIF because students in that group have not had course exposure that would assist

in solving the item successfully. In any case, a DIF analysis can provide preliminary evidence about the degree to which certain test items are biased for or against particular groups.

Method

For both institutions, testing was mandated for students and held on designated days on which classes were cancelled. Two-hundred eighty-six community college students and 1174 four-year college students participated in testing, yielding the data used for this project. The raw data were scored for each group and the two data sets were concatenated following the addition of a group-ID variable. To determine which items on the SR/QR demonstrate DIF between the community college and four-year college groups, item parameters were first estimated by item response theory (IRT). DIF was then calculated using these item parameters.

In IRT three main models are used to estimate item parameters. These models are, in order of complexity, the one-parameter logistic model (1-PL), 2-PL and 3-PL (Hambleton et al., 1991). Researchers decide which model is most appropriate for their studies by considering the sizes of their samples and evaluating model fit. The 1-PL only takes item difficulty into consideration, the 2-PL takes difficulty and discrimination into account, and the 3-PL models item difficulty, discrimination, and guessing. The first parameter is b (item difficulty), the second is a (item discrimination) and the third is c (guessing). As a general rule of thumb one should not apply a 1-PL model unless the sample has at least 200 participants. Four hundred and 1,000 are the suggested sample size minimums for the 2-PL and 3-PL models respectively. The size of our sample (1173) and the nature of our data (multiple choice items with a variety of difficulty and discrimination levels) suggested that a 3-PL model would be a logical starting point, and an analysis comparing the 1-PL, 2-PL, and 3-PL models confirmed that the 3-PL model did indeed result in the best fit.

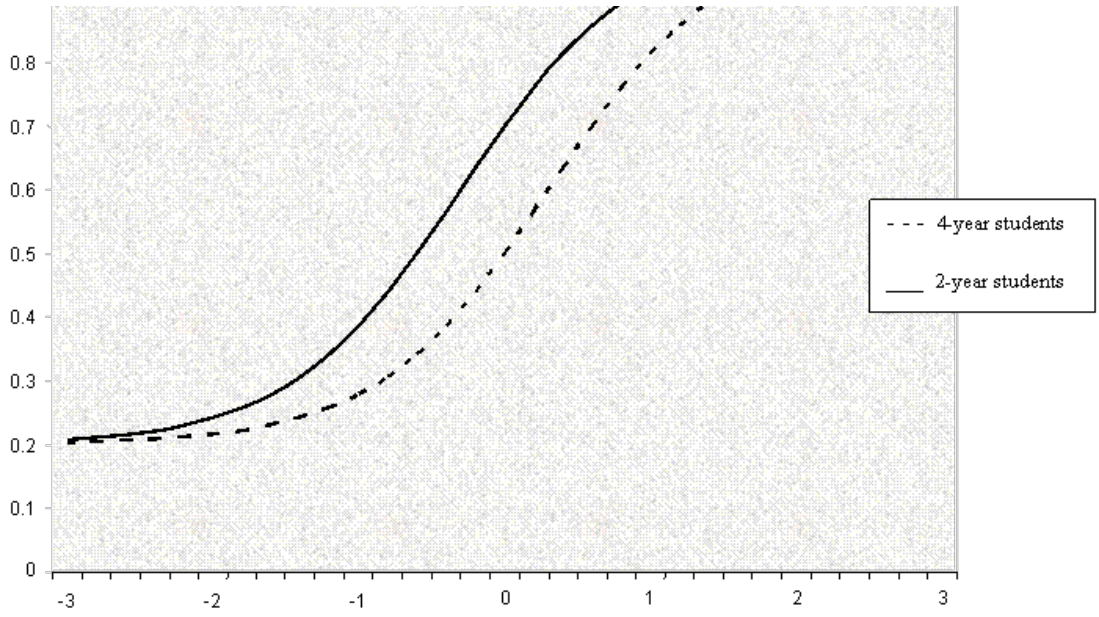

Figure 1. Example of an item showing DIF between two-year students and four-year students.

In IRT, ability (denoted by θ) is measured on a scale with 0 representing average ability and with each point above or below representing a standard deviation. For example a score of "+1" would represent ability at one standard deviation above the average and a score of "-2.5" would represent ability at two-and-a-half standard deviations below the average. The b parameter reflects at what ability level 50 percent of test takers get the item correct. When these values are aggregated over items and averaged, the result is the difficulty value for the entire test.

For this study, two methods of detecting DIF were employed. The first uses IRT to determine whether the item response characteristics look different across testing groups (Hambleton et al, 1991). Es-

sentially, a null hypothesis is being tested to determine whether there are significant differences when groups are compared.

Using the output generated by BILOG-MG (Zimowski, Muraki, Mislevy & Bock, n.d.), the appropriate values were input into the equation $(b_1 - b_2)/ \sigma bdiff$, where b_1 and b_2 are the difficulty values for groups 1 (community college students) and 2 (four-year students), respectively, and $\sigma^b diff$ is the standard error of the difference between the two b values in the numerator. The solution to this equation is distributed as a z-score (M=0, SD=1). Based on the results of the equation above for each item, any items with an absolute value z-score greater than 2.58 (corresponding to a two-tailed $p \leq .01$) were pulled out to examine for DIF, since these z-scores flagged a significant difference between the b values between groups for those item.

The second method involved the calculation of Mantel-Haenszel (M-H; Hambleton et al, 1991) statistics for each of the items that exhibited high absolute z-scores to confirm the presence of DIF. The M-H value is a common-odds ratio that represents a proportion with Group 1 in the numerator and Group 2 in the denominator. For this research, when this value was greater than 1 then the item favored Group 1, and when the value was lower than 1 it favored Group 2.

To determine effect sizes of the difference between difficulties, delta (Δ) values were evaluated. Delta values are calculated by locating the odds-ratio, or α, value on the output resulting from the M-H procedure, and substituting that value into the equation $\Delta = -2.35 \ln(\alpha)$.

Based on the Educational Testing Service scale for effect size, these Δ values are classified into A, B and C categories (Dorans, 1989). If the absolute value of Δ is less than 1, the magnitude of the effect is negligible; this is considered an "A" item. When the absolute value of Δ is between 1 and 1.5, the item is placed in the "B" category. Items that show the most DIF have an absolute Δ value greater than 1.5; these items are placed in the "C" category.

Results

Out of the 50 SR/QR items, 23 items had high absolute z-scores. Using M-H statistics, the presence of DIF was confirmed in all of these items, and the group the item favored was ascertained. Out of the 23 items that showed DIF, 13 of the items favored the two-year college, while 10 favored the four-year school. Calculation of effect sizes revealed that 22 out of 23 of the items were placed in category C connoting the highest amount of DIF. Table 1 provides a summary of these results.

A review of the item content revealed that the items biased in favor of the community college students pertained to higher order reasoning skills such as evaluating a claim or ascertaining the relationship between variables by interpreting a graph. In other words, controlling for ability, community college students did better than expected on these items. Many of these positively biased items were also part of testlets. Testlets are two or more items related to a single stimulus. Conversely, the items that two-year students missed more than expected controlling for ability (i.e., biased against the community colleges) were those related to performing routine algorithms.

Discussion

According to Anderson and Sundre (2005) examining DIF between two-year and four-year students is important because many assessments used by two-year institutions were developed for and normed on four-year students. When selecting an established instrument, colleges will want to review the fit of the items to the institution's objectives. However, exploring DIF after initial use of the instrument will assist with identifying items that have more subtle problems associated with bias. It is worth noting that just because an item favors the community college group does not necessarily mean that this group scored higher on that item. Indeed, for many items the two-year students still scored lower, but they did not score as low as expected.

Table 1

Results from both DIF Analyses (IRT and M-H) for Items Showing DIF

Item #	Group 1 *b* value	Group 2 *b* value	*b* difference	z-score	MH Odds Ratio	Group favored	Delta (Δ)	Category
31	0.524	3.500	-2.976	-9.537	0.173	1	4.120	C
10	-0.232	2.482	-2.713	-7.713	0.234	1	3.413	C
37	-0.384	2.248	-2.633	-6.169	0.264	1	3.130	C
22	0.957	2.669	-1.712	-5.388	0.390	1	2.212	C
40	0.909	8.882	-7.972	-4.773	0.120	1	4.981	C
24	-1.414	0.408	-1.822	-3.861	0.359	1	2.405	C
35	1.964	6.344	-4.380	-3.417	0.369	1	2.344	C
42	-0.949	0.575	-1.523	-3.246	0.437	1	1.944	C
26	-1.297	6.269	-7.565	-3.236	0.526	1	1.509	C
32	-0.018	1.235	-1.253	-3.156	0.467	1	1.791	C
14	-0.088	2.480	-2.569	-3.095	0.498	1	1.638	C
36	0.000	2.577	-2.577	-2.773	0.512	1	1.571	C
20	0.709	1.941	-1.232	-2.751	0.554	1	1.386	B
38	-0.781	-2.227	1.447	2.795	2.490	2	-2.144	C
21	0.585	-1.333	1.918	3.172	2.239	2	-1.895	C
30	0.688	-1.334	2.022	3.308	2.021	2	-1.654	C
39	0.763	-2.180	2.942	4.029	3.565	2	-2.987	C
43	-0.247	-2.005	1.759	4.087	3.079	2	-2.642	C
25	-0.589	-2.573	1.985	4.631	4.478	2	-3.523	C
50	11.551	-1.320	12.871	5.161	479.799	2	-14.507	C
7	8.957	-0.477	9.434	5.541	38.276	2	-8.565	C
47	3.816	-1.728	5.544	8.024	38.087	2	-8.554	C
45	0.939	-2.032	2.971	8.261	11.198	2	-5.677	C

Note. *Z*-scores ≥ |2.58| suggest DIF; **Δ** values greater than 1.5 signify high amount of DIF (category C). Categories based on Educational Testing Service classification (Zieky, 2004)

As mentioned earlier, when different groups have unequal probabilities of getting a test item correct after controlling for ability, DIF is present. Indeed, in this study many items, almost half of the total, showed DIF for and against community colleges students. They performed better than expected on 13 items and worse than expected on 10 items.

While reviewing these items, the author speculated about what factors may have contributed to DIF. Since many of the items are one part of a testlet, it is conceivable that community college students are more persistent and less likely to get bored or fatigued, and therefore do not skip items or answer carelessly as often. Persistence across groups may be worthy of future investigation.

Another factor is that these two groups represent two very different institutions, with varying curricula and objectives. So in some cases the two-year group may have actually covered certain material to a greater extent than the students at the four-year school and less of other curricular components. Relatively speaking, perhaps the community colleges spent more time on the reasoning components of science and less on applying algorithms. A counter argument is that reasoning may be acquired outside of traditional classroom learning. Given that these community college students were older and likely have had a wider array of experiences, this scenario cannot be ruled out. Such a situation would illustrate Messick's (1995) concept of construct irrelevant variance: performance on test items is due to an influence outside of the learning arena at which the instrument is aimed.

Following the administration of this test, new items were introduced to the SR/QR test form, while some of the previous items were removed due to low scoring or inappropriateness to the curriculum. Items showing DIF that were not removed were retained on a provisional basis, with the test's advisory team committing to continually analyze test data to determine the appropriateness of including these items on later versions. If these changes in the exam had not been made for the community college group, the results of this DIF study would have presented great urgency for further test review before using the exam in its original state for this population.

Summary

When performance on items is different than expected for a group, DIF is present. This article describes how DIF was indentified when comparing results of two and four-year students on the same test, and explored reasons for its presence.

As testing for the purpose of gauging student learning becomes more common, many postsecondary schools will find themselves in need of already developed instruments that are appropriate for their own testing programs. This DIF study serves as an important cautionary reminder about comparing test results of two different groups of students. Since students are exposed to a variety of instructional styles, classroom sizes, and campus cultures, it is unlikely that their performances on test items will be similar, even after controlling for any differences between the groups in overall test score. So by gathering information about differential item functioning, more appropriate comparisons can be made between or among groups.

A DIF study is a useful way to determine whether test items created for one student group yield comparable information when administered to another group. The analysis is relatively quick and only requires a data set for each diverse group, but the information that is produced is essential to the validity of the scores generated by the assessment. If the students in your school are not performing as expected as indicated by DIF, then the validity of the inferences made by the test scores, particularly comparisons among groups of students, are likely suspect.

References

American Association of Community Colleges. (2005). Fast facts: Community college fact sheet. Retrieved May 19, 2005, from: http://www.aacc.nche.edu.

Anderson, R. D., & Sundre, D. L. (2005). Assessment partnership: A model for collaboration between two-year and four-year institutions. *Assessment Update*, *17*(5), 8-16.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*(3), 216-233.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

National Center for Education Statistics. (2003). Retrieved May 18, 2005, from http://nces.ed.gov/programs/coe/2003/section5/indicator32.asp.

Zieky, M. (2003). *A DIF Primer.* Princeton, NJ: Educational Testing Service.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (n.d.). BILOG-MG [Computer software]. St. Paul, MN: Assessment Systems Corporation.