4-2004

Item Parameter Drift: The Impact of the Curricular Area

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Item Parameter Drift: The Impact of the Curricular Area

Christine E. DeMars

James Madison University

(2004, April). Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Abstract

The items from tests from two content areas, information literacy and global issues, were examined for item parameter drift across four years. The items on the information literacy test were expected to show more drift because the content of this field is changing more rapidly and because the test changed from low to high stakes for students while the other test remained low stakes. More items did show drift on the information literacy test, but the drift was not always readily explained. Further, some items did not fit the drift model available in BILOG-MG, either because the drift was a one-time shift rather than a gradual change or because both the discrimination and difficulty changed over time.

Item Parameter Drift: The Impact of the Curricular Area

Purpose

In item response theory (IRT), item parameter drift occurs when the difficulty or discrimination of an item systematically changes over time. Drift impacts the ability estimates and creates complications for equating. Drift might be especially prevalent in rapidly changing subject fields. Easy items could become more difficult as the knowledge or skill tapped by the items becomes less common, and difficult items could become easier as formerly specialized knowledge becomes more generally known. In this study, drift was examined across four years on an information literacy test and a global issues test. These tests differed in content and context; after identifying drift items I attempt to explain how these factors might influence drift.

Framework

Item drift found in other studies has sometimes been attributed to shifts in curricular content. Sykes and Fitzpatrick (1992) found that items in one of four content areas on a licensure exam were more likely to show shifts in difficulty than items in the other content areas. Similarly, Bock, Muraki and Pfeiffenberger (1988) suggested curricular differences could be responsible for the item difficulty drift they found in some items on a College Board physics test. Over the course of 10 years, basics mechanics items became easier relative to the other items, which was consistent with a curriculum survey that showed this area was the most heavily covered. An item with metric units became easier over time, as an item with English units became more difficult. In the present study, the content of items will be explored as a possible explanation for drift.

Method

Instruments

The Information Seeking Skills Test (ISST) is a test of information literacy for college students. Reference librarians, working with faculty, developed the curricular objectives, classroom

instruction and on-line tutorials to teach those objectives, and the ISST to assess whether the objectives were met. Students at James Madison University are required to pass the test during their first year; the passing score was set by general education faculty using the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001) and was put into effect between the first and second years the test was used. It contains 53 multiple-choice items about using and evaluating information sources, including paper and electronic library materials. Some items require students to look up information in databases or evaluate web pages. Though multiple test forms have since been developed, a single test form (with the exception of one item) was used for the first four years. The test is administered by computer, in a proctored lab, to facilitate access to electronic sources as well as to make the test available on-demand.

The Global Experience test was developed to assess knowledge of global issues. The test covers global political, social, cultural, and economic issues addressed by the common objectives taught in several courses designed to meet the university's Global Experience requirement. The test has 32 multiple-choice items and is administered by paper and pencil. Students are not required to pass this test and it does not count toward course grades; rather, a random sample of students takes the test each year on a day when all students with 45-70 credit hours participate in assessment. All students participate in testing that day, but they are randomly assigned to different tests. Students are informed that the results are used for program evaluation, not individual accountability.

The information literacy test assessed knowledge of a variety of information sources, including electronic databases, internet sources, and traditional brick-and-mortar library sources. Due to rapid changes in the availability and use of these information sources, it was hypothesized that some of the test items would exhibit parameter drift. Further, during the years the data were gathered, information literacy instruction was expanded and refined, with both the instruction and the test designed to address objectives developed by reference librarians and faculty. To the extent

that this instruction impacted some items more than others, this could also cause parameter drift. Finally, after the first year the test was implemented, students were required to earn a passing score. If these increased stakes led to higher performance on all items, this would not lead to parameter drift, but if some items, perhaps the more time-consuming items, were affected more than others, drift could result. The global issues test, in contrast, assessed a field that was not changing as rapidly. Further, changes in instruction were expected to have less of an impact. While the test and the curriculum were linked by the same objectives, the objectives were quite broad and could be met by any of five courses in anthropology, economics, geography, political science, or sociology. Also, the test was administered in a low-stakes context throughout the four years studied. Thus, the information literacy test was expected to show more parameter drift than the global issues test.

Data Sources

Data were available from 8721 students who took the ISST in the 1998-1999, 1999-2000, 2000-2001, or 2001-2002 academic years (353 students in year 1, 2671 in year 2, 2741 in year 3, 2956 in year 4). The ISST can be taken multiple times if needed; only first-time examinees were used in this study. The 1998-1999 group was the final pilot sample of students; they were not required to obtain passing scores on the test as the successive groups were. Data were available from 2361 students who took the Global Experience test in the spring of 2000, 2001, 2002, or 2003 (324, 356, 743, 938 students in years 1-4, respectively). This test was used only for program evaluation, not student accountability, all four years.

Item Calibration

Item parameters and drift were estimated using BILOG-MG 3.0 (Zimowski, Muraki, Mislevy, & Bock, 2003). Each academic year was treated as a test-administration group. A 3-parameter logistic model was used, with the discrimination and lower-asymptote parameters held constant over time points (the standard drift model in BILOG-MG) and linear trends estimated for

the difficulty parameters. One of the Global Experience items was eliminated from the analysis, leaving 31 items, because the parameters for this item did not converge in 200 iterations (this item had a slightly negative item-total correlation). Items were discarded from the analysis if they did not fit the model at the .0001 level, based on the log-likelihood χ^2 printed in the output. The .0001 alpha level was selected because the log-likelihood χ^2 tends to produce high levels of Type I error with short tests or large examinee samples (Orlando & Thissen , 2000; Glas & Suárez Falcón , 2003). Nine of the ISST items were eliminated due to misfit, leaving 44 items with adequate fit. The Global Experience items showed adequate model fit.

Results

Linear Parameter Drift

On the ISST, using a significance level of .01, 23 items showed significant linear trends; 11 of these items became easier and 12 became more difficult. Some of these significant differences were quite small, though. Using a criterion of a change in difficulty averaging at least 0.2 per year as well as statistical significance to flag items that showed meaningful amounts of drift, five items were flagged as becoming easier and one was flagged as becoming more difficult. Item characteristic curves are shown for these items in Figure 1. Item 29 showed an average difficulty change of -1.84 per year, from 4.18 to -1.37 over the four years. This change was readily explained; the reference librarians had been surprised that this item, which asked how to determine whether the library carried a periodical, was so difficult and they began emphasizing this content in their instruction and on-line materials. Evidently, their efforts were successful. The item that became more difficult (.40 per year, item 5) and one of the items that became easier (.22 per year, item 4) both gave students a reference and asked them to identify the type of reference it represented (book, government document, etc.). Both items had been disclosed to some faculty between years 3 and 4 during an informational session about the test. There was no obvious reason one item became

harder and the other grew easier. The other three items that became easier (.22, .21, and .20 per year, items 33, 42, and 22) involved locating a book review in an electronic database, identifying the credentials of a web page author, and finding a book's call number using the library's electronic database. This might suggest that students became more knowledgeable about electronic resources, but it should be noted that similar items showed little change over the years.

On the Global Experience test, three items showed significant linear trends at the .01 level; all three items became more difficult. Two of these items showed drift of at least .20 per year and are plotted in Figure 2. Item 15 increased from a difficulty value of -0.07 to 0.69 over the four years. This item reads: "Which of the following countries would you examine if you were studying the problems involved in moving from a centrally planned economy to a market economy?", with a correct answer of "Russia". While still relevant, this item could logically be seen as less salient to successive cohorts of students. There was no obvious reason why the other item was becoming more difficult.

It is not surprising that some items showed drift for no apparent reason. This is a common situation in DIF studies as well. For example, Skaggs and Lissitz (1992), discussing a set of items identified for gender DIF, noted "an examination of the content of these six items, is not particularly revealing, in our opinion, about why they are biased against males or females. In fact, one of the items consistently identified as biased favoring males shares the same stimulus as one of the items favoring females" (p. 235).

Misfitting Items

As noted in the Method section, nine ISST items were omitted from the drift analyses due to misfit. Misfit could occur because the item did not fit the 3PL model within one or more years, or because the item did not meet the equal slopes constraint of the drift model or because the difficulty shift did not occur in a linear fashion. To check for the latter possibilities, items were also calibrated

separately within each year with only the lower asymptotes constrained to be equal. All nine of the ISST items which misfit the drift model had acceptable fit in the individual years. The scales were equated and put on the metric set in the first year, and the ICCs were plotted graphically to compare how they changed over time.

In Figure 3, Items 17, 27, and 53 showed drift between the 1st and 2nd years, followed by little change. The linear drift procedure models drift as changing in constant increments, so this pattern could explain the poor fit for these items. The difference in item parameters, then, might be due to the context shift from a low-stakes context in year 1 to a high-stakes context in years 2, 3, & 4 rather than to gradual changes in technology or curriculum. Further, item 17 seemed to show drift in discrimination, which was not modeled.

Item 52 became easier in year 2 than year 1, but the difficulty for years 3 and 4 was between that of years 1 and 2. For item 25, the difficulty for years 2 and 4 was between that of years 1 and 3. These patterns would seem to be chance instability rather than drift.

Items 14, 28, 32, and 51 showed drift in discrimination. Item 28 is a difficult item that shows unpredictable shifts in both discrimination and difficulty. Because this item was so difficult, the standard errors for the parameter estimates were large and thus the shifts in parameters (and resulting misfit) may not be reliable.

Conclusions

This study illustrates the importance of checking for parameter drift; parameters may not be invariant over time. The technology test had more items impacted by drift than the global issues test, as expected. However, the changes found could not always be readily explained by curriculum, technology, or context. This parallels the findings for DIF; bias review panels often can not explain why one item shows DIF and similar items do not (Angoff, 1993; Engelhard, Hansche, & Rutledge, 1990; Ramsey, 1993; Shepard, Camilli, & Williams, 1984; Skaggs, & Lissitz, 1992). This does not

mean that the drift can be ignored; the effects of drift on the measurement scale need to be taken into account as new items are pilot-tested and calibrated to the existing scale. Drift can also be used as a starting point to flag items that ***might be*** becoming less relevant, just as DIF is often used as a starting point to flag items that ***might be*** assessing an irrelevant construct that disadvantages one gender or ethnic group.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presences of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*, 347-360.

Glas, C. A. W., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Ramsey, P. A. (1993). Sensitivity review. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, L., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93-128.

Skaggs, G., & Lissitz, R. W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement, 29*, 227-242.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT *b* values. *Journal of Educational Measurement, 29*, 201-211.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3.0* [computer software]. Lincolnwood, IL: Scientific Software International.

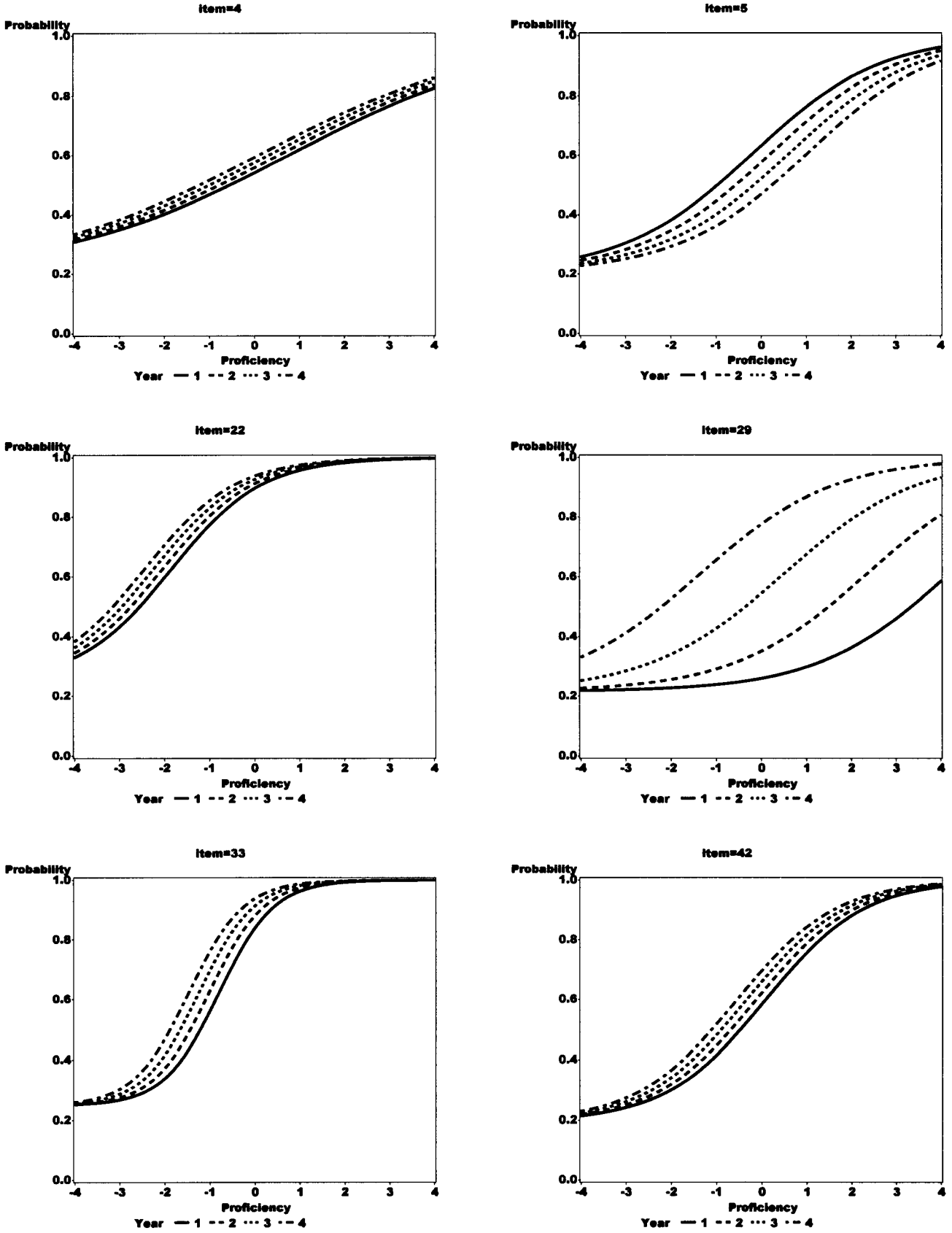


Figure 1: *ISST items*

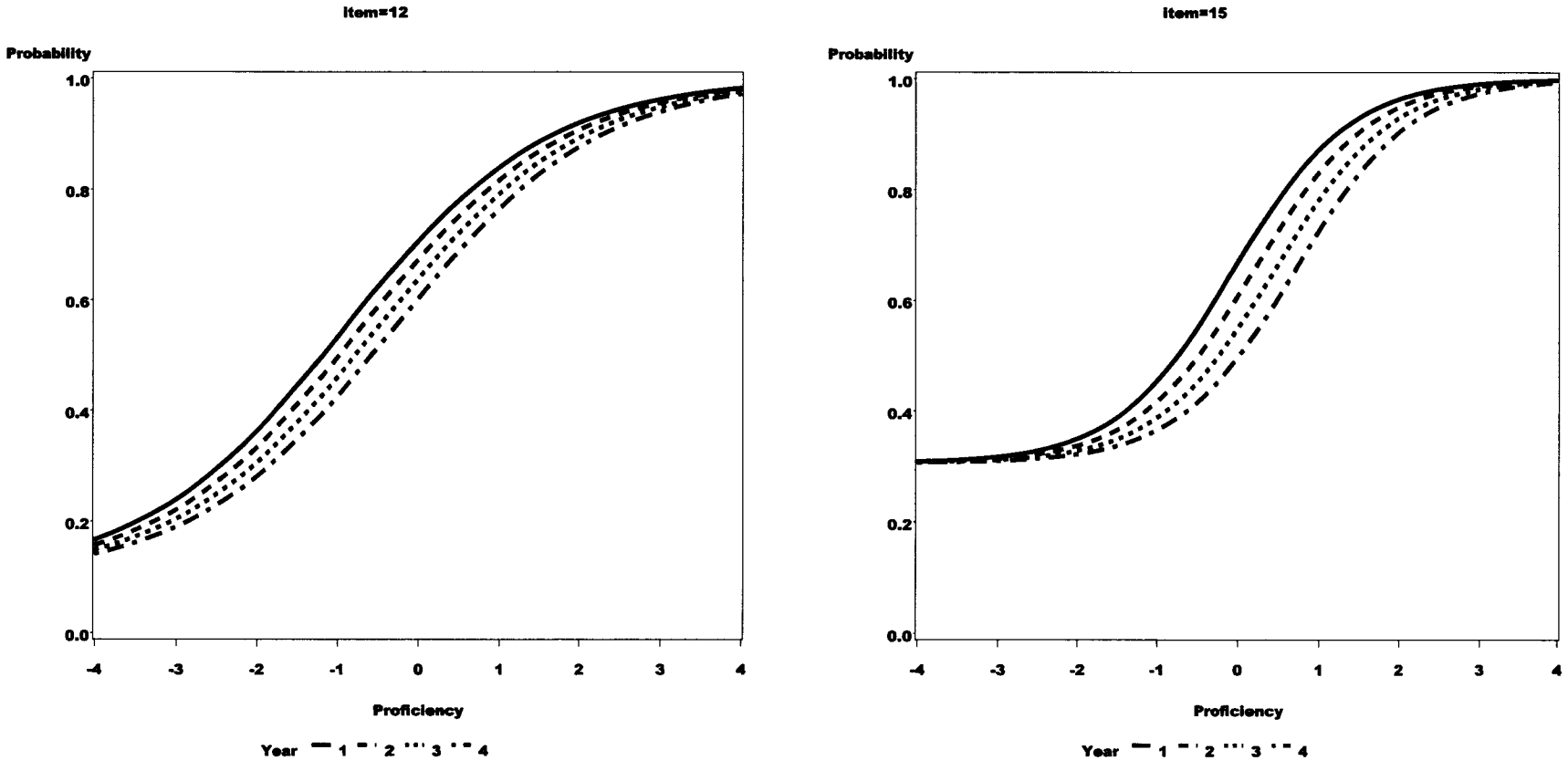


Figure 2: *Global Experience items*

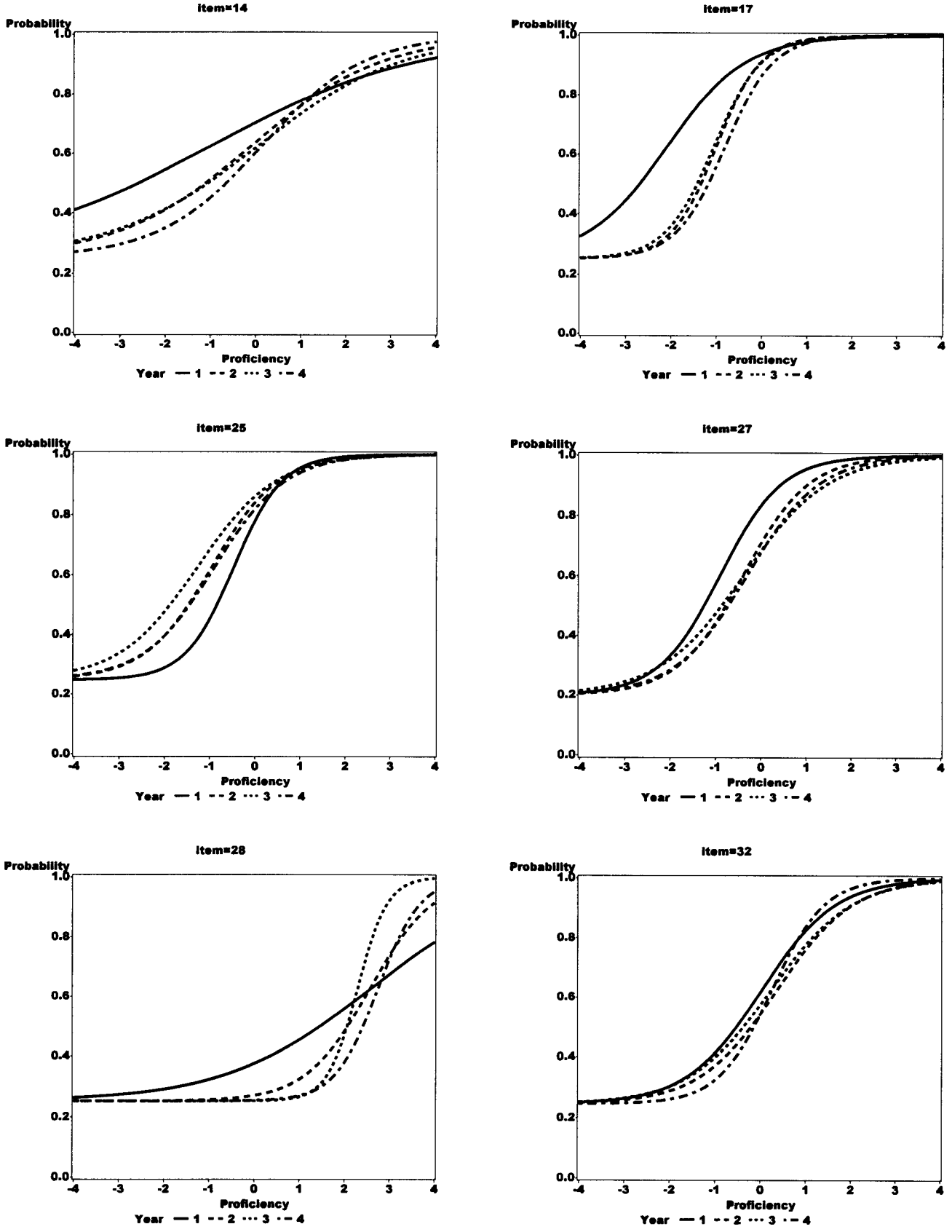
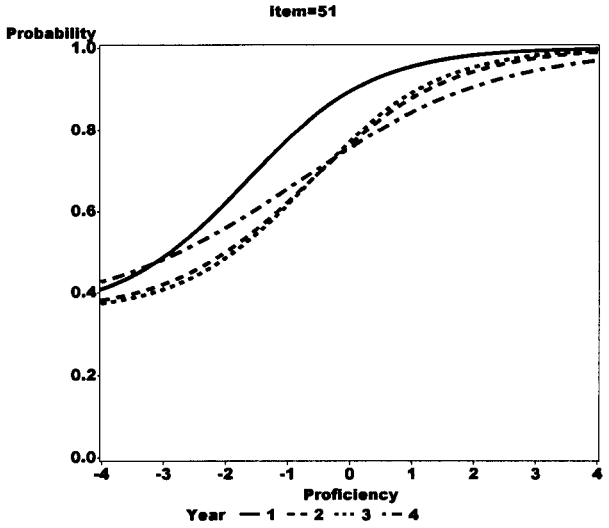
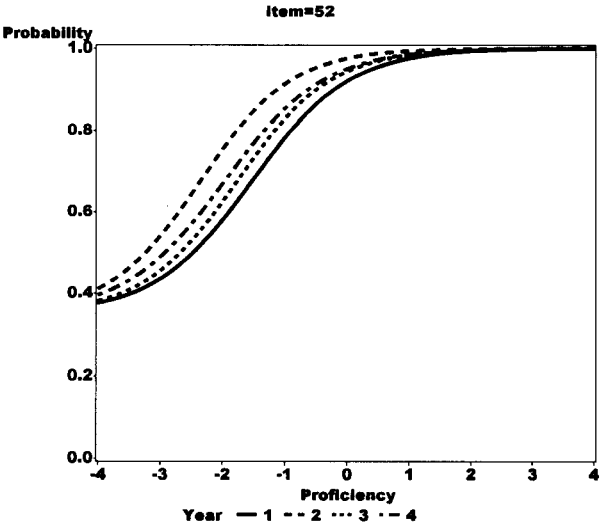


Figure 3: Misfitting Items





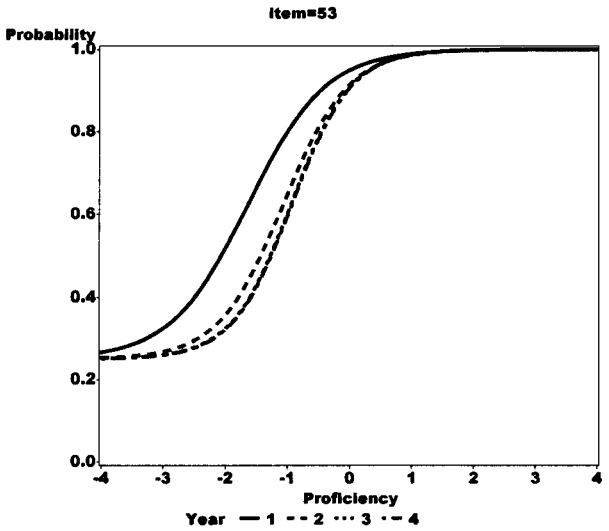


Figure 3 (continued)