DOCUMENT RESUME

ED 481 656 TM 035 357

AUTHOR Witt, Elizabeth A.; Stahl, John A.; Bergstrom, Betty A.; Muckle, Tim

TITLE Impact of Item Drift with Non-Normal Distributions.

PUB DATE 2003-04-00

NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC01 Plus Postage.

DESCRIPTORS *Item Response Theory; *Statistical Distributions; *Test Items

IDENTIFIERS *Item Parameter Drift; Nonnormal Distributions; *Rasch Model

ABSTRACT

The focus of this simulation study was to investigate the effects of item difficulty drift on the stability of test taker ability estimates and pass/fail status under the Rasch model. Real, non-normal distributions of test taker abilities and item difficulties were used to represent true parameters. Test taker responses for 18 conditions of item drift were simulated; ability estimates were obtained using the WINSTEPS program (J. Linacre, 1999) and compared with baseline data. Results are encouraging in that they provide further evidence of the robustness of Rasch model ability estimates in the face of undetected item drift. A fairly large number of item difficulties (e.g. 25%) must be altered before even a hint of possible distortion of ability estimates appears. (Contains 7 tables and 12 references.) (SLD)

IMPACT OF ITEM DRIFT WITH NON-NORMAL DISTRIBUTIONS

Elizabeth A. Witt
John A. Stahl
Betty A. Bergstrom
Tim Muckle

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 2003.
Correspondence should be sent to Elizabeth Witt, American Board of Emergency Medicine, 3000 Coolidege Road, East Lansing, MI 48848, e-mail EWitt@ABEM.com, or to John Stahl, Promissor, 1007 Church St., 7th Floor, Evanston, IL 60201, e-mail John_Stahl@Promissor.com.

INTRODUCTION

The invariance of item parameters is essential to the validity of many IRT procedures, including equating and adaptive testing (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). Unfortunately, item parameters are sometimes found to "drift" over time due to a variety of factors (Bergstrom, Stahl, & Netzky, 2001). Curriculum changes can result in a shifting of content emphasis, effectively making some items easier and others more difficult (Wells, Subkoviak, & Serlin, 2002). Items may be overexposed due to heavy usage or cheating (Stahl, Bergstrom, & Shneyderman, 2002; Cizek, 1999). Changes in laws or policies can affect item difficulties, especially in licensure and certification testing. Even widely publicized historical events can result in a shifting of item parameters. Consider, for example, an item testing the definition of the word "chad" before and after the 2000 U.S. presidential election.

Previous research has found evidence that item drift does occur in a variety of assessment situations (Chann, Drasgow, & Pfeiffenberger, 1988; Goldstein, 1983; Mislevy, 1982). To date, however, research investigating the impact of item drift on test taker measures such as ability estimates and pass/fail status has not been extensive. Wells, Subkoviak, and Serlin (2002) found a small effect on ability estimates when drift was simulated for the *a* and *b* parameters under the two-parameter logistic model. Stahl, Bergstrom, and Shneyderman (2002) simulated drift under the Rasch model. They found only a negligible shift in ability estimates, and errors in pass/fail status occurred at rates comparable to those resulting from measurement error, suggesting that Rasch model equating is robust in the face of item difficulty drift.

The Wells, et al. (2002) study simulated item parameters based on distributions from real-life data. Ability estimates, on the other hand, were simulated according to a normal distribution. Stahl et al. (2002) simulated both item and person parameters as normal distributions. Yet there are many assessment situations, particularly in the context of licensure and certification testing, in which non-normal distributions are common. This is especially true of test taker abilities. A test developer can exercise some control over the distribution of item difficulties in building an exam. The distribution of abilities, however, is governed by factors related to the self-selection of test takers, factors that vary from one profession to another.

The purpose of this paper is to investigate the effects of item drift on test taker ability estimates using realistic distributions of both test taker ability and item difficulty. Non-normal distributions of these

measures are commonly observed in many assessment contexts. In part, this study replicates the Stahl et al. (2002) study, examining the impact of item drift on Rasch ability estimates and pass/fail status, but this investigation begins with "true" distributions of ability and item difficulty that are modeled on empirical data from typical credentialing exams.

METHODOLOGY

The focus of this simulation study is to investigate the effects of item difficulty drift on the stability of test taker ability estimates and pass/fail status under the Rasch model. Real, non-normal distributions of test taker abilities and item difficulties were used to represent true parameters. Test taker responses for 18 conditions of item drift were simulated; ability estimates were obtained using WINSTEPS (Linacre, 1999) and compared with baseline data.

Approximately 50 empirical distributions of item difficulty and test taker ability from an assortment of certification testing programs were examined to determine what types of distributions typically occur in real life testing situations. Both difficulty and ability were estimated under the Rasch model and expressed as logits. Although normal distributions on both measures were found for some programs, test taker abilities were often skewed, and item difficulties often differed from the normal distribution in skew and/or kurtosis. For abilities, significant skew occurred in both directions, positive and negative, with a negative skew being slightly more common. Negatively skewed distributions were much more common for item difficulties. This makes sense, as most licensure and certification exams include a few items covering essential content that all test takers are likely to know. These items would represent the lower tail of the difficulty distribution.

Two commonly occurring combinations of non-normal distributions were selected to represent true item difficulties and test taker abilities from which the simulated test taker responses would be generated. The first was a 100-item test administered to 187 test takers. Abilities were negatively skewed, while item difficulties were normally distributed (not significantly different from normal). The second distribution selected was a 200-item test administered to 260 test takers. Abilities were negatively skewed, and item difficulties exhibited a negative skew and a positive kurtosis. Descriptive statistics for both distributions are shown in Table 1.

To establish a baseline, for each of these two combinations of "true" ability and difficulty distributions, 100 sample matrices of test taker responses to items were simulated using *Promissor Test Simulator*

(Schnipke, D., Popkins, N., Kinsella, K & Stahl, J., 2001) and test taker abilities were estimated under the Rasch model via WINSTEPS (Linacre, 1999). These represent sample distributions of ability estimates that might be expected to occur under conditions of no item drift. Pass/fail standards of 1.65 logits for the 100-item test and 1.16 for the 200-item test were selected to simulate the pass rates typical in certification testing (approximately 70%). Baseline data was characterized in terms of correlations between "true" abilities and the ability estimates resulting from the 200 simulations (100 each for the 100-item and 200-item test) and also in terms of number of test takers misclassified according to pass/fail status among the simulated samples.

True item difficulty distributions were then modified to simulate 18 conditions of item drift for each of the two tests. The difficulty of items was shifted by +.10, +.25, +.50, -.10, -.25, and -.50 logits. Positive shifts represent items becoming more difficult; negative shifts represent items becoming easier. The percentage of items affected by drift was simulated for 5%, 10%, and 25% of the items at each of the six magnitudes of drift. Because the results of previous research (Stahl, Bergstrom, and Shneyderman, 2002) suggest that mixing positive and negative drift within a test is less likely to result in informative effects, mixed drift was not simulated in this study. For each condition, the appropriate percentage of items was randomly selected, and the logit difficulties of these items were manually altered.

The resulting 36 datasets (18 for each test) served as input item files to generate simulated response string matrices using *Promissor Test Simulator* (Schnipke, et al., 2001). Rasch ability estimates were then calculated via WINSTEPS (Linacre, 1999) for each set of simulated data. In this step of the analysis, items were anchored at their original, non-drifted difficulties in order to model a realistic situation in which item difficulties have changed but the testing organization is not yet aware of any drift. In computer based testing, in order to provide immediate, on-site scoring, the test is commonly built with pre-calibrated items and equated using the stored item statistics. Where items have actually become easier, test takers have a higher probability of answering correctly. Their ability estimates are thus inflated because the stored item difficulties are now higher than the true, drifted difficulties. The opposite effect occurs when items drift in a harder direction; ability estimates are deflated because the stored item difficulties are now lower than the true ones.

The effect of item drift on ability estimates was investigated, first, by examining correlations and mean differences between true abilities and the abilities estimated under conditions of drift and, second, by noting the number of pass-fail misclassifications in comparison with the baseline data. Some simulated test takers' true abilities fell below the passing standard, yet their estimated abilities were high enough to

allow them to pass; these are designated false positives. Conversely, some had true abilities above the passing standards, but their estimated abilities would result in failing the exam; these are the false negatives.

RESULTS AND DISCUSSION

Baseline data were summarized first to provide a picture of the relationship of sample ability estimates to true abilities under no-drift conditions as well as the number of misclassifications that might be expected to occur due to measurement error alone. Among the 100 baseline samples simulated for the 100-item exam, correlations between true and estimated abilities ranged from .813 to .943, with a median of .922. The number of false positives among the 187 simulated test takers ranged from 5 (2.7%) to 19 (10.2%); the number of false negatives ranged from 3 (1.6%) to 12 (6.4%). The total number of misclassifications for any one sample ranged from 9 (4.8%) to 27 (14.4%). The mean number of misclassifications across the 100 samples was 17.4, or 9.3%.

For the 200-item test, baseline correlations ranged from .951 to .967, with a median of .961. False positives for the 260 simulated test takers ranged from 3 (1.2%) to 12 (4.6%), false negatives from 4 (1.5%) to 17 (6.5%), and total misclassifications from 11 (4.2%) to 24 (9.2%). The mean number of misclassifications across the 100 samples was 17.0, or 6.5%.

Table 2 shows the correlations between true and estimated abilities for the 100-item test under the 18 conditions of drift as well as the mean differences between estimated and true abilities. Item difficulties for this test were normally distributed, and true abilities were negatively skewed. Mean differences were calculated by subtracting the true ability from the estimate so that positive figures reflect estimated abilities above the true values and negative figures reflect estimated abilities below true values. Correlations range from .855 to .943 and thus fit within the range found among the 100 baseline samples. On average, estimated abilities differ from the true abilities by only a small amount, even under maximum conditions of item difficulty drift; all mean differences but one are less than ±.05. One-sample t-tests found no significant mean differences between true and estimated abilities for the 18 conditions of drift; significance values are shown in the final column of Table 2. If item drift has affected ability estimates, one might expect to see a pattern in which mean differences increase in absolute value as the number of drifted items and the magnitude of the shift increase. In addition, one might expect to see negative mean differences in ability measures for conditions in which items were shifted upward (more difficult) and positive mean differences where items were shifted downward (easier). Yet no consistent pattern appears

among the mean differences shown in Table 2, suggesting that item drift had little, if any, effect on ability estimates.

Correlations, mean differences, and significance values (p) for the 200-item test under the 18 conditions of drift appear in Table 3. True abilities were negatively skewed for this test; item difficulties exhibited negative skew and positive kurtosis. Correlations are all very high, ranging from .955 to .967, and fit within the range of correlations among the baseline samples. As with the 100-item test, none of the mean differences is significant, and it is difficult to find a consistent pattern. However, for this test, the mean differences are noticeably larger in absolute value when 25% of the item difficulties were shifted up or down. The reason this occurs only for the 200-item test may be because of the precision gained in estimating abilities with a larger number of both items and test takers. Although mean differences of approximately .05-.06 approach a level of practical significance, they are still quite small, and it is reassuring to find that differences of this magnitude do not appear until a fairly large number of items (25%) have drifted.

Licensure and certification exams are used to make high-stakes decisions about individuals. The specific score a test taker receives is not so important as the pass/fail status determined by the score. The impact of item difficulty drift on pass/fail decisions was evaluated by examining the number of test takers misclassified under the various conditions of drift.

On the 100-item test, for the drifted samples, the number of false positives among the 187 simulated test takers ranges from 6 (3.2% of 187 simulated test takers) to 14 (7.5%); the number of false negatives ranges from 2 (1.1%) to 12 (6.4%). The total number of misclassifications for any one drift condition ranges from 10 (5.3%) to 23 (12.3%). These figures are very much in line with the no-drift, baseline samples described above.

On the 200-item test, for which both true abilities and item difficulties were negatively skewed, the number of false positives among the drifted samples ranges from 2 (0.8% of 260 simulated test takers) to 12 (4.6%); the number of false negatives ranges from 5 (1.9%) to 17 (6.5%); and the total number of misclassified test takers for any one drift condition ranges from 12 (4.6%) to 25 (9.6%). Again, these figures are in line with the number of misclassifications occurring in the baseline samples.

A representative sample of the misclassification results appears in Tables 4 and 5. A portion of the distribution of true and estimated abilities is shown, along with the pass/fail status determined by each of

these values. (The tails of the distribution are omitted as they contribute no useful information here.) False positives are marked with a single asterisk, false negatives with a double asterisk. The lightly shaded (yellow) section in the center represents the 68% confidence interval around the passing standard (1.65 logits for the 100-item test, 1.16 for the 200-item test) based on the standard error at the cut on the distribution of true abilities. The darker shaded section extends to the limits of the 95% confidence interval.

Across all 18 conditions of drift, only four of 187 simulated test takers on the 100-item test were misclassified due to ability estimates large (or small) enough to fall outside the 95% confidence interval. For the 200-item test, seven of 260 simulated test takers were thus misclassified. Tables 6 and 7 show the number of test takers misclassified, broken down into false positives (should have failed, but passed) and false negatives (should have passed, but failed) by drift condition for the 100-item and 200-item tests, respectively.

In all cases the number of misclassifications is well within what might be expected as a result of measurement error alone. Nevertheless, the pattern of misclassifications is somewhat informative. In licensure and certification testing, false positives are generally considered more serious in their consequences than false negatives are. For both tests, there are slightly more false negatives than false positives when item difficulties are shifted upwards. This is to be expected; items have become harder, yet they are anchored at their original, easier difficulties. Thus test taker ability estimates are likely to be lower than their true abilities, leading to more failures among those having the ability to pass.

One might also expect the reverse to be true when item difficulties are shifted downwards; false positives might outnumber false negatives. This does not appear to be the case however, for either of the tests, and that may be due to the shape of the ability distributions. Since both ability distributions are negatively skewed and passing standards were selected to reflect pass rates of approximately 70%, fewer test takers are found just below the cut than just above it. Thus fewer test takers will move above the cut when abilities are overestimated than will move below it when abilities are underestimated.

For the 100-item test, false positives and false negatives occur at approximately the same rate when item difficulties are shifted downward. For the 200-item test, however, false negatives actually appear to occur more often than false positives when five to 10 percent of the items become easier. Since the number of misclassifications is very small, this may be nothing more than random noise. It is possible, however, that the shape of the item difficulty distribution (negatively skewed with positive kurtosis) has also had

some small influence. It would be interesting to see if the patterns observed in Tables 6 and 7 are reversed when true distributions are positively skewed. However, there is no reason to expect that the overall number of misclassifications would be any greater.

CONCLUSION

The results of this study are encouraging in that they provide further evidence of the robustness of Rasch model ability estimates in the face of undetected item drift, even when true abilities and item difficulties are not normally distributed. A fairly large number of item difficulties (i.e., 25%) must be altered before even a hint of possible distortion of ability estimates appears. The conditions of drift investigated in this study were selected because they were assumed to be realistic. Within the context of licensure and certification testing, it is difficult to imagine a situation in which more than 25% of the test content would drift by more than about .50 logits per item--except, of course, in the case of rampant cheating. And in that case, statistical issues would not be our greatest concern.

Even at the most extreme levels of item drift considered here, mean differences, correlations, and numbers of misclassified test takers indicated no greater distortion of ability estimates than might be expected to occur as a result of chance factors. It is especially encouraging to note that skewed ability distributions do not appear to have a strong influence. If non-normal item difficulty distributions were found to exacerbate the effects of item drift, we could control this through better management of our item writing and test construction processes. But if non-normal ability distributions had been found to aggravate the consequences of item drift, there would be little we could do about it.

As long as credentialing exams are used to make high-stakes decisions about individuals, it is still important to make reasonable efforts to avoid item drift and to detect it when it occurs. Nevertheless, it is reassuring to find additional evidence, based on realistic distributions, that the effects of undetected item drift are negligible in terms of their impact on pass/fail decisions and that test takers are no more likely to be misclassified as a result of item drift than they are as a result of measurement error.

Future research along these lines will attempt to evaluate the impact of item drift on test taker abilities in additional realistic settings, including adaptive testing.

REFERENCES

Bergstrom, B. A., Stahl, J. A., & Netzky, B A. (2001, April). *Factors that influence item parameter drift.* Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Chann, K-Y., Drasgow, F., & Pfeiffenberger, W. (1988). What is the shelf life of a test? The effect of time on psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84,* 610-619.

Cizek, G. L. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369-377.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). *Item response theory.* Vol. 2. Hillsdale, NJ: Lawrence Erlbaum.

Linacre, J. M. (1999). *Winsteps.* Chicago: Mesa Press.

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1982, March). *Five steps toward controlling item parameter drift.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Schnipke, D., Popkins, N., Kinsella, K & Stahl, J. (2001) *Promissor Test Simulator*, Promissor Inc., Evanston, IL.

Stahl, J. A., Bergstrom, B. A., & Shneyderman, O. (2002, April). *Impact of item drift on test-taker measurement.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77-87.

Table 1

Representative Distributions of Test Taker Ability and Item Difficulty

	N	Mean	SD	Skew (SE)	Kurtosis (SE)
100-Item Test					
Abilities	187	2.05	.80	-.558(.178)	.399(.354)
Difficulties	100	-.03	1.27	.002(.241)	.182(.478)
200-Item Test					
Abilities	260	1.45	.66	-.446(.151)	.002(.301)
Difficulties	200	-.00	1.03	-.663(.172)	1.034(.342)

Table 2
Correlations and Mean Differences Between True and Estimated Abilities Under 18 Conditions of Item Difficulty Drift

100-Item Test
N=187
Abilities Negatively Skewed
Item Difficulties Normally Distributed

% of Items Shifted	Logit Shift	Correlation	Mean Difference	Significance Value (p)
5%	+.10	.943	.0225	.729
	+.25	.904	.0245	.703
	+.50	.927	-.0328	.604
10%	+.10	.937	-.0221	.729
	+.25	.926	.0196	.746
	+.50	.922	-.0234	.730
25%	+.10	.928	-.0495	.461
	+.25	.929	-.0148	.819
	+.50	.930	.0060	.928
5%	-.10	.943	-.0027	.967
	-.25	.935	.0437	.491
	-.50	.915	.1030	.115
10%	-.10	.940	-.0261	.691
	-.25	.925	-.0074	.907
	-.50	.937	-.0275	.658
25%	-.10	.939	.0311	.625
	-.25	.907	-.0125	.842
	-.50	.855	-.0299	.651

Table 3
Correlations and Mean Differences Between True and Estimated Abilities Under 18 Conditions of Item Difficulty Drift

200-Item Test
N=260
Abilities Negatively Skewed
Item Difficulties Distributed with Negative Skew and Positive Kurtosis

% of Items Shifted	Logit Shift	Correlation	Mean Difference	Significance Value (p)
5%	+.10	.955	.0024	.957
	+.25	.961	-.0036	.933
	+.50	.962	-.0205	.633
10%	+.10	.963	.0012	.978
	+.25	.963	-.0162	.703
	+.50	.962	-.0076	.858
25%	+.10	.964	.0569	.188
	+.25	.957	.0498	.248
	+.50	.963	.0520	.236
5%	-.10	.962	.0151	.731
	-.25	.959	-.0097	.815
	-.50	.964	-.0117	.786
10%	-.10	.964	-.0119	.778
	-.25	.961	.0295	.493
	-.50	.959	.0166	.698
25%	-.10	.960	.0408	.317
	-.25	.962	.0673	.111
	-.50	.967	.0604	.170

Table 4

Misclassifications:
5% of Items Shifted Up (Harder) by
0.10 Logits
100-Item Test

True Ability	Pass/Fail True	Pass/Fail Under Drift	Estimated Ability	
2.21	Pass	Pass	2.85	
2.21	Pass	Pass	2.26	
2.21	Pass	Pass	1.98	
2.21	Pass	Pass	2.47	
2.21	Pass	Pass	3.35	
2.13	Pass	Pass	1.82	95%CI
2.13	Pass	Pass	1.98	
2.13	Pass	Pass	1.90	
2.13	Pass	Pass	2.36	
2.13	Pass	Pass	2.71	
2.13	Pass	Pass	1.98	
2.06	Pass	Pass	1.74	
2.06	Pass	Pass	2.16	
2.06	Pass	Pass	2.36	
2.06	Pass	Pass	1.98	
2.06	Pass	Pass	1.98	
2.06	Pass	Pass	1.98	
2.00	Pass	Pass	1.67	
2.00	Pass	Pass	2.47	
2.00	Pass	Pass	2.16	
2.00	Pass	Pass	2.07	
2.00	Pass	Pass	2.07	
1.93	Pass	Pass	2.16	
1.93	Pass	Pass	1.82	68%CI
1.93	Pass	Pass	1.82	
1.87	Pass	Pass	2.16	
1.87	Pass	Pass	1.74	
1.87	Pass	Pass	2.16	
1.87	Pass	Pass	1.90	
1.87	Pass	Pass	1.82	
1.81	Pass	Pass	1.90	
1.81	Pass	Fail**	1.59	
1.81	Pass	Pass	1.98	
1.75	Pass	Fail**	1.59	
1.75	Pass	Fail**	1.59	
1.75	Pass	Fail**	1.52	
1.75	Pass	Pass	1.67	
1.75	Pass	Fail**	1.59	
1.75	Pass	Pass	1.90	

table continues

Table 4, continued

1.75	Pass	Pass	1.82	
1.75	Pass	Pass	2.16	
1.69	Pass	Fail**	1.26	
1.69	Pass	Fail**	1.52	Passing Standard
1.63	Fail	Pass*	2.26	
1.63	Fail	Pass*	1.98	
1.63	Fail	Pass*	1.67	
1.63	Fail	Fail	1.32	
1.63	Fail	Pass*	1.82	
1.58	Fail	Fail	1.52	
1.58	Fail	Fail	1.46	
1.58	Fail	Pass*	1.74	
1.52	Fail	Fail	1.46	
1.52	Fail	Pass*	1.98	
1.52	Fail	Pass*	1.82	
1.52	Fail	Fail	1.46	
1.52	Fail	Fail	1.46	
1.52	Fail	Pass*	2.26	
1.52	Fail	Pass*	1.82	
1.52	Fail	Fail	1.26	
1.52	Fail	Fail	1.59	
1.52	Fail	Fail	1.52	
1.52	Fail	Fail	.90	
1.47	Fail	Fail	1.52	
1.47	Fail	Fail	.84	
1.47	Fail	Fail	1.13	
1.47	Fail	Fail	1.59	
1.42	Fail	Fail	1.20	
1.42	Fail	Fail	1.26	
1.37	Fail	Fail	1.59	
1.37	Fail	Fail	1.26	68% CI
1.32	Fail	Fail	1.52	
1.32	Fail	Fail	1.39	
1.32	Fail	Fail	1.39	
1.32	Fail	Fail	1.39	
1.32	Fail	Fail	.90	
1.32	Fail	Fail	1.26	
1.17	Fail	Fail	.90	
1.17	Fail	Fail	.84	
1.17	Fail	Fail	1.26	
1.17	Fail	Fail	1.01	
1.13	Fail	Fail	.96	95% CI
1.08	Fail	Fail	.96	
1.03	Fail	Fail	1.01	
.99	Fail	Fail	1.07	
.99	Fail	Fail	1.01	

Table 5

Misclassifications:
10% of Items Shifted Down (Easier)
by 0.25 Logits
200-Item Test

True Ability	Pass/Fail True	Pass/Fail Under Drift	Estimated Ability	
1.58	Pass	Pass	1.69	
1.54	Pass	Pass	1.48	
1.54	Pass	Pass	1.69	
1.54	Pass	Pass	1.58	
1.54	Pass	Pass	1.38	
1.51	Pass	Pass	1.87	
1.51	Pass	Pass	1.58	
1.51	Pass	Pass	1.51	
1.48	Pass	Pass	1.38	95% CI
1.48	Pass	Pass	1.69	
1.48	Pass	Pass	1.65	
1.48	Pass	Pass	1.51	
1.45	Pass	Pass	1.35	
1.45	Pass	Pass	1.55	
1.45	Pass	Pass	1.69	
1.45	Pass	Pass	1.72	
1.41	Pass	Pass	1.76	
1.41	Pass	Pass	1.23	
1.41	Pass	Pass	1.58	
1.41	Pass	Pass	1.62	
1.41	Pass	Pass	1.42	
1.41	Pass	Pass	1.62	
1.41	Pass	Pass	1.42	
1.41	Pass	Pass	1.45	
1.41	Pass	Fail**	1.12	
1.41	Pass	Fail**	1.15	
1.41	Pass	Pass	1.42	
1.38	Pass	Pass	1.58	
1.38	Pass	Pass	1.29	
1.38	Pass	Pass	1.38	
1.38	Pass	Pass	1.35	
1.38	Pass	Pass	1.17	
1.38	Pass	Pass	1.45	
1.38	Pass	Pass	1.17	
1.38	Pass	Fail**	.95	
1.35	Pass	Pass	1.51	
1.35	Pass	Pass	1.58	
1.35	Pass	Pass	1.26	
1.32	Pass	Pass	1.35	68% CI

table continues

Table 5, continued

1.32	Pass	Fail**	1.15	
1.29	Pass	Fail**	1.15	
1.29	Pass	Pass	1.26	
1.29	Pass	Pass	1.35	
1.29	Pass	Pass	1.35	
1.29	Pass	Pass	1.42	
1.26	Pass	Fail**	.90	
1.26	Pass	Fail**	1.12	
1.26	Pass	Pass	1.45	
1.26	Pass	Pass	1.55	
1.23	Pass	Fail**	1.09	
1.23	Pass	Pass	1.42	
1.23	Pass	Pass	1.23	
1.23	Pass	Pass	1.23	
1.23	Pass	Pass	1.45	
1.23	Pass	Pass	1.20	
1.20	Pass	Pass	1.20	
1.20	Pass	Pass	1.26	
1.17	Pass	Pass	1.42	
1.17	Pass	Pass	1.23	
1.17	Pass	Pass	1.29	Passing Standard
1.14	Fail	Pass	1.32	
1.14	Fail	Pass*	1.48	
1.14	Fail	Fail	.95	
1.14	Fail	Fail	1.01	
1.14	Fail	Fail	.77	
1.12	Fail	Pass*	1.17	
1.12	Fail	Fail	.85	
1.12	Fail	Fail	1.09	
1.12	Fail	Fail	.95	
1.09	Fail	Fail	1.09	
1.06	Fail	Pass*	1.20	
1.03	Fail	Fail	.90	
1.03	Fail	Fail	1.01	
1.03	Fail	Fail	.95	
1.01	Fail	Pass*	1.17	
1.01	Fail	Fail	.95	
1.01	Fail	Fail	1.12	
1.01	Fail	Fail	1.03	
1.01	Fail	Pass*	1.32	68% CI
.98	Fail	Fail	.72	
.95	Fail	Fail	.67	
.95	Fail	Fail	1.09	
.95	Fail	Fail	1.09	
.90	Fail	Pass*	1.23	
.90	Fail	Fail	1.06	
.90	Fail	Fail	.57	

table continues

Table 5, continued

.87	Fail	Fail	.98	
.87	Fail	Fail	1.09	
.87	Fail	Fail	.77	
.87	Fail	Fail	.92	
.87	Fail	Fail	.62	
.84	Fail	Fail	.64	
.84	Fail	Fail	.62	
.84	Fail	Fail	.74	
.82	Fail	Fail	.67	
.82	Fail	Fail	.47	95% CI
.79	Fail	Fail	.98	
.79	Fail	Fail	.92	
.79	Fail	Pass*	1.32	
.77	Fail	Fail	.77	
.77	Fail	Fail	.72	
.74	Fail	Fail	.69	
.74	Fail	Fail	.72	
.72	Fail	Fail	.74	
.69	Fail	Fail	.49	

Table 6
Number of Misclassified Test Takers With Ability Estimates Outside the 95% and 68% Confidence Intervals Under Various Conditions of Item Drift

100-Item Test
N=187
Abilities Negatively Skewed
Item Difficulties Normally Distributed

% of Items Shifted	Logit Shift	Outside 95% Confidence Interval		Outside 68% Confidence Interval	
		P False Positives	F False Negatives	P False Positives	F False Negatives
5%	+.10	0	0	0	0
	+.25	0	1 (0.5%)	1 (0.5%)	3 (1.6%)
	+.50	0	2 (1.1%)	0	4 (2.1%)
10%	+.10	0	0	0	5 (2.7%)
	+.25	0	0	2 (1.1%)	3 (1.6%)
	+.50	0	0	2 (1.1%)	0
25%	+.10	0	0	2 (1.1%)	2 (1.1%)
	+.25	0	0	1 (0.5%)	3 (1.6%)
	+.50	0	0	0	2 (1.1%)
5%	-.10	0	0	0	0
	-.25	0	0	1 (0.5%)	1 (0.5%)
	-.50	0	0	1 (0.5%)	0
10%	-.10	1 (0.5%)	0	1 (0.5%)	1 (0.5%)
	-.25	0	0	2 (1.1%)	1 (0.5%)
	-.50	0	0	1 (0.5%)	0
25%	-.10	0	0	1 (0.5%)	1 (0.5%)
	-.25	0	0	1 (0.5%)	1 (0.5%)
	-.50	0	0	1 (0.5%)	0

Table 7
Number of Misclassified Test Takers With Ability Estimates Outside the 95% and 68% Confidence Intervals Under Various Conditions of Item Drift

200-Item Test
N=260
Abilities Negatively Skewed
Item Difficulties Distributed with Negative Skew and Positive Kurtosis

% of Items Shifted	Logit Shift	Outside 95% Confidence Interval		Outside 68% Confidence Interval	
		P False Positives	F False Negatives	P False Positives	F False Negatives
5%	+.10	0	1 (0.4%)	2 (0.8%)	6 (2.3%)
	+.25	1 (0.4%)	0	2 (0.8%)	2 (0.8%)
	+.50	0	0	0	3 (1.2%)
10%	+.10	0	0	0	4 (1.5%)
	+.25	0	0	0	6 (2.3%)
	+.50	0	1 (0.4%)	1 (0.4%)	4 (1.5%)
25%	+.10	0	0	2 (0.8%)	0
	+.25	0	0	0	2 (0.8%)
	+.50	0	0	0	2 (0.8%)
5%	-.10	0	0	0	3 (1.2%)
	-.25	0	1 (0.4%)	1 (0.4%)	2 (0.8%)
	-.50	0	2 (0.8%)	0	6 (2.3%)
10%	-.10	0	0	1 (0.4%)	2 (0.8%)
	-.25	1 (0.4%)	0	2 (0.8%)	3 (1.2%)
	-.50	0	0	0	5 (1.9%)
25%	-.10	0	0	1 (0.4%)	1 (0.4%)
	-.25	0	0	2 (0.8%)	1 (0.4%)
	-.50	0	0	1 (0.4%)	2 (0.8%)

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

TM035357

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: IMPACT OF ITEM DRIFT WITH NON-NORMAL DISTRIBUTIONS

Author(s): Elizabeth A. Witt, John A. Stahl, Betty A. Bergstrom, Tim Muckle

Corporate Source: American Board of Emergency Medicine
Promissor

Publication Date: Presented @ AERA April, 2004

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Elizabeth A. Witt	Printed Name/Position/Title: Elizabeth A. Witt, Ph.D., Senior Statistician
Organization/Address: American Board of Emergency Medicine, 3000 Coolidge Road, East Lansing, MI 48823	Telephone: (517) 332-4800 FAX: (517) 332-3943
	E-Mail Address: ewitt@abem.org Date: 11/10/2004

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
University of Maryland, College Park
1129 Shriver Lab
College Park, MD 20742