Running Head: ITEM PARAMETER DRIFT ON A TAKE-HOME EXAM

An Analysis of Item Exposure and Item Parameter Drift on a Take-home Recertification Exam

Carolyn Giordano and Raja Subhiyah

National Board of Medical Examiners

Brian Hess

American Board of Internal Medicine

Paper Presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 2005.

Abstract

There are few certifying or recertifying examinations in the medical field that are given in a take-home format. This stems from a concern that examinees may discuss items with peers, or save copies of items on the exam and then pass them on to others. This study examined if item exposure on take-home examinations influences the difficulty of the exam and subsequent performance of examinees. To assess item exposure, sixty items were used repeatedly for three consecutive administrations on a take-home recertification examination. An item parameter drift analysis was conducted for those 60 repeated items as well as the non repeated items by using the differential item function in the computer software Winsteps, as well as a comparison of the *p*-values and displacement statistics. Results showed that only 12 items out of the 60 study items had significant differential item functioning; only six of these 12 items got easier over repeated administrations. These data suggest that candidates did not use knowledge of these items to their advantage. Given the context of the examination used in this study, reusing items on an unproctered high-stakes examination did not lead to widespread aberrant behavior.

An Analysis of Item Exposure and Item Parameter Drift on a Take-home Recertification Exam

It is an exam administrator's responsibility to the medical community to make sure that each exam administered correctly measures abilities from year to year. This is easily done with traditional proctored exams in a secure environment; however, with take-home exams there is a concern of not being able to correctly measure candidates' abilities due to cheating. As a result, few knowledge tests in the medical field are given in a take-home format. On take home examinations, it's expected that candidates will use the resources available to them to answer questions. However, test administrators do not want candidates to have copies of the questions and not review the material themselves. Essentially, the fear of using any old items on a take home exam is that examinees may save copies of items on the exam, and then pass them on to others on future administrations. This in turn would compromise the validity of the exam, making the exam easier and less discriminating (Donoghue & Isham, 1998).

While they are rarely administered, there are some benefits to take-home examinations. An advantage to taking an exam in an open-book take-home format is that it reduces test anxiety as well as allows examinees to work at their own pace in a comfortable environment. In addition, take-home examinations can evaluate subject matter that is difficult to memorize but is key to becoming a successful medical practitioner.

In the case of recertification examinations, the goal is to reassess basic knowledge and skills of individuals who already met certified requirements. Recertification makes certain that practitioners are as skilled as they were when they were first certified, and are knowledgeable about the ever-changing advances in their field (Benson, 1991; Norcini, 1999). The take-home exam design especially makes sense on recertification exams, on which candidates have already passed the certification exam and have been working professionals for some time. This format allows certified candidates to use the skills and resources that are used in their various medical occupations (Norcini, Lipner & Downing, 1996).

While the take-home format is justified as a good measure for recertification candidates, it does not mean the exam format is impervious to candidate cheating. To combat this concern, currently, many take-home exams do not contain any items that have been used previously. The decision to use only new items for each administration is costly, but stems from the fear that examinees in different cities or states could discuss items from past exams and simply fill out the answer sheet without having to think about the question (Luecht, 1998; Smith, 2004). With the massive popularity and ease of the Internet, Luecht (1998), states that telecommuting examinee collaboration networks (ECN) save and share items, but given the scope of the internet these networks are not easily quantified, therefore the extent is unknown.

There are, however, several known Internet websites called ‘braindumps’. These websites are pervasive in the Information Technology (IT) realm and may contain advice from candidates as well some actual items (Smith, 2004). In 2002, the Educational Testing Service (ETS) had a major security breach when candidates posted actual answers of the Graduate Records Examination Computer Science Test on the Internet. This was met by ETS canceling the computerized version for a period of time in some geographic areas (ETS Press Release, 2002). Computer, or IT, exams are at greater risk of security breaches from candidates posting items on these ‘braindumps’ (Smith, 2004). This may be because the candidates of these computer certification exams are more technologically savvy, or because the exams are more generalizable and given to such a large number of candidates.

Medical recertifying exams target a niche group, on the other hand, and there are far fewer candidates taking each exam. An Internet search of over 25 ‘braindump’ sites, as well as search engines, found no mention of any items from recertifying medical exams. Even with a thorough search of the known ‘braindumps’ and other Internet sites, there is no way to gauge item expose via word of mouth, that is, telephone, email, or face to face conversations. Therefore, there is still a major concern with the issue of item exposure, especially on take-home examinations. However, this concern has not been justified in the literature. Smith (2004) placed six information technology items on a ‘braindump’ site, and then tested those items for parameter drift versus the unknown items and found no statistical difference. It seemed that in this case, even when candidates had an unfair advantage, they were not able to capitalize on that and affect the validity of the exam.

Purpose of the Study

Many test administrators of take-home examinations err on the side of caution, knowing that cheaters exist, and that cheating is becoming easier with the help of the Internet. The current study investigated item exposure and how it would affect the validity of a take-home recertifying exam by selecting 60 items to be used repeatedly for three administrations. Item parameter drift statistically measures the change in item parameters over time (Bock, Muraki, & Pfeiffenberger, 1988; Goldstein, 1983; Mislevy, 1982), and is a useful tool to detecting if item exposure is affecting the validity of scores. To examine the influence of item exposure, this study investigated the differential item functioning statistics and *p*-values of new items and used study items on a take-home recertification exam.

Method

Data

Data were from responses on a take-home national recertifying medical exam that is given twice a year to professionals to reassess their competency, knowledge and skills in their medical field. Typically, each administration of the exam consists of three hundred new multiple-choice items. Examinees are instructed to complete their work alone, and are allowed a three-month period to complete the exam at home, and mail back their examination booklet and answer sheet.

To assess item exposure for this study, sixty items were selected to be used repeatedly for three consecutive administrations, leaving 240 new items each administration. These 60 items were similar in content to the other 240 new items for each administration. The recertifying exam used in this study is scored based on item response theory (IRT) principles, specifically, using the Rasch model (Rasch, 1980). According to this model, the probability of an examinee possessing proficiency level θ to correctly answer an item of difficulty *b* is given by the formula:

$$P(\theta) = (1 + \exp(b - \theta))^{-1}$$

The Rasch model has been well researched and is being successfully used for scoring many multiple-choice tests in the medical field (Clauser, Ross, Luecht, Nungester, & Clyman, 1997; McKinley, Julian, & Nungester, 1991). Winsteps (Lincare, 2003) is a software program used widely in testing practices for item calibration and proficiency estimation. This computer program utilizes the joint maximum likelihood estimation method and unconditional procedure (UCON) (Wright & Stone, 1979) and allows for a calibration of items and simultaneous estimation of abilities.

There are several software programs that can be used to examine item parameter drift. In this study, item parameter drift was analyzed for all items (study and live) by using the differential item functioning (DIF) procedure in Winsteps (Lincare, 2003). According to Rasch modeling, DIF occurs for an item if the response probabilities for that item cannot be attributed to the ability of the candidate and a set of difficulty parameters for that item. In other words, the DIF analysis identifies items that are abnormally easy or difficult. In addition to several different software packages available to detect DIF, there are also several different methods to compute DIF. The Winsteps DIF analysis is similar to the Mantel-Haenszel method (Linacre, 2003). Moderate to large DIF was detected if the DIF difference is .43 logits - .64 logits, or above 1.5 delta difference, and *t* is greater than 2 (Zieky, 1993).

In Winsteps, all items and candidates from the first administration were calibrated with the study items used as an anchor, then the next administration was calibrated with the difficulties from the first anchored calibration, and then the third administration was calibrated with the calculations from the first calibration. The item statistics from the three calibrations, for both study items and live items were compared.

To understand the influence of item exposures on the item's difficulty further, *p*-values were reported for both sets of items. The p-value is a measure of average item difficulty (Camilli & Shepard, 1994). Mean difficulty scores are influenced by both the difficulty of the items and the proficiency of the candidates; therefore, they cannot be meaningfully compared across administrations. However, mean difficulty scores can be used to compare the average difficulty of the study versus the non study new items within each administration. In addition, item displacement was investigated for the 60 study and non study items.

Examinees

Over the three administrations, a total of 1018 first time candidates were exposed to the 60 study items. Table 1 shows the examinee count by administration.

Results

Table 2 shows the mean difficulty by administration for both the 60 study items and the 240 new items. The comparative results are not consistent from administration to administration. Specifically, the 60 study items are less difficult than then 240 new items for the first two administrations, but are more difficult for the third administration. Additionally, the mean difficulty for the 60 items was higher for the second and third administrations than the first administration. Item displacement statistics are shown in Table 3, and demonstrate that there is very little displacement for any of the items over all three administrations.

As stated above, moderate to large DIF is detected if the DIF difference is .43 logits - .64 logits, or above 1.5 delta difference, and *t* is greater than 2 (Zieky, 1993). Using this criterion, of the 60 items that were used repeatedly for the three administrations, only 12 of those items exhibited significant DIF (Table 4). Of those 12 items that showed significant dif, half of those items got more difficult with repeated administrations.

Discussion

Only 12 items out of the 60 study items, or 20%, had significant differential item functioning as detected by WINSTEPS. This small number suggests that if candidates saw the repeated items, they did not use this knowledge to their advantage. Furthermore, of those 12 items that showed significant dif, 6 of those items got more difficult with a repeated administration. This seems to contradict the idea that repeated exposure of used items would effect cheating, and make the exam easier, giving an unfair advantage to candidates taking newer versions of the exam.

In addition to the DIF analysis not returning results pointing to cheating, the item statistics were similar for both live and study items over the course of the three administrations. For the first and second administration, the 60 repeated items were actually statistically less difficult than the new items. Taken together perhaps cheating via item exposure is not an issue and it would be acceptable to utilize some used items on take home recertifying exams.

The candidate population for the recertification exam in this study was small when compared to other large scale recertification examinations. Further, the material covered in this examination is targeted to a small niche group. Therefore, the results of this study aren’t meant to be generalizable to larger scale examinations, certainly not to high stakes examinations. It is impossible to determine precisely if examinees are cheating on the exam or not using this method. That said, looking at the item statistics and the DIF statistics this study found no evidence that using items repeatedly is affecting the overall validity of a re-certification examination of this size and type. If items are re-used over time on small scale take-home recertification exams it would greatly reduce the burden of creating all new items each administration and save both time and money for the test administrators. Given the context of the examination used in this study, reusing items on an unproctered high-stakes examination did not lead to widespread aberrant behavior.

References

Angoff, W. H. (1993). Differential Item Functioning Methodology. In P.W. Holland and H. Wainer (Eds.), *Differential Item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Benson, J. A. (1991). Certification and recertification: one approach to professional accountability. *Annals of Internal Medicine, 114,* 238-242.

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

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items (Vol. 4)*. Thousand Oaks, CA: Sage Publications.

Clauser, B. E., Ross, L. P., Luecht, R. M., Nungester, R. J., & Clyman, S. G. (1997). Using the Rasch model to equate alternative forms for performance assessments of physicians' clinical skills. In J. A. Scherpbier, C. M. P. Van der Vleuten, J. J. Rethans, & A. F. W. Steeg (Eds.) *Advances in medical education* (pp. 416-419). Dordrecht: Kluwer.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*, 33-51.

Educational Testing Service Press Release. (August 26, 2002). Security breaches force GRE board to cancel computer science test administrations. Website: www.ets.org/news/02082602.html.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *33*: 315-332.

Lincare, J. M. (2003). WINSTEPS. Chicago: MESA Press.

Luecht, R. M. (1998, April). *A framework for exploring and controlling risks associated with test item exposure over time.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

McKinley, D. W. Julian, E. R., & Nungester, R. J. (1991). *IRT model application to a national medical certifying examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

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, NY.

Norcini, J. J. (1999). Recertification in the United States. *British Medical Journal. 319,* 183-1185.

Norcini, J. J., Lipner, R., & Downing, S. (1996). How meaningful are score on a take-home recertification examination? *Academic Medicine, 71(10 Supplement),* S71-73.

Rasch, G. (1980). *Some Probabilistic Models for Intelligence and Attainment Tests*, Chicago: University of Chicago Press.

Smith, R. W. (2004). *The impact of Braindump sites on item exposure and item parameter drift.* Paper presented at the annual meeting of the American Education Research Association, San Diego, CA.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development, In P. W. Holland & H. Wainer (Eds.), *Differential Item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Table 1. *Examinee Count by administration of the 60 Study Items*.

Administration	**Number of Examinees**
First Administration	292
Second Administration	438
Third Administration	288

Table 2. *Mean Difficulty by Administration*

	60 repeated study items			**240 non repeated items**		
Administration	Mean Difficulty (SD)	Minimum	Maximum	Mean Difficulty (SD)	Minimum	Maximum
First Administration	.77 (.15)	.32	.98	.83 (.17)	.14	1.0
Second Administration	.81 (.14)	.32	.98	.82 (.20)	.01	1.0
Third Administration	.79 (.14)	.27	.97	.78 (.19)	.03	1.0

Table 3. *Mean Item Displacement by Administration.*

	60 repeated study items			**240 non repeated items**		
Administration	Mean Displacement (SD)	Minimum	Maximum	Mean Displacement (SD)	Minimum	Maximum
First Administration	.0007 (.0002)	.0000	.0000	.0007 (.0003)	.0000	.0000
Second Administration	.0005 (.0001)	.0000	.0000	.0006 (.0001)	.0000	.0000
Third Administration	.0006 (.0002)	.0000	.0000	.0006 (.0003)	.0000	.0000

Table 4.*Items that showed Significant Differential Item Functioning*.

Item	Administration	DIF Measure	DIF S.E.	Administration	DIF Measure	DIF S.E.	DIF Contrast	Joint S.E.	*t*	D.F.
1	First Administration	4.81	.16	Second Administration	3.62	.23	1.19	.28	4.21	728
1	First Administration	4.81	.16	Third Administration	4.21	.21	.61	.27	2.28	578
2	First Administration	5.50	.14	Second Administration	4.68	.15	.82	.21	3.99	728
2	First Administration	5.50	.14	Third Administration	4.95	.16	.55	.21	2.58	578
3	First Administration	5.73	.13	Third Administration	5.26	.15	.47	.20	2.35	578
3	Second Administration	5.75	.11	Third Administration	5.26	.15	.48	.19	2.56	724
4	First Administration	4.68	.17	Second Administration	5.34	.13	-.66	.21	-3.13	728
4	First Administration	4.68	.17	Third Administration	5.29	.15	-.61	.23	-2.70	578
5	Second Administration	6.62	.10	Third Administration	6.06	.13	.56	.16	3.41	724
6	First Administration	2.65	.39	Third Administration	3.80	.25	-1.14	.46	-2.49	578
7	First Administration	4.70	.17	Second Administration	4.14	.19	.57	.25	2.25	728
8	First Administration	6.56	.12	Second Administration	5.98	.11	.58	.16	3.51	728
8	First Administration	6.56	.12	Third Administration	6.07	.13	.49	.18	2.73	578
9	First Administration	5.34	.14	Second Administration	4.87	.14	.47	.20	2.32	728
10	First Administration	5.08	.15	Second Administration	4.53	.16	.56	.22	2.52	728
10	Second Administration	4.53	.16	Third Administration	5.05	.16	-.52	.23	-2.32	724
11	Second Administration	5.64	.12	Third Administration	6.12	.13	-.48	.17	-2.79	724
12	First Administration	4.73	.17	Second Administration	4.21	.18	.53	.25	2.13	728
12	Second Administration	4.21	.18	Third Administration	4.87	.17	-.66	.25	-2.67	724

**Significant DIF if contrast is .43 - .64 and t greater than 2 (Zieky, 1993).*