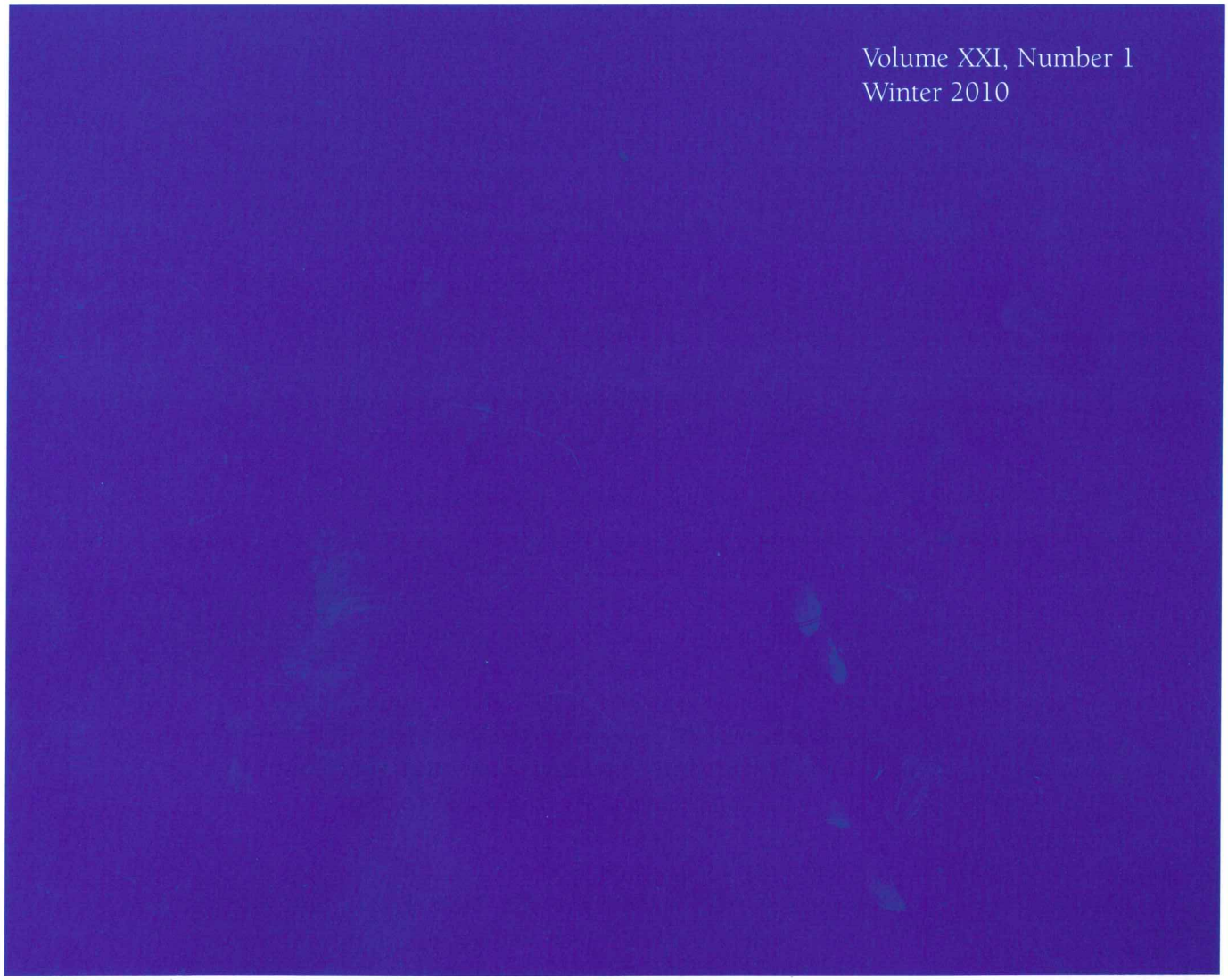
CLEAR Exam Review

Volume XXI, Number 1
Winter 2010



A Journal

CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 403 Marquis Ave., Suite 200, Lexington, KY 40502.

Editing and composition of this journal have been written by Prometric, which specializes in the design, development, and full-service operation of high-quality licensing, certification and other adult examination programs.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $30 per year to others. Contact Stephanie Thompson at (859) 269-1802, or at her e-mail address, sthompson@clearhq.org, for membership and subscription information.

Advertisements and Classified (e.g., position vacancies) for CER may be reserved by contacting Janet Horne at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page, and cost $100 or $200, respectively, per issue.

Coeditor
Michael Rosenfeld, Ph.D.
Educational Testing Service
Princeton, NJ 08541-0001
mrosenfeld@ets.org

Coeditor
F. Jay Breyer, Ph.D.
Prometric
1200 Lenox Drive
Lawrenceville, NJ 08648
jay.breyer@prometric.com

CLEAR Exam Review

VOLUME XXI, NUMBER 1 WINTER 2010

Contents

 ISSN 1076-8025

Understanding the Impact of Enemy Items on Test Validity and Measurement Precision

ADA WOO, Ph.D.
Ada Woo, Ph.D., is Senior Psychometrician at the National Council of State Boards of Nursing, Chicago, IL

JERRY L. GORHAM, Ph.D.
Jerry L. Gorham, Ph.D., is Psychometric Manager at Pearson, Bloomington, MN

Effective sampling of statistically independent items from a well-specified test plan is a key characteristic for any good standardized examination. When developing items for a test bank, it is often beneficial to produce items that are similar to one another yet not identical. This practice of producing "variant" items helps avoid memorization of items by examinees and allows test developers to use equivalent though not identical items across different forms of the examination or across published item banks (Wendt, Kao, Gorham, Woo, 2009a). Items that are produced having these characteristics are so similar in content, however, that in most cases, the items would not be administered together on one test to the same examinee. The term "enemy item" is often reserved for two or more items that should not be placed together on one test because of the adverse effects on content sampling and item independence.

Enemy items detract from the validity and the reliability of a test. Insofar as enemy items are testing duplicate content, the presentation of two similar items to an examinee on a single test reflects degradation in face validity, in content validity, and in measurement precision. Face validity is affected because the examinee may perceive the items as being redundant, which could distract the examinee or cause the examinee to question the credibility of the examination (Kane, 2006).

Content validity is also adversely affected because the domain being sampled is not consistent; the areas of the content domain represented by enemy items can be thought of as "oversampled" areas. There is little defense for such oversampling of content in standardized examinations since only a limited number of items may be sampled in any test plan area. A near duplicate sampling of content is therefore inappropriate and suboptimal (Nunnally & Bernstein, 1994).

Measurement precision is likewise affected because the two items are most likely correlated. That is, the probability of an examinee answering the first enemy item correctly is related to the probability of an examinee answering the second enemy item correctly. Conditional independence is a fundamental assumption of most IRT models. Measuring

the same content domain with more than one item violates this assumption (Mislevy, 2006). Moreover, administering two enemy items to the same individual on one test implies that two dependent pieces of information are being sampled from an examinee and are being used to calculate a score or an ability estimate, thus "double-counting" that domain and leading to a biased ability estimate.

Common Categories of Enemy Items

Enemy item equivalence can be thought of in terms of extent to which two items are similar. The more item components in common between two items, the more explicit the enemy relationship.

Duplicate items (all item components are virtually identical) – This category of item pairs describes items that are virtually identical, with the exception of punctuation, notation, or other trivial differences. These items are true duplicates. In large item banks, where close monitoring of enemy characteristics is nearly impossible without some form of automated processes, items are being developed continuously to meet test specifications. In such cases, it is possible for duplicate items to be developed, pretested, and even used operationally on the same test without test developers or psychometricians being aware of the existence of these enemy counterparts. Recent advances in software and processes for checking linguistic similarity can assist test developers in avoiding the problem of exact duplicates in an item bank (Becker & Kao, 2009).

Duplicate stems (identical stems and different options) – This category often describes items that have been scaled to production-level based on successful pretesting of a prototype. One effective use of a prototype item that has met content and psychometric criteria is to reproduce the item with slight variations in the options. Compared to developing an item "from scratch", this approach increases the likelihood of producing a useable item and allows the test developer better control of item difficulty of the new items produced (Wendt, Kao, Gorham, & Woo, 2009b). The variant items are intended for use on different tests and different candidates. They afford scalability to the item banks and may be beneficial for testing similar item content across multiple item pools or test forms. This use also reduces the risk of overexposure, thus item compromise, of one particular item across time. One consequence of using duplicate stems is that this item type may appear to be the most obvious example of duplicate items to examinees taking the test. Even if options are entirely different, the presence of identical stems may create confusion and distraction among candidates. All of which may lead to follow-up complaints by examinees and the resulting staff time investigating the legitimacy of such complaints.

Similar or duplicate options (different stems and near identical options) – This type of enemy item may occur when a subset of response options can be constructed for a variety of different stems. The advantage from a production perspective is that the group of options can be validated as plausible (but incorrect) from research sources and used in multiple items for efficiency. This will ultimately reduce the average per item production cost. The risk again is that examinees may become confused or distracted, thinking that they have seen the same item previously on their exam, and will require additional staff time for investigation and resolution of the complaint.

Duplicate stimuli (identical graphics, exhibits, sounds or reading passages) – This category describes a fairly common practice for large-scale item production. Supplemental item stimuli such as custom-produced graphics, exhibits, custom sounds, or costly copyrighted reading passages or other elements can be reused in multiple items to reduce cost and to make the best use of staff resources for research and collection of stimuli elements. The use of duplicate stimuli on an examinee's test, even if used in completely different items, may create confusion or distraction for the examinee. In a pretest setting in which groups of items using the same stimuli elements might be pretested together, this issue becomes even more important to acknowledge and to manage. For example, an examinee who receives three pretest items on one test, all using the same graphic stimulus, will likely become confused and overwhelmed from seeing the same graphic in three different items. This is especially true in computer-based testing, where candidates are not allowed to review items that have already been answered. In addition, the validity of pretest response data in a context such as this should be scrutinized carefully.

Overlapping content (different phrasing of stem and/or options, but the specific concept being tested is equivalent) – This category of item enemies is more difficult to detect in a large item bank, since most automated techniques identify enemy items based on lexical similarity and might not be able to detect overlapping content (e.g., Lin & Hovy, 2003). This category of enemies may not pose the same risk for confusing the examinee or memorization of components of the item, but nonetheless affects the content validity and measurement precision of the exam.

It is possible for two items to be enemies without sharing item format. For example, a traditional multiple choice item may

potentially be an enemy of a non-MC item (such as a "Select All that Apply" multiple response item) if there is sufficient testing equivalence between the two items. For instance, if the stem and two or three of the same options are identical between a MC and non-MC item, it may be prudent to consider the items enemies in order to avoid presenting both items to one examinee on the same test.

Methods for Managing Potential Enemy Items

1. Tagging of Enemy Items by Content Specialists

   Identification of the potential enemy relationships is the first step in avoiding duplicate content and item inter-dependence. For a fixed length traditional test form, this is a relatively simple process that is already part of most test developers' form quality checklists. For small item banks (a few hundred items or less), the task is also a relatively simple one. However, as item banks grow, the process of identifying potential enemy item pairs and subsets becomes impractical and ineffective with manual search alone. Automated software that can process all item pairs in a large item bank becomes the practical means of flagging potential enemy items. The flagged items still must be reviewed manually by subject matter experts (SMEs). To reduce burden on the SMEs, test developers can organize the potential enemy items list in descending order from most probable enemy pairs to least probable pairs. If items are tagged in the item bank database, test forms and item pools can be constructed to include only one item per enemy family, thus controlling the enemy issue at the global form or pool level.

2. Blocking Enemy Items in Delivery

   If enemy item pairs or subsets are identified and tagged in the item bank database, test driver software can often take input based on the tagged items so that selection of one item would prohibit items from its enemy set from being selected for any one particular examinee's test. This avoids many of the issues of duplicate content and item inter-dependence. One advantage of this method is that entire families of enemy items can be placed in one CAT or CBT pool without risk of any particular examinee being administered more than one enemy from each family. This feature can also be used when pretesting large groups of variant or enemy items during one testing cycle to avoid administering more than one pretest item from each enemy family to an examinee.

3. Evaluating Item Intercorrelations

   Another method that can be used following delivery is to analyze item intercorrelations among a sample of examinees and to investigate items with high intercorrelations. The high intercorrelations may provide test developers with clues about what items appear to be interdependent. Two items with significant intercorrelations that are sampling the same content, yet are not identically written, would likely be considered enemies. Although this method identifies potential enemies only "after the fact," the advantage is that items that may not be obvious candidates for enemy status can be identified based on observed data and reviewed by psychometricians and content specialists.

Conclusions

Items in any standardized examination should contain a good representation of the domain of interest as specified by the test plan. In that sense, oversampling a specific content area by asking multiple questions on the same topic may reduce measurement precision and increase a bias in estimating an examinee's ability. From a statistical perspective, placing enemy item pairs or sets in the same test violates conditional independence, an assumption of most IRT models. The present paper highlighted a few ways to detect enemy items and to prevent putting these items in the same test form or operational item pool, while maintaining item production efficiency. As testing programs grow and item pools become larger, it is essential to monitor possible enemy item sets in the item inventory.

References

Becker, K. A., & Kao, S. (2009). Finding stolen items and improving item banks. Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* 4th ed. (pp. 17-64). Westport, CT: Praeger.

Lin, C. and Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (Vol. 1, pp. 71-78). Morristown, NJ: Association for Computational Linguistics.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment, In R. L. Brennan (Ed.), *Educational Measurement* 4th ed. (pp.257-306). Westport, CT: Praeger.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory.* New York: McGraw-Hill.

Wendt, A., Kao, S., Gorham, J., & Woo, A. (2009a). Developing item variants: An empirical study. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved January 19, 2010 from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09woo.pdf.

Wendt, A., Kao, S., Gorham, J., & Woo, A. (2009b). Developing models that impact item development. *CLEAR Exam Review,* 20(2), 15-18.