Defining Item Compromise[1]

Anthony R. Zara

Pearson VUE

As more tests are administered to more examinees for more reasons, the validity of individuals' test scores becomes more important (not to suggest it was not important previously). It is incumbent on the agency or individuals using test scores to be confident that the score proffered by a test taker is a true indication of that person's level of knowledge, skill, or ability on the construct of interest. The impact of individual-level score inflation is very difficult to detect systematically, but the outcomes of inflated scores are very important for examination programs relying on testing outcomes to determine individual competency, eligibility for licensure, placement decisions, admissions decisions, etc. Much of the current research designed to address concerns related to score invalidity issues does not address item compromise per se, but rather focuses on item exposure issues. These may be related issues, but they are not the same. This paper develops a framework for defining item compromise in an operational testing program, discusses issues of item exposure and its relationship to item compromise, and discusses methods for determining whether or not test items and, by extension tests, have become compromised

High-stakes tests, by definition, lead to important outcomes for test-takers. As more tests are given to more people more often to learn more information about more their knowledge, skills, and abilities, test-takers are seemingly more willing to shade ethical boundaries to gain an advantage in the testing process than they have been in the past. This is not a new phenomenon, but with modern testing practice (e.g., continuous computer-based testing) and increasing competition for success, it seems that a non-commensurate increase in cheating behavior by test takers has been created (cheating reference). In fact, a for-profit company (Caveon) has recently incorporated for the express purpose of dealing with testing security issues.

As a consequence, both test developers and testing information consumers are becoming more sophisticated in their thinking about test validity issues. Valid test scores as an accurate reflection of test-takers' performances are becoming a more important currency throughout peoples' lifespan. Decisions based on testing information are increasing in both frequency and importance. Obviously, the increased focus on test security issues by test developers presumes that the examinations have already been planned and developed to meet traditional validity concerns.

The concerns of test security and its impact on validity as actualized in test takers achieving test results not actually reflective of their trait levels are well-placed and there are a number of ways that test validity as related to test security issues may be addressed. There has been a significant amount of psychometric research in the area of item exposure issues (see, e.g., Chan & Twu, 1998; Davey & Parshall, 1995; Stocking & Lewis, 1995; Sympson & Hetter, 1985; van der Linden & Veldkamp, 2005). It seems logical to assume that item exposure has been interesting because of the probabilistic link

[1] Paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco. Correspondence regarding this paper should be addressed to Anthony Zara, Pearson VUE. Email: tony.zara@pearson.com

between item exposure and item compromise (e.g., the more items have been administered, the higher the likelihood that they may be compromised leading to the higher the likelihood that test-takers may enter a testing opportunity with prior knowledge of specific test content). The implicit hypothesis above being that prior knowledge of specific testing content may cause test-takers to receive inflated scores on the construct of interest as compared to their actual competence level, causing some amount of score invalidity. Also, it may be true that item exposure is researched so often because it presents a more tractable set of issues than those suggested by the interesting questions regarding true item compromise and its impact on test score invalidity.

Although the argument chain for the importance of item exposure is somewhat logical, the research community's emphasis on item exposure feels misplaced when thinking about its impact on practical applications. For example, a guarantee of no item compromise can be attained through the minimum-level item exposure; that is, if the item is never administered to test-takers, it cannot be compromised (it also cannot be used, but let's not quibble). Obviously this limiting case is silly, but thinking about the real issues related to item compromise lead quickly to the conclusion that using item exposure as a proxy for true item compromise is an incomplete strategy. Another example of this weakness is that a single administration to the wrong candidate can cause an item to become compromised. Conversely, a program could administer an item to hundreds of thousands of candidates and the item may never become compromised. So, item exposure counts alone provide insufficient evidence of item compromise and are therefore only tangentially important to the issue.

This paper will develop a framework for considering issues of item compromise that expand beyond the concept of item exposure as a good proxy for compromise. It is hoped that this framework may prove useful in thinking about research issues related to test score invalidity in the future. The ambition of this paper is to raise again some of the same issues that were originally discussed in Way (1998) in which he conceptualized the issues of item pool protection in a much broader context than being related to statistical psychometrics alone.

Defining Item Compromise

This section will provide a definition of item compromise and propose several methods for determining the presence of item compromise within a testing program. Item compromise can be defined as occurring when evidence exists that an item's performance has changed during some defined timespan (generally becoming less difficult) and it is reasonable to believe that the performance changes are due to its content having been distributed beyond its defined valid usage boundaries (e.g., published in unauthorized review guides, on the web, etc.) or due to overexposure to test takers. In this definition, an item compromise has not occurred based on the item exposure level alone. It has also not occurred if the item's performance parameters have changed due to knowledge increases in the test taker population, an increased awareness of its general content (e.g. items with HIV content in 1983 as compared to the same items in 2006), or a change in the educational preparation leading to an opportunity to test.

This definition of item compromise also provides a distinction which suggests a separate definition for test compromise. Test compromise occurs when item compromise leads to improperly inflated (and thus invalid) test taker scores. An item may show parameter changes and evidence of distribution (and thus be defined as compromised), yet have no material impact on test takers' scores (which is possible through some combination of the various testing program design factors). A compromised item with minimal impact on test takers' scores suggests different operational actions than a compromised item that has a material effect on test takers' scores. The remediation strategy for these two situations would be different.

As it is defined above, potential methods for identifying item compromise should proceed in parallel along two tracks. One is a more standard statistical, psychometric type of inquiry for determining any aberrations in item performance; the second is a less psychometric type of investigation to determine the overall exposure level of an item in relation to the other factors of test program design and determining the possibility that an item was distributed outside of valid program uses and that test takers would have reasonable access to this "harvested" item.

There are a number of promising statistical/psychometric methods for investigating the first track for determining item compromise (see e.g., Karabatsos, 2003; McLeod, Lewis, & Thissen, 1999; McLeod & Schnipke, 1999). Depending on the testing program's size and stability, looking at changes in item p-values across defined time periods may prove promising. Other investigators have hypothesized that changes or differences in item latency may suggest that the item has been a victim of improper item exposure (van der Linden & van Krimpen-Stroop, 2005). A common approach has looked at using person fit or model fit as a method for identifying potentially problematic items (Karabatsos, 2003). A recent study (Smith, 2004) has applied a DIF approach for investigating items on test forms that may exhibit a change in behavior over time. Each of these approaches may provide some information on the statistical behavior of specific items, but in isolation they do not provide meaningful information about item compromise nor test score invalidity.

Item compromise is best identified through a combination of the statistical methods and other types of investigation which can link the statistical results with environmental evidence of improper item distribution or overuse. This type of holistic view provides a more balanced grounding for any decisions that must be made concerning potentially compromised items.

A Proposed Framework for Full Contemplation of Item Compromise

As outlined above, item exposure alone is not sufficient to determine the likelihood of an item which has been compromised. It is also important to investigate the continuing validity of test taker scores in the context of test security along two dimensions: the likelihood that an item is compromised and the <u>impact</u> of compromised items on test-takers' scores. Testing programs that are designed to minimize either one or both these factors can feel more comfortable that the scores produced by test-takers are an accurate reflection of their actual competence on the construct being measured. Within this framework, there are several features of examination program policy, design, and operations that can either exacerbate or minimize the likelihood and impact of item compromise on test-takers' scores. These program design features may address the either the likelihood or the impact of item compromise (or both).

Examination Design Factors

Specific item format. There are some item formats that present content in ways that are more memorable and thus more easily remembered (and perhaps then communicated). The content richness that makes these item formats appealing also cause them to become better candidates for increasing the likelihood of test takers inappropriate sharing of specific item information, perhaps then leading to item compromise. This is not to suggest that vanilla item formats are preferred, but rather that testing programs utilizing memorable content need to be diligent in their efforts to assure that specific item content is not shared and that sufficient items are produced to minimize the potential impact of that sharing.

Specific test format. Although item exposure is not equivalent to item compromise, issues related to item exposure as it is affected by the overall assessment design is an important factor in the likelihood of item compromise. For example, at one far end of the spectrum is a testing program that is designed as a forms-based assessment with items that are considered disposable, e.g. only being used in a single administration, then being retired. The other extreme can be illustrated by a testing program design that reuses items in an ongoing basis, perhaps even making the test available daily as a computerized adaptive testing (CAT) assessment. Obviously these test formats would have very different concerns about item exposure and item compromise. The psychometric research on methods for limiting item exposure has the most impact in relationship to a specific test format or design. The effect of any particular item exposure control method can be measured directly as applied to a specific test format.

Test overlap rates. Another important part of the assessment design decision-making is the determination by the testing program of how much overlap in test items is permitted across testing instances. For example, the number of common items across forms in a forms-based design, or the number of common items across operational item pools in a CAT-based design has a direct impact on the probability that test-takers will be given the same items in multiple administrations. It also determines the across-taker rates of being administered items in common. CAT item pools need to be designed to take into account not only the overlap across instances of operational item pools, but also the specific number of items within ability strata (to minimize test overlap across test takers who test within one operational item pool) (Way, 1998).

This factor (and the next two factors) is related to item exposure, not in a simple "number of exposures" context but rather in the context of types of exposure and the potential impact of the exposures on the likelihood of item compromise.

Item pool sizes. It is easy to understand the issues of item pool size as it relates to number of total item exposures. As the operational item pool moves toward containing smaller numbers of items, test-takers have an increased probability of being administered items that they have seen before; and the common item exposure across test takers also increases. This relationship is only strictly true within the structural design of a single item pool being out in the field. If the assessment design includes multiple operational pools and meaningful rotation schedules, the relationship between item pool size and item compromise becomes more complicated. This factor interacts with the item pool rotation schedule in important ways and they cannot be unwound and looked at in isolation of each other.

Item pool rotation schedules. Again, issues concerning item pool rotations are not complicated to understand in isolation. That is, it seems obvious that daily item pool rotations would produce the impact that test-takers never be administered the same items across time. However, that is only strictly true if the rotating pools do not contain common items. It is very difficult to contemplate an operational testing program with sufficient item resources to conduct daily item pool rotations without significant numbers of common items across the pools (in fact, there are no such programs). Therefore it makes sense to think about this design factor in the context of some reasonable number of item pool rotations per year. This factor also interacts with some policy factors described below, such as examination retake rules.

Test-taker Policy Factors

In terms of contemplating the likelihood and impact of item compromise, the specific assessment design factors as outlined above are important, but provide an incomplete picture. Research in these areas provides only a small window into the considerations that operational testing programs must address in the constant effort to guard against item compromise. As important as specific test design, are the policies that govern the operations of an ongoing testing program.

Examination retake rules. This is the policy that the testing organization sets regarding how often test-takers may be administered a specific assessment. Generally based on non-measurement considerations, there are some examination programs that permit examinees to take the examination every time it is offered; yet other high-stakes programs limit the number of times that examinees may take their tests. The examination retake rules vary across testing program, but may specify total number of lifetime takes, number of administrations per defined time period or some other limitation factor. This policy factor interacts with the design of the assessment program in obvious ways. For example, if the testing program is forms-based, it would be logical to limit examinees to a single exposure to any single form of the test, requiring a retake rule sensitive to the forms rotation schedule. Similarly, a CAT-based program may want to limit the test-takers exposure to a specific operational item pool and its retake rules may be designed to limit the number in conjunction to its operational item pool rotation schedule.

This factor will affect both the likelihood of items being compromised (through exposure factors) and the impact that compromised items may have on the validity of test scores. For example permitting test-takers to retake within the same form increases both the likelihood of item exposure leading to compromise and the likelihood that test-takers may inflate their scores through prior exposure to the items in a form.

Turning items "off." This refers to the examination policy whereby all items that a test-taker has been administered are logged and during any subsequent retake event, the items are "turned off" and not available to be administered to the test taker. The operationalization of this factor is dependent on both the design and the technology of the testing program (obviously not possible in a paper-and-pencil testing program). The ability to turn items off affects the possibility that test-takers will utilize previous test experience to learn about their specific test items leading to subsequent score inflation. This possibility is certainly more than speculative. There is one national high-stakes program that is gathering evidence that testing candidates are using their first administration as a "study" opportunity and only really trying on the second and subsequent administrations. The capability to turn items off has a real impact for this program.

Test-taker eligibility policies. Testing programs that are designed to have tighter control on who is permitted to take their tests have an advantage in controlling access to the test (and thus increased control over situations that raise the likelihood of item compromise). Conversely testing programs that permit anybody to test have a more difficult challenge in this regard. Testing programs that require evidence of specific educational preparation and/or sponsorship by a third-party agency (e.g., state licensure testing programs) will have the ability to control test-taker access to items in ways that programs that require only valid payment do not. These programs also have an increased capability to enforce other policy-related testing rules (see e.g., examination retake rules) based on their increased "knowledge" about the test takers.

Examination Program Operational Factors

Although the above factors related to testing program design and policy have important impact on the likelihood of item compromise, the specific operational procedures that are utilized during the testing event have an equally important place in the defense against item compromise. The visual of a "three-legged stool" as a construct for contemplating minimizing item compromise might be useful. If any of the test design factors or the testing policy factors or the testing program operational features is developed without a core goal being to reduce the likelihood of item compromise, then the three-legged stool will not stand. Having one really long leg does not balance the second or third leg being shorter. The three sets of testing factors all work together to build a solid defense against item compromise.

Test-taker identification procedures. It is easy to see how this factor is related to safeguarding the item assets of a testing program. This factor certainly interacts with some of the policy factors mentioned above (examination retake rules, test-taker eligibility rules) to help assure that only appropriate individuals are administered the test. There are plenty of current instances of test takers not being who they purport to be and taking a test with the express purpose of "harvesting" items to share with others (a prima facie case for the item compromise factor of unauthorized distribution of item content). The operational procedures for enforcing test taker identification rules are not as straight-forward as it may seem on the surface. Although it seems like obvious good practice to know who is taking a test, the specific identification steps enforced at testing events differ widely depending on the type of testing program, purpose of the examination, testing methodology (paper-and-pencil or computer-based), and even test administration provider (e.g., school testing lab, Prometric, Pearson VUE, Promissor, or others). We have had experiences with real differences in interpretation at test centers of the (what seemed like) specific directive "check government-issued i.d." For example, does it have to be current? Must it be from the local government?, etc.

The importance of being thorough in creating clear and enforceable rules cannot be overstated. Although not usually in the interest-sphere of research psychometricians, these procedures do have material impact on all testing programs' ability to protect their examinations. They also interact with the outcomes of topics of more traditional research. Research studies are only illuminating in providing practical advice to measurement programs to the extent that the features of the total testing system are taken into account.

Test-taker matching data and procedures. This design factor is very program specific operationally and it is dependent on the database sophistication of the testing program and its data collection and retention policies. The design of this factor impacts the ability of the testing program to protect its item assets from overexposure both due to test takers who attempt to take the exam more frequently than policy permits and from those who "legally" take the examination on multiple occasions (exam repeaters). It is defined here as the testing program's capability to match test-taker records with the purpose of combining the information from separate instances of test taker exam applications. It is in a testing program's interest that each time "John Q. Public" applies that his record is recognized and merged with his previous applications and data from his previous administrations. It is not good when separate data instances of the same actual "John Q. Public" exist in the database.

Although non-trivial, it is a much easier problem to find and match candidate records that contain the exact same identifiers (e.g., first and last name, birthdate, social security number, etc.) than records from those candidates who subtly change aspects of their identifying information in an attempt to subvert the matching process. Rather sophisticated algorithms for detecting and merging matching candidate records have been developed and deployed at the various testing companies, but each testing program will have test takers who will try, in different ways, to "beat" this system so individualized matching rules should be developed for each testing program.

This merging of candidate information based on matching application record information allows the program to prevent the same examinee from taking the examination at a frequency that subverts program policy. Understanding each examinee retake application also permits the testing program to merge subsequent to previous application records, thus providing the capability to utilize information about the items administered previously which permits the testing program to “turn off” items that a test taker has already been administered.

Testing event proctoring and physical security. One of the most visible operational program design factors defining a testing program’s commitment to securing its item assets (thus preventing item compromise) is its procedures for physically protecting its items during test administrations. There are a number of best practices for securing paper-and-pencil and computer-based testing events. It is beyond the scope of this paper to define a full recitation of appropriate physical security procedures, but obvious visible human proctoring is a good first step. Also, it helps testing programs in their enforcement of proper examinee testing behavior to have clear communications regarding test taker responsibilities provided to examinees in advance of the testing day.

Enforcement of item copyrights. This operational factor provides a very important deterrent to the improper sharing of item information (no matter how it has been harvested). Testing programs should always register copyrights for their content-based intellectual property. Testing programs that do not enforce their item copyrights are almost inviting individuals to harvest and share specific item content with others. Asserting intellectual property ownership and improper usage by the item harvesters or commercial item sellers is almost impossible without copyrights. Testing programs that rigorously defend their intellectual property rights send notice to test takers and others who would profit from the improper sharing of item content that that behavior will not be tolerated and will be prosecuted.

Conclusions

As testing program design and procedural elements tilt toward a more comprehensive and higher security model, the likelihood of improper item content sharing decreases, thus decreasing the likelihood of item compromise. Lowering the likelihood of item compromise is a good goal, however compromised items still need to be identified in order to learn about potential score invalidity. Also, merely identifying compromised items is not enough. Testing program administrators require valid information about compromised item impact on test scores in order to make important decisions about testing program design changes, potential changes in procedures or policy, or the need for additional item development. There is only scattered research addressing the issue of compromised item impact in sufficient detail to provide much direction for testing programs. There is even less research that provides guidance in addressing how testing program design factors can mitigate the impact of compromised items. Also, studies addressing how potential trade-offs in test design factors can affect test score validity have been lacking. For example, in terms of increasing the likelihood of producing valid scores, I have not seen research that clearly describes whether it is better (in a CAT context) to expose larger amounts of the operational item pool less times, or a lesser numbers of items more times. The answer to that type of question has a direct impact on issues related to item pool design, rotation policy and size.

As psychometricians we are generally interested in creating knowledge to assist practitioners in their mission to produce valid test results. It seems that there are several research holes in this area and that our emphasis needs to be refocused on the impact of item compromise. Many interesting studies could be designed which utilize simulations of compromised item impact including realistic inputs representing the design features of an entire testing system. There would be challenges in generalizing the results of such a study due to the specific design elements under consideration, but as these types of applied studies increase, a solid core of information would be developed from which more general knowledge could be distilled.

References

Chan, S. & Twu, B. (1998, April). *A comparative study of item exposure control models in computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Davey, T. & Parshall, C.G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16(4)*, 277-298.

McLeod, L.D., Lewis, C. & Thissen, D. (1999). *A bayesian method for the detection of item preknowledge in CAT.* (Computerized Testing Report LSAC-R-98-07). Princeton, NJ: Law School Admissions Council.

McLeod, L.D. & Schnipke, D.L. (1999, April). *Detecting items that have been memorized in the computerized adaptive testing environment.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Smith, R. W. (2004, April). *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.

Stocking, M.L. & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing.* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Sympson, J.B. & Hetter, R.D. (1985, October). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973 - 977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W.J. & Veldkamp, B.P. (2005, December). *Constraining item exposure in computerized adaptive testing with shadow tests.* (Law School Admissions Council Computerized Testing Report 02-03). Newtown, PA: Law School Admissions Council.

van der Linden, W.J. & van Krimpen-Stoop, E.M.L A. (2005, December). *Using item response times to detect aberrant responses in computerized adaptive testing.* (Law School Admissions Council Computerized Testing Report 01-02). Newtown, PA: Law School Admissions Council.

Way, W.D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*, 17-27.