2^{nd} Annual
Conference on Statistical Detection of Potential Test Fraud

Abstracts

SESSION 1: DETECTION OF ANSWER SIMILARITY AND PERSON MISFIT

Hellinger Distance and its Use to Identify Answer Copying

Authors: Seonho Shin, Larissa Smith, & Jaehoon Seol

The similarity relationship between the response vectors of two examinees is an equivalent relationship: reflexive, symmetric and transitive. The Hellinger distance is a statistic that can be used to measure the similarity between two probability distributions. The Hellinger distance $H(p,q)$ is similar to the Kullback-Leibler Divergence (KLD) index $D(f||g)$ (Cover & Thomas, 1991, Kullback & Leibler, 1951) used by Belov et. al. (2007) and Belov and Armstrong (2009) to provide an index of similarity between the two distributions and to develop an algorithm to detect answer copying. Unlike the KLD index, $D(f||g)$, which measures the relative entropy difference between the two response vectors and satisfies only reflexibility, the Hellinger distance satisfies all three properties of an equivalent relationship, making it a better statistic to measure similarity between two responses. It can also be used for a likelihood ratio approach to identify aberrant examinee responses when an examinee copies from another examinee with a different pretest block.

This study compares the performance of the Hellinger distance and the KLD index in the framework of the two-stage algorithm to detect answer copying. First we will investigate the experimental distribution and other properties of the Hellinger distance $H(p,q)$. Then we will simulate item responses for 10,000 examinees with ability levels distributed $N(0,1)$. To simulate answer copying, we will add 100 aberrant response patterns with varying percentages of copied answers. We will evaluate pretest/operational partitioning as well as other partitioning options. Finally, we will provide a comparison of Type I and II error rates between the two similarity measures.

Using a Hypergeometric-Binomial Compound Distribution for Evaluating Potential Collusion on Multiple-Choice Tests

Author: Dennis Maynes

Several methods have been proposed and used for detecting potential collusion on multiple-choice tests (Examples are: Frary, Tideman & Watts, 1977; Holland, 1996; Belleza & Belleza, 1989; Wesolowsky, 2000; van der Linden & Sotaridona, 2006; Wollack, 1997). Heretofore, all of these methods have neglected the fact that the statistical distribution of the number of identical responses is constrained. This paper derives the nature of the constraints and then shows that under the (simplified) assumptions of equally difficult items and constant conditional probabilities of identical incorrect responses that a compound hypergeometric-binomial distribution provides a very good probability model for the number of identical responses. The advantage of this approach is that a fast-computational algorithm may be implemented for screening large data sets (e.g., millions of tests and billions of pairs). After potential pairs (or clusters) have been detected, a more computationally intensive algorithm can be used in a second pass to refine the extracted anomalies. The paper discusses theoretical considerations resulting from the simplified assumptions and how those assumptions may impact false-positive and false-negative rates. By taking advantage of the compound distribution, a suggestion is provided for testing the alternative hypothesis of non-independent test taking. In other words, the discussion will suggest a way to compute the probability of non-independent test taking given the observed data, instead of computing the probability of the data given the assumption of independent test taking.

Detecting Unusual Item Response Patterns Based on Likelihood of Answer

Authors: Kyoungwon Lee Bishop & Christopher Neil Stephens

A test is to measure examinees' true ability. When cheating occurs, the test results will be inflated beyond the true ability. This study proposes a model that can detect overly scored examinees based on the likelihood of item response pattern given ability levels. This proposal is based on a desire to combine multiple extremely small likelihood values to create an interpretable person fit statistic with minimal statistical assumptions.

Using Rasch model, probabilities of getting items correct are provided given theta levels of examinees for a test. There should be a discrete theta value for each total number correct value. Test items are grouped into three based on difficulty levels: high, middle, and low. In each theta value for that number of correct items, an overall likelihood value for all possible combinations will be computed. For example, if a theta value for getting 38 items out of 40 correct is present, then all possible combinations of missing 0, 1, and 2 items would be calculated. Then a relative likelihood for each combination will be calculated based on the likelihood of that particular combination, calculated by multiplying the item p-values (based on the item being correct or incorrect), over the sum of all possible combinations based on the given parameters of the specific theta value. Each theta will then have its own population distribution of probabilities of item response patterns. Using the relative likelihood value, the combinations will be ordered from largest to smallest likelihood. All of the values with lower likelihood values than the examinee's actual combination will be summed. That value will represent the person fit statistic for that examinee. A value less than 5% would be flagged. Multiple examinees flagged in a given classroom/ school would be of heightened concern. Simulated and empirical data will be used for computation.

Bayesian Checks on Cheating on Tests

Authors: Wim J. van der Linden & Charles Lewis

Posterior odds of cheating on educational tests are presented as an alternative to current statistical hypothesis testing p-values. A major advantage of the use of posterior odds is the possibility to account for the typical incidence of cheating in the population of test takers through the specification of prior odds. In addition, their use prevents likely problems among the stakeholders of cheating analyses in the form of misinterpretation of significance levels, seemingly conflicting results between hypotheses at different levels of aggregation (e.g., individual students and school classes), and the arbitrary choice between conditional and unconditional statistical tests. The calculation of the posterior odds is demonstrated for several of the current probabilistic models for the detection of answer copying and fraudulent erasures on answer sheets.

Keynote Address

Three Eras in Cheating Detection

Presenter: Gregory J. Cizek

This presentation will describe three historical phases of quantitative methods for investigating the integrity of test data in the U.S. primarily focusing on assessment integrity in K-12 education contexts, but also touching on cheating in licensure and certification contexts. Primary attention will be paid to Phase III -- the era that those concerned with statistical detection of unethical test behavior are now entering and the challenges facing detection in increasingly distributed computer-based learning and testing environments.

SESSION 2: DETECTION OF PREKNOWLEDGE AND BRAINDUMPS

The Impact of Test Characteristics on Kullback-Leibler Divergence Index to Identify Examinees with Aberrant Responses

Authors: Jaehoon Seol & Jonathan D. Rubright

In information theory, the Kullback-Leibler Divergence (KLD) index D(r||s) measures the information loss when probability density function s(x) is used to approximate r(x) (Cover & Thomas, 1991; Kullback & Leibler, 1951). Belov et. al. (2007) and Belov and Armstrong (2009) used the KLD-index as the foundation of a two-stage algorithm to detect aberrant examinee responses. More specifically, they use the KLD-index to compare the information differences between the posterior probabilities for the operational and pretest portions of an examination. Large values of D(r||s) indicate a divergence in examinee performance between these portions of an exam. Yet, quality performance of this method is based on two preconditions: (1) The operational parts for test takers sitting in close proximity are generally identical. This helps find the asymptotic/experimental distribution of the KLD-index in advance. (2) The operational and pretest parts of the exam should be similar to each other to ensure the compatibility of an examinee's performance on the two parts.

However, examinations vary in the extent to which they satisfy the first condition listed above. The operational and pretest portions may have notably different lengths, especially in exam formats such as CAT and CBT. Additionally, the statistical properties of pretest items are generally unknown in advance, making it difficult to build a form to satisfy the second condition.

As a first step to expand the applicability of the two-stage KLD algorithm to various examination structures, we analyze, via simulation, the performance of the two-step algorithm for exams with mixed total form lengths and different operational-to-pretest length ratios. Additionally, we present performance results for exams that have pretest parts with varying levels of difficulty in comparison to the operational portion.

The results of this study will be important in identifying test characteristics where the KLD two-step algorithm may be appropriately applied to identify aberrant examinee responses and answer copying.

Identification of Test Collusion by the Methods of Information Theory and Combinatorial Optimization

Author: Dmitry I. Belov

Item preknowledge occurs when some examinees (called "aberrant examinees") know answers in advance to some subsets of items from an administered test. A result of item preknowledge is that aberrant examinees perform better on items in these subsets (called "aberrant item subsets") as compared to other items. When the number of aberrant examinees is large, the corresponding testing program and its users are negatively affected.

The detection of different types of test collusion can be reduced to the detection of item preknowledge. For example, the case of a teacher correcting answers for a group of students can be detected as large-scale item preknowledge. There are numerous person-fit statistics exploiting differences in item performance to detect aberrant examinees. Usually, these statistics assume that the aberrant item subset is the same for all aberrant examinees and is defined by assigning to each item a probability of preknowledge. This assumption is not realistic, and computer simulations demonstrate that when the number of items with a high probability of preknowledge grows, the detection rate drops dramatically.

This naturally raises the general question of how to identify aberrant item subsets, which can be formulated as a combinatorial optimization problem. The developed algorithm combines information theory and combinatorial search to detect aberrant examinees and their corresponding aberrant item subsets. This algorithm is applicable to all types of testing programs: paper-and-pencil testing (P&P), computer-based testing (CBT), multiple-stage testing (MST), and computerized adaptive testing (CAT). Computer simulations demonstrate the advantages of using the algorithm in CAT.

Detecting Examinees with Preknowledge: Examining the Robustness of the Deterministic Gated IRT Model

Authors: Carol Eckerly & James A. Wollack

The Deterministic Gated IRT Model (DGIRTM) detects examinees involved in collusion by estimating examinees' true abilities and score gains due to cheating based on differences in performance on secure items versus compromised items. This simulation study examines the robustness of the DGIRTM by looking at the power and Type I error rate of the model when item compromise status is not perfectly known by the user.

Scoring Tests with Contaminated Response Vectors

Authors: Arnond Sakworawich & Howard Wainer

Test scoring models vary in their generality, some even adjust for examinees answering multiple choice items correctly by accident (guessing), but no models, that we are aware of, automatically adjust an examinee's score when there is internal evidence of cheating. In this study we use a combination of jackknife technology with an adaptive robust estimator to reduce the bias in examinee scores due to contamination through such events as having access to some of the test items in advance of the test administration. We illustrate our methodology with a remarkable dataset collected, in part, by the FBI on a raid on an illegal cache of stolen test items thus allowing us to know that the security of some items has been breached for a subset of the examinees.

Free Proven Tools to Help Nail the Brain Dumps

Authors: Lawrence Rudner, Daniel Eyob, & Layne Pethick

Brain dumps, web sites posting unauthorized intellectual property, not only reduce the validity of your examination program, but they seriously tarnish your brand. Two software tools that test sponsors can license free of charge to help you fight this battle will be demonstrated.

BadGuyFind is a tool to help you identify web sites with copyright infringing material. The system searches the web for rogue sites that appear to have your test questions. It makes a list of such sites and then analyzes that list. Sites with large numbers of suspected infringements can then be further investigated.

Once sites are found, potentially infringing questions can be downloaded and analyzed using ItemFind. This patented tool compares downloaded files against your database of questions to identify and document copyright infringements. Its output is a set of side-by-sides showing each downloaded question along with your question and, optionally, your question's history and copyright filing information.

The session will explain how these tools work, provide a demonstration of how these tools have been used with great success at the Graduate Management Admission Council, and identify how your testing company can obtain a no-cost license.

SESSION 3: DETECTION OF UNUSUAL GAINS

Detecting Unexpected Changes in Pass Rates: A Comparison of Two Statistical Approaches

Authors: Matthew N. Gaertner & Yuanyuan Z. McBride

There is a growing body of literature focused on examining students' scale scores to detect possible test security violations. These approaches have proven useful for identifying unexpected changes in scores over time – potential test security violations related to student growth or proficiency targets under a school accountability framework. Research on pass rate analysis, however, is more limited. Approaches in this family of statistical techniques focus on the number or percentage of students school-wide who reach established proficiency targets (e.g., performance standards). Pass rate analyses are particularly sensitive to small unexpected changes in students' scale scores that (1) are undetectable via examination of scale score patterns alone, and (2) nonetheless result in substantial changes in campus-level proficiency rates.

Our study examines the comparative efficiency of two statistical approaches for detecting abnormal changes in pass rates over time. Specifically, we investigate a two-proportion Z-score and multilevel logistic regression, and our analyses are driven by a central, organizing research question: Which technique successfully identifies schools where cheating has occurred, while limiting false-positives?

The data used for this research were 2011 and 2012 grade 10 mathematics scores from a large statewide assessment program. By simulating schools where test security violations have occurred, we were able to compare the performance of the two statistical approaches under various conditions (different school sizes, different initial-year pass rates, and different amounts of cheating).

Preliminary results indicate (1) both methods increase in efficiency as cheating becomes more extensive; (2) cheating at large schools is detected more easily than cheating at small schools; and (3) the initial-year pass rate may determine which statistical approach is preferred. These findings could help states and districts understand how best to employ pass rate analysis and identify suspicious proficiency patterns, while minimizing the costs of campus monitoring and avoiding false-positives.

Analysis of Ability Changes for Repeating Examinees Using Growth Models

Authors: Shu-chuan Kao, Ada Woo, & Jerry Gorham

This study explores how much growth time repeating examinees need to improve their ability estimates and to pass a high stakes licensure exam. This information can in turn be used to decide if a Large Score Difference (LSD) associated with the exam lag is within a reasonable range. Identification of LSDs can be useful in flagging examinees who appear to be improving too quickly, and may highlight potential compromises in test security.

Many testing programs have monitored LSDs to detect significant changes in test scores for repeating examinees, but there isn't enough supporting information available to interpret a LSD appropriately. The purpose of this study is to use growth modeling to depict the growth trajectory of repeating examinees for a standardized, nation-wide licensure exam. The starting point of the model is examinees' graduation dates from their education programs and the end point of the model is the last test date when examinees pass the exam. The historical performance data and the demographic data will be implemented in the model to estimate growth trajectory. In this growth model, it is assumed that repeated measures of test scores at a specific time point are functionally related to time and other measures. It is also assumed that not every repeating examinee follows the same growth path over time. Once the model fits the data, the systematic change (both growth and decay) can be detected and the correlation of the growth parameters can be identified. It is hoped that the results from this study will help the testing program to better understand the characteristics of repeating examinees. It is expected that the methodology proposed in this research will provide a way to examine testing data and to enhance test security.

A Bayesian Hierarchical Linear Model for Detecting Group-Level Cheating and Aberrance

Authors: William P. Skorupski & Karla Egan

The purpose of this study is to demonstrate, through Monte Carlo simulation, the utility of a recently developed method for detecting group-level cheating and aberrance (Skorupski & Egan, 2011, 2012). The method relies on vertically scaled test scores across grade levels, or at the very least, linked assessment scores over time points. Using these data, the change in individual scores, nested within groups (classrooms, schools, or other units) over time may be modeled. The approach is based on a hierarchical linear model (HLM), and evaluates unusually large group-by-time interaction effects as evidence of potential cheating or aberrance. Parameters are estimated within a fully Bayesian network in order to make direct inferences about the probability of cheating behavior, given the size of the aberrance. Effect sizes for likely cheating are then developed, which are much more intuitive than frequentist null hypothesis inferences about low probabilities for aberrance given that no cheating has occurred. The authors have previously demonstrated this method using real data from a large, statewide, testing program (Skorupski & Egan, 2011) and through a small simulation study (Skorupski & Egan, 2012). These studies have provided encouraging success, through cross-validation with other methods, and through reasonable power and Type I error rates. The purpose of the current study is to directly evaluate the validity of this method under a comprehensive series of simulated cheating and non-cheating situations, using a variety of conditions to completely cross true effect size, the number of groups, the size of the groups, and the nature and persistence of the aberrance (which could be evidence of cheating or something else unusual). This evaluation will be conducted by considering marginal recovery of the known parameter values from the HLM, as well as a determination of power and Type I error rates for identifying aberrant versus non-aberrant groups.

Using Nonlinear Regression to Identify Unusual Performance Level Classification Rates

Authors: J. Michael Clark III, William P. Skorupski & Stephen T. Murphy

Test misconduct can take numerous forms and involve examinees or third-parties, such as educators. Due to the complicated and varied nature of misconduct, an assortment of tools should be used in any forensic investigation of test data. One indicator of potential misconduct is greater-than-expected gains in examinee performance across years. An approach for identifying unusual performance gains might involve comparing changes in examinees' test scores across years, perhaps using raw differences, t-tests, or some form of linear regression model. These approaches will be most sensitive to large gains in examinee scores across years, but they will be less sensitive to small gains that are due to potentially targeted misconduct (e.g., educators changing enough answers for examinees to pass the test). Modeling performance level outcomes is desirable in forensic investigations because such a technique is sensitive to both large score gains across years, as well as systematic, small increases in individual-level test scores that result in greater-than-predicted changes in performance level counts at the unit level (i.e., districts, sites, or classrooms).

The method we will discuss in this presentation employs a cumulative logit regression model to predict examinees' Year 2 performance levels from their Year 1 test scores. Individual probabilities for each Year 2 performance level are obtained from the cumulative logit model and summed across examinees within the unit to obtain expected counts for each performance level as well as an estimated unit-level standard deviation. Using these values, a standardized test statistic is created to identify units with unexpectedly large differences between observed and expected counts at each performance level. In the presentation, we will describe the characteristics of the statistic, contrast this proposed method with other approaches for identifying unusual performance gains, and share results of a study based on actual test data.

POSTER RECEPTION

Identifying Cheaters: An Experiment in Anomaly Detection

Presenters: Kathleen A. Gialluca & Paul Reiners

While the validity of a standardized examination may be compromised in numerous ways, the focus of this investigation was on the specific situation in which test-takers benefitted from access to item content prior to test administration.

Test-takers who have prior access to exam content may interact with the individual items in ways that are different than test-takers who do not have that access. There may well be a number of features – or combinations of features – of the response data (such as item latencies, correct responses to difficult items, etc.) that serve to distinguish the honest test-takers from ones who have studied the items in advance of the exam. We investigated this situation using a data set that consisted of exam results from a relatively small number of fraudulent – i.e., anomalous – cases embedded within a larger set of results from honest test-takers. This study was an application of standard supervised-learning anomaly detection methods to identify the cheaters in this large set of exam data.

The data analyzed for this study were obtained from a large computer-based testing program; response data were examined from 10,000 presumed honest (i.e., non-suspect) test-takers plus 20 test-takers who bragged on the Internet that they had seen items in advance of the exam. Various features of the response data, hypothesized to show differences between the groups, were extracted from the data set. The overlapping distributions of these features were examined and analyzed for the two groups of test-takers, with the goal of identifying the point in the distribution that best discriminated between the honest and the fraudulent cases. In this study, "best" was defined by the F1 statistic; other standard indices of classification accuracy were also computed.

Answer-Changing Statistics and Covariates

Authors: Vincent Primoli & Djibril Liassou

This paper uses data from multiple state testing programs and years to provide empirical results and comparisons related to the type and frequency of answer-changing (AC) behavior by testing mode: online and paper-and-pencil. So, erasure data from paper-and-pencil exams will be compared with the analogous telemetry data captured from online exams. The data will also be analyzed to see how the behavior might vary by item difficulty, student ability, grade level, content area, state testing program, economic status, and district/school.

It is hoped that this paper furthers the field's understanding of AC behavior as an assessment construct and the factors associated with it. In addition to providing base-rate information regarding the prevalence of AC, it is possible such information and analysis might help improve related data forensics procedures.

On the whole, AC occurs rarely (about one out of every 50 items for one program studied). Items with very high rates can occur and deserve additional scrutiny. A small negative relation between AC and item difficulty was observed. A distinct relationship between student ability and AC rates was observed. AC rates were lower for low and high ability students but higher for middle ability students. Relations with state testing program and economic status were also observed.

Answer Changing and Response Time on Computer-Based High Stakes Achievement Tests

Authors: Gail C. Tiemann & Neal M. Kingston

The validity and integrity of high-stakes achievement test scores calls for continued evolution of fraud detection methods. While answer changing in general has been studied for decades, little published research exists related to answer changing on state accountability assessments. Literature that does exist has focused solely on paper-pencil tests, using hand or optical scanner detection of changed items. The growth of computer-based testing, however, provides new context for study, since richer information on response patterns exists than has ever been available before.

The purpose of this study is to explore the efficacy of combining response time with wrong-to-right and total answer-changing flagging rules. One state's computer-based, summative assessments in reading and math will be examined. The distribution of wrong-to-right answer changes and response time will be documented, including wrong-to-right answer changes at various time points. A flagging rule will be considered which combines wrong-to-right count and answer time. Actual versus expected flagged records will be reported, along with a comparison of general count-based flagging rules to combined count and time flagging rules.

This study will expand understanding of answer changing and response-time behaviors as constructs, as well as their utility as statistical means of detecting test fraud. Since State Education Agencies (SEAs) are primarily responsible for monitoring score integrity, states and the research community must be prepared with practical and useful detection methods, before questionable situations arise. To this end, developing a thorough understanding of expected as well as unlikely patterns of answer changing is critical. This study will respond to this need by documenting answer-changing patterns and evaluating a potential new approach to detecting test fraud.

On False Key Analyses

Authors: Jennifer A. Lawlor, Charles Lewis, & Peter J. Pashley

Consider a test fraud case in which a test taker is suspected of copying from his or her neighbor during a pencil-and-paper assessment consisting of sections of multiple-choice items. If the two test takers are actually working within different sections of the assessment, scoring the suspect's responses by employing the source's responses as a false key can provide insight into the probability of copying behavior. In this case, the key appropriate to the source's section could also be applied to the suspect's responses as a false key. Other false keys could include response strings found in the possession of a test taker before an assessment has commenced, or response strings discovered on the Internet. False keys can be applied to individuals who are suspected of malfeasance or to all test takers as part of routine security sweeps. This paper provides an approach to generating empirical distributions related to typical false key analyses, which can then be used to evaluate the probability of false key scores occurring by chance. Related graphical displays that can be employed to uncover copying behavior are also discussed. These approaches to false key analyses will be illustrated with both simulated data and real data obtained from a large-scale high-stakes assessment.

Data Forensics Using Conditional Expectations to Identify Possible Harvesters and Cheaters

Authors: Sanjeev Sethi & Lawrence Rudner

Test takers with inappropriate prior knowledge can, in theory, be identified by their answering multiple difficult questions correctly and very quickly. Individuals seeking to harvest questions can, in theory, be identified by their taking an inordinate amount of time with multiple questions, ostensibly to memorize a handful of questions. But defining very quick or very slow response time is often problematic.

Studies of response times have consistently shown a great deal of variability. A short latency, for example, can be indicative of a question that is very easy for a test taker and therefore can be answered quickly, or it can be indicative of one that is so hard for the examinee that they are not going to waste any time with it. When aggregated across all ability levels the variance is almost always extremely large often to the point that latency data is almost meaningless.

In this study, expected difficulty and latency were defined for each examinee as the median difficulty and latency for all examinees within one standard error of the examinee's final score. In other words, the expected values are based on examinees that are similar in ability to the target examinee. Extreme values were defined in terms of within group percentiles, thereby avoiding distributional assumptions. The approach was applied to a dataset containing known cheaters. The data for suspected cheaters and harvesters are presented graphically which makes the aberrance of their data very obvious.

Key advantages of this approach are 1) all the requisite data is contained in the response data file – one does not need to refer to prior calibration data, 2) examinees are compared to examinees with similar scores, 3) expected values are robustly estimated, 4) it can be applied to data from a computer adaptive test, and 5) it can be easily explained to non-psychometricians.

Impact of CAT Settings on Person-Fit Index Performance

Authors: David Shin & Yuehmei Chien

Computerized adaptive tests (CAT) have been widely used in licensure, achievement, and placement tests. From the view of psychometrics, besides the content balance, item exposure control, and ability estimation issues, another important issue to be considered in CAT is the aberrant response patterns caused by reasons such as memorization (pre-knowledge of some items) and random guessing. Many statistics called person-fit statistics (PFS) have been developed to detect the misfit for CAT. Research that compared these methods showed that the performance of the PFS varied in different CAT settings. Since CAT settings varied greatly in its applications, it is necessary to investigate how the CAT settings affect the performance of the PFSes. For example, licensure tests (such as the NCLEX-RN and NCLEX-PN) are longer tests (100+) with simple content constraint structure (i.e., each item only belongs to one content category) and simple item selection algorithm. K-12 testing programs usually have CAT with median test lengths (~ 50) but complicated content structure and item selection algorithms. College placement tests such as Accuplacer tests have relatively shorter tests (< 20) but complicated content constraint structure. The purpose of this study is to conduct a series of simulations using CAT settings as studied factors to investigate their impact on the performance of the PFSes.

The CAT settings focused in this study are test length, CAT pool size, and item exposure rate. Three *operational* item pools alone with their content constraint tables from a CAT licensure exam, a K-12 state testing program, and a college placement test will be used to simulate the data. Two aberrance behaviors, cheating and random guessing, will be simulated with several aberrance rates. Three CUSUM PFS (C^-, C^L, and C^{LR}), the l_z, and the z statistics are used for analyses. Results are evaluated using the detection and type-I error rates.

Practical Application of Testing Irregularities in a Large Urban District

Authors: Joshua Looser & Marc Sanders

Milwaukee Public Schools is the 39th largest school district in the nation, with over 80,000 students (NCES, 2011). 83% of these students qualify for free/reduced-price lunch and as of 2011 8th grade MPS students scored second lowest on the NAEP in both math and reading compared to all other urban districts. Now more than ever educators are being held accountable for academic performance, and with the high demands for proficiency from NCLB, 33 states have now been approved for the NCLB waiver. Key priorities of the NCLB waiver and Race to the Top Grant (Obama's prominent education initiatives) have encouraged the use of district, administrator, and teacher evaluation procedures that incorporate student performance on high stakes tests, with a focus on growth as opposed to absolute attainment. A latent effect of these initiatives is the increased impetus for systemic falsification of test results in order to demonstrate growth. Given the increasing incentive to cheat more systematically, it isn't surprising that reports of educator cheating on high-stakes tests are becoming a significant and growing problem; for example, from 2005-2006, more than 30 incidents of educator cheating were reported (Thiessen, 2007). In order to address this growing concern, it is imperative that school districts develop protocols to address cheating at the systems level. This presentation will focus on the procedures developed by MPS to address these concerns. Procedures include preventative methods such as ongoing training to ensure rigorous test administration and test security, in addition to post-hoc procedures such as statistical detection of testing irregularities and administrative follow-up procedures.

Empirical Methods of Cheating Detection

Authors: Amin Saiar & John Weiner

Test security continues to be a central concern for all high-stakes testing programs. While technology advances have led to innovations in test item types and delivery methods that help improve security, technology has also enabled more powerful and sophisticated methods for content piracy and cheating. Thus, testing organizations must expend additional resources to protect their intellectual property and preserve the integrity of their examination programs.

Organizations can address test security via prevention, enforcement (during the test) and/or detection (after the test). This session focuses on the latter class of methods for forensic monitoring of operational testing programs. Such methods have grown in popularity in the past decade as technology has become better and cheaper, providing ready access to candidate data and making it cost-effective to apply data-mining techniques in conjunction with a variety of statistical algorithms to model various cheating and security breach scenarios.

This session will explore empirical methods and results of cheating and aberrant response detection. The presenter will describe analytic approaches that have been applied and found to be useful in monitoring large scale credentialing examination security and will share results of analyses that contrast key statistical indicators (e.g. candidate response similarity, group performance changes, and item parameter changes) in datasets in which cheating was known to have occurred vs. data in which no cheating occurred (as indicated by non-statistical indices). The session will include a discussion period to allow the audience to raise additional issues and interact with the presenter.

Item-Level Analysis of Response Similarity

Authors: Arianto Wibowo, Leonardo S. Sotaridona, & Irene Hendrawan

Statistical tests for flagging schools/classes of cheating on standardized assessments, also known as group-level analysis, have been proposed recently, e.g., Jacob & Levitt (2003), Skorupski & Egan (2011), Simon (2012), Sotaridona, Wibowo, & Hendrawan (2012). When a group is flagged, accumulation of both statistical and anecdotal evidence is a necessary next step to corroborate the initial finding from group-level analysis. Examples of additional statistical evidence can be: student-level analysis of erasures (van der Linden & Jeon, 2012), student-level response similarity analysis (van der Linden & Sotaridona, 2006; Wollack, 1997). This paper presents an *item-level* statistical test, which flags an item with unusually large number of identical responses. A large similarity index could be used as an indicator that an item might have been compromised and also as additional supporting evidence of group-level test irregularity. The method is applicable to a multiple-choice item from a paper & pencil test format. For a given examinee pair $r=(s,v)$ and item *i*, the match indicator m_{ri} indicating matching item response for item *i* of examinee *s* and examinee *v* is a Bernoulli random variable where the probability of a match, $\Pr(M_{ri}=1)=p_{ri}$ is computed conditional on the abilities of *s* and *v* and the characteristics of an item *i*. A challenge when modeling the match indicators $m_{r_1 i}$ and $m_{r_2 i}$ is that they are not independent when $r_1 \cap r_2 \neq \varnothing$, for example, matches for pair (1,2) and pair (1,3) are correlated. Hence, computation of the variance of $M_i=\sum_r m_{ri}$ is very complex. It is shown how the probability of a match for a triplet $t=(s,v,z)$, $\Pr(m_{ti}=1)=p_{ti}$ can be used to derive an expression for the variance of M_i that is useful for practical purpose, e.g., less computation time. Consequently, an item-level statistical test is derived taking into account the dependency of m_{ri} . The power and error rates of the test will be investigated in simulation studies.

An investigation of the detection of cheating on the Multistate Bar Exam

Authors: *Seo Young Lee & Mark Albanese*

The Multistate Bar Exam (MBE) is a high-stake test to determine if an examinee is competent to practice law. Since cheating on the test can result in giving a license to under-qualified person, it should be prevented in advance and by all means. If it occurs during the test, it must be caught and reported by proctors. In addition, further investigation using statistical indexes should be followed as necessary.

Cheating on the MBE is being investigated by four indexes: the number of identical responses (IR), the longest consecutive sequence of items with identical responses (IR-LCS), the number of identical responses on items with different keys (IR-DK), and the number of identical incorrect responses on items with the same key (IIR-SK). In this study, the indexes used for cheating detection in the MBE are introduced since there are few documents describing the indexes. Furthermore, the performance of the indexes to detect cheating is investigated and compared with indexes used in large-scale tests such as B index, H index (Angoff, 1974), and K index (Holland, 1996). ω index (Wollack, 1997) is served as the standard for comparison of the different indexes since it is known to be robust to real data (Wollack, 2003). The indexes are applied to the cases reported by proctors as irregular behavior from the MBE administered in February 2011.

Student Cheating: Best Practice for Deterrence, Detection, and Disposition

Authors: *Mark Albanese, Sharon McDonough, George Mejicano, Elizabeth Petty, & Jasna Vuk*

The objectives of the poster are to discuss best practices for deterring cheating, detecting unusual similarity (cheating), and dealing with students who have cheated. The investigators include medical educators, pharmacy educators, the director of research for a national standardized testing agency and senior associate deans for academic affairs and education. The poster will discuss the session objectives followed by each investigator's perspective on each of those objectives

CONCURRENT SESSION 4A: ROBUSTNESS AND SENSITIVITY ISSUES

Robustness of Common Data Forensics Methods in the Presence of Model Misfit

Authors: Jessalyn Smith & Karla Egan

Model misfit and cheating on tests are two of a countless number of threats to test validity. Even though both procedures to detect test cheating and model misfit common in the psychometric literature, few works have examined the effectiveness of these procedures when both model misfit and cheating occur. The proposed simulation study evaluates the robustness of select data forensics procedures (such as gains score analysis, response similarity analysis, and erasure analysis) to different degrees of model misfit. Operational data from a K-12 testing program will be used to determine realistic values for all IRT item characteristics. For the simulated cases, the proportion of examinees that exhibit cheating behavior will vary. Additionally, the degree and type of model misfit will vary between study conditions. A baseline condition will be used to identify a best-case scenario. Two types of model misfit will be explored in this study: 1.) data will be simulated to exhibit item parameter drift (IPD); and 2.) data will be simulated to emphasize a situation where widespread cheating has occurred resulting in a high ability examinee groups. This study will increase understanding of how reliable the data forensic methods are under conditions of model misfit. Results will be presented to help identify any strengths and weakness of the data forensics methods.

Realistic Simulation Study to Investigate Sensitivity of Data Forensic Methods for Various Cheating Types

Authors: Mayuko Simon, Christie Plackner, Vincent Primoli, & Djibril Liassou,

As the need for data forensics increases, more research has been conducted in the area. However, the sensitivity of the methodologies is often ignored. Simulation studies have been conducted by many researchers; however simulated data are typically with a known distribution. The truth is that real data are not often normally distributed and moreover, we do not know how often actual cheating is happening and how well we are detecting instances. When research studies used empirical data and reported results obtained by the methodologies being investigated, the sensitivity of the methods were ignored (e.g., Mroch, Lu, Huang, & Harris, 2012; Simon, 2012). A study by Satoridona, Wibowo, and Hendrawan (2012) used real data for their simulation of classroom level copying and reported the sensitivity of their copying index, but focused on only copying behavior.

The main purpose of the study is to conduct a realistic simulation study to examine sensitivity of several data forensic methodologies under various cheating types. The simulation used real data from more than 1000 schools with student responses manually changed to reflect test fraud.

The factors investigated in this study are school sample size, school average scale score levels, the number of items cheated, and the various types of cheating (e.g., random, consistent, copying). These cheating types were included to reflect possible teacher/administrator modification of items with varying levels of modification.

The proposed study will show how to simulate using real data and examine the performance of data forensic methodologies, including erasure analysis, scale score analysis, modified Jacob and Levitt methods (Plackner and Primoli, 2012). A preliminary study has been conducted and the results show that different methodologies have varying strength of detection with different types of cheating. The simulation study will summarize results from 20 replications.

The Performance of Statistical Indices in Detecting Answer Copying on Multiple-Choice Examinations Using Dichotomous Item Scores

Authors: Cengiz Zopluoglu, Troy T. Chen, Chi-Yu Huang, & Andrew A. Mroch

Several indices have been proposed to detect unusual answer similarity between two examinees taking a test. In previous simulation studies of the empirical power and type I error rates of these answer copying indices with multiple-choice items, examinee response option chosen for each multiple-choice item (e.g., “A”, “B”, “C”, “D”) was simulated using a nominal response IRT model. Then, the response option chosen was used when computing the answer copying indices. Using response option provides more information for distinguishing unusual answer similarity between examinees. However, in practical applications a dichotomous IRT model is often used for calibrations, and the parameter estimates from these calibrations are readily available for use in security analysis.

The primary purpose of this study is to adapt existing answer copying indices (generalized binomial test, ω, K, K1, K2, S1, S2, lz, and modified caution index) to dichotomous IRT models and to examine the utility of applying these indices in detecting answer copying on multiple-choice tests, specifically comparing the power and type I error rates across different indices under simulated conditions.

We manipulated the following factors in this study: test length (30-item and 50-item), dichotomous IRT model (2-PL and 3-PL), types of copying (random and random-string), ability levels of the examinees in answer copying pairs (low-low, low-medium, low-high, medium-medium, medium-high, and high-high), and amount of copying (20%, 40%, and 60%). All factors were fully-crossed yielding a total of 144 conditions for empirical power analysis. For power analysis, 5,000 answer copying pairs were simulated within each condition. For type I error analysis, test length and IRT model conditions were crossed yielding a total of four conditions and 180,000 "honest" examinee pairs (with no simulated copying) were simulated within each condition.

Not So Neat? Evidence of Fraudulent Preparation on Internal Anchor Items

Author: N. Scott Bishop

Educators who are motivated to do so have many ways to cheat. Although there are some commonly applied procedures to flag suspicious outcomes (e.g., erasure analysis) these are not sensitive to all types of misconduct. For testing programs that use a common-items, nonequivalent-groups (CINEG) equating design, a subset of items that appeared on previous years' tests also appears on the current year's test. In practice, these items generally contribute to the student's operational scores in the second administration (i.e., they are *internal* anchor items). It is possible that school personnel might copy or reproduce items from the older tests and then use that information in their instruction, test preparation, and/or coaching activities before future testing. Students who receive such preparation will have an unfair advantage over other students on the anchor items, and thus will earn higher test scores. This paper explores the use of Rasch item residuals over linking items and non-linking items (aggregated at the school level) as a potential approach to identify such unethical practice.

An Investigation of Unusual Response Patterns on a Large Scale Assessment

Authors: Hulya Yurekli, Xinya Liang, Jin Koo, Betsy Becker, Insu Paek, & Salih Binici

This project involves investigating unusual response patterns on a large-scale state assessment. The purpose is to investigate unusual response patterns for reading and mathematics items using person-fit statistics and to examine the impact of unusual responses on the calibration. The research questions include: What percent of students have unusual responses? Does the unusual response pattern differ based on grade, achievement level, gender, race, and curriculum groups? The data used have been collected from all students who took a statewide large scale reading and mathematics assessment in 2011. In this study, we will calculate several person-fit statistics and will draw person response curves using the WPerfit program. We will also estimate the item and person parameters using a 3 parametric logistic item response theory method using MULTILOG software.

CONCURRENT SESSION 4B: TEST TAMPERING

Compelling Illustrations of Test Score Tampering

Author: Cynthia Butler

The focus of this paper is to present the results of a comprehensive data forensics analysis using illustrations of the data or statistical methods, and images of actual erasures. The data forensics analysis was conducted to explore the possibility that tampering occurred in order to increase test scores. Attendees will learn what kinds of illustrations are useful and compelling for presenting their own analyses to non-test professionals like judges, attorneys, jurors, peers, clients, and students. Hand-tabulated erasure counts were used for this analysis. Due to privacy concerns, some details concerning the analysis will not be shared.

Compelling Illustrations: (a) images of actual erasures on multiple-choice and hand-written items, (b) effective comparisons of test results using dotplots, (c) effective comparisons of erasure counts using box-n-whisker plots, (d) displays of matched students' erasure rates from year to year useful for a longitudinal study or the paired t-test, and (e) displays of the relationship between erasure rates and test scores that provide evidence of targeting answer documents to increase pass rates.

If there is time, the measurement error of hand tabulated erasures will be shown and the need for vendors to provide similar erasure measurement error from scanning equipment will be discussed.

Multilevel Detection of Possible Test Tampering

Authors: Shanshan Qin & Allan S. Cohen

Erasure analyses in educational accountability testing programs usually investigate potential tampering by teachers or administrators on the base of between-group differences on the prevalence of wrong-to-right (WTR) erasures, without controlling for the impact of covariates, such as examinees' ability. Methodologies reported in the literature mainly focus on person level detection with no probabilistic statement that a test has been tampered. In this study, we extend the IRT-based tampering detection method by Wollack, Cohen, and Eckerly (2013) to a group-level analysis in the context of hierarchical linear models with explanatory covariates. Type I error and power (i.e., detection) rates of this approach are examined with a simulation study. Of particular importance are determining the criteria for flagging a suspected case at each level as indicative of erasures.

Detection of Test Tampering at the Group Level

Authors: James A. Wollack & Carol Eckerly

When test tampering occurs, although it is the individual examinee who is associated with the unusual response pattern and who receives the spuriously high test score, it is actually the administrator(s) engaging in the answer changing behavior that we would like to detect. Therefore, test tampering investigations are rarely done at the examinee-level; rather, analyses are conducted at the teacher, school, or district level to identify consistent within-group patterns that are consistent with a tampering hypothesis.

In this study, we develop an extension of the erasure detection index (EDI; Wollack, Cohen, and Eckerly, 2013) to identify test tampering beyond the level of the individual. Using simulated data, we demonstrate that this method has acceptable Type I error control and offers strong power to detect tampering at both the class and school level. Results indicate that power is affected by class size, the number of tampered students, and the number of tampered questions, while school-level tampering is further influenced by the number of tampered classes.

An Initial Exploration into the Use of Erasure Analysis Results to Target Monitoring and Investigations

Authors: Steven G. Viger, Dong Gi Seo, & Shiqi Hao

Traditionally, the size and scope of large scale K-12 testing does not permit a State Educational Agency (SEA) to have personnel on-site for the millions of individual tests administered during a typical testing window. As such, the actual test administration in addition to any monitoring or proctoring activities becomes the responsibility of the Local Educational Agency (LEA). This practice makes it extremely difficult for potential instances of dishonesty that would be detectable in the results, to be directly observed by the SEA. As one would expect, the self-reported instances of misadministration are few and far between and are largely the result of failure to follow protocol or student refusal. Operating behind the scenes are numerous policy and legal barriers that can greatly limit the ability of the SEA to enforce anomalies in data without direct corroborating evidence that supports misconduct. Therefore, in order to actually be able to enforce or challenge results based solely on data characteristics, rules would have to be promulgated, vetted publicly, and then applied with criteria for action being a policy decision, and one needing considerable data to establish baselines.

One common set of results scrutinized are erasure analyses from paper/pencil administrations. In our proposed session, we will share with the audience one state's first attempt to identify anomalies in erasure data utilizing multiple methods or criteria. We will also share what we found in follow-up investigations, including the resulting additional analyses conducted to better understand the empirical phenomenon. While the generalizations will be limited, we hope that the audience will appreciate the pragmatic context in which the investigations and techniques were applied in addition to discussing the policy and legal limitations that often hinder significant follow-up.

Item-Level Analysis of Wrong-to-Right Erasures

Authors: Leonardo S. Sotaridona, Arianto Wibowo, & Irene Hendrawan

A statistical analysis of the number of wrong-to-right erasures (WTR) in statewide assessments data is becoming customary practice as part of test security protocol adopted by state education agencies to identify possible occurrences of test fraud. The unit of analysis is often on a group of examinees, e.g., school or class. The general consensus is that once a group is flagged for suspicious test taking behaviour, qualitative analyses have to be undertaken to rule out alternative explanations for the flag. Furthermore, collection of collateral information to substantiate or refute the allegation is necessary in order to minimize false accusations. The focus of the present paper is also on the analysis of WTR but with an emphasis on the *item level* instead of on examinee groups. The information obtained from item-level analysis of erasures, in combination with other independent analyses, e.g., item-level similarity analysis (Wibowo, in preparation), student-level similarity analysis (van der Linden & Sotaridona, 2006; Wollack, 1997) can be used as part of substantive analyses following the group-level analysis of test irregularity. Additionally, an item that is flagged from different units, either by using the method presented in this paper or other methods, e.g., item-fit and parameter drift, may indicate that an item has been seriously compromised for future operational use. In this paper, a statistical test of WTR on item *i* in unit u (W_{iu}) is proposed. The distribution of W_{iu} is a realization from a series of independent Bernoulli random variables, each with different success probability p_{is}, where $s = 1, \ldots, S_u$ denotes examinees in unit *u*. The success probabilities are computed non-parametrically conditional on incorrect initial responses and demographic variables, e.g., examinee abilities and AYP designation of the school. Results of simulation studies investigating the statistical properties of the proposed method will be reported.

Concurrent Session 5A: Computer-Based Testing

Data Forensics: Enhancing Security in a Computerized Adaptive Exam

Authors: Xiao Luo, Melissa Franke, & Logan West

Vigilant test security is one of the essential components to maintaining the integrity of a high-stakes examination. A significant part of upholding that integrity is evaluating the exam's security at the multiple levels in which test fraud can occur. Data forensics has become a valuable security tool for administrators seeking to secure the integrity of their exams and discover patterns of potentially fraudulent activity.

Commonly, potential test fraud situations are dealt with on a case-by-case basis: a candidate's scores or responses are unusual, that candidate's exam is reviewed and the result of any investigation determines the course of action. While this means of detecting cheating has traditionally been acceptable, as candidates have become more sophisticated in their means of committing test fraud, test security has also evolved to include new ways to ensure and reinforce security. Computerized adaptive tests (CAT) are one way in which test security has been integrated into test design. However, the test security properties inherent in CAT also make it difficult to detect fraudulent candidates and results. The use of data forensics allows for the consideration not just of individual cases, but historical data that have been collected over time. These data can be used to decipher patterns at multiple points of the exam process, including information gathered at registration (e.g., educational program, test location or time of day) or during exam administration (e.g., item responses, ability estimates or break times). The purpose of this study is to combine these multiple facets of data from an operational high-stakes CAT examination to determine patterns that may indicate candidates whose performance does not reflect their actual ability. Analysis of the patterns of data may also point toward and prevent potentially fraudulent test takers or situations.

Detection of Compromised Items Using Response Time

Authors: Muhammad Naveed Khalid, Christopher Neil Stephens. Ardeshir Geranpayeh, & Farah Shafiq

An important problem faced by high stakes CAT programs is test item pool security. It is challenging to maintain security, however, because items are repeatedly being exposed to examinees who might remember items they received. Examinees are gaining access to the test content prior to the examination being administered via illicit means. Response times are a most intriguing source of data for the computation of aberrance indicators related to item pre knowledge. In present study we have used various descriptive measures such as task duration by centre and date, count of correct and incorrect, task exposure and task exposure by centre and date to screen out the potential compromised items. Additionally, we have examined the proportion correct, proportion response time and the response latency values for the potential compromised items.

Utilization of Response Time in Data Forensics of K-12 Computer-Based Assessment

Authors: Lucy Liu, Christie Plackner, & Vincent Primoli

In the current study, the authors investigated the procedures and utility of using item response time to detect aberrant test behaviors on K-12 on-line state tests. When an unexpected short item response time produces an unexpected correct response, it may indicate that the examinee has some preknowledge of the item. Given the ability of a person (estimated by the 3PL model) and the speededness of an item (predicted by a loglinear response time model), the expected response time (called effective response time, ERT) was estimated through an ERT data set selected by two criteria: the item was answered correctly and the probability of item being answered correctly is large enough. For each item, the divergence of the observed response time from the expected ERT was tested by a chi-square statistic against the standard normal distribution. Furthermore, the item-level statistics were aggregated to subgroups of items classified by item difficulty and item type to examine whether there was differential impact of these item properties on the likelihood of aberrance responses. In addition, the item-level statistics were aggregated to student-level and school-level to identify suspicious examinees or schools.

The proposed procedures were applied to a real data set from a state assessment test and provided insightful findings regarding on-line test fraud detection. Higher likelihood of aberrant responses seems to be more related to multiple-choice items rather than constructed-response items. No clear relationships were found between item difficulty and the tendency for cheating. The findings from the item response time approach were consistent (or cross-validated) with those from the other statistics to detect aberrant behavior, such as wrong-to-right answer changes and counts of item visits, which indicates that the ERT approach is a promising method to distinguish examinees with aberrant test behaviors from those with regular test behaviors.

Detecting Aberrant Test Response Patterns Through Modeling Expected Response Time with Item and Examinee Characteristics

Authors: Jiyoon Park & Yu Zhang

Security breaches undermine the validity and fairness of test scores, and detecting compromised scores and items has become a top concern for test developers, especially for high-stakes tests. Despite advantages of using person-fit analysis to detect aberrant response patterns, implementing person-fit analysis in practice has limitations due to low power. Therefore, methods relying on other test information, such as item response times, have been considered complementary. In this study, we examine and compare the efficiency of a series of regression models that use response times to detect aberrant responses.

A regression framework with which to model response times was introduced by van der Linden and Krimpen-Stoop (2003). Meijer and Sotaridona (2006) added an examinee-ability variable to van der Linden's model to detect item pre-knowledge. In this approach, the expected response times of individual examinees are evaluated at item level and are compared to the individual examinee's observed response times. Our study uses this approach as the base model but tries to improve model fit in a hierarchical linear structure.

In the proposed models, we incorporate examinee and item characteristics that might affect the structure between examinee ability and response time. Two variables are added at the examinee level (location of education and test-taking history) and at the item level (item difficulty and cognitive level). For candidates who have similar knowledge or skills, U.S. educated examinees may be able to solve an item faster than non-U.S. educated examinees are able to, due to advantages in language. Candidates who take tests for the purpose of harvesting items may be more likely to fail in their first attempt due to a low level of motivation and intention to re-take the examination than legitimate examinees. By estimating response times for multiple items simultaneously, the proposed models are expected to lead to more accurate estimates of expected response times, which may result in higher detection rates of aberrant responses.

CONCURRENT SESSION 5B: POLICY ISSUES

Cheating on Tests: A Threat to Response Validity

Authors: Mark Albanese & James A. Wollack

Cheating on standardized tests, how it is detected and what is done about it are confidential and methods of detection are often treated as proprietary. In this paper, we propose that cheating be treated as a threat to response validity. In this conceptualization, the validity argument approach (Kane, Crooks and Cohen, 1999) would be used to support taking four levels of action: 1. purge the responses from the test data-base; 2. launch a "forensic" search of the evidence; 3. withhold reporting scores/require retaking the exam; 4. pursue charges of fraud, or other serious allegations.

Treating cheating as a response validity concern changes the nature of the process from being an action taken against individuals and the potential backlash that such action can yield to developing an argument for whether the responses should be pulled from the data-base. This is a low stakes decision that merits the development of tools to detect low levels of cheating and their impact on test analysis data and other aspects of the test development process. In the process of developing that argument, it may come to light that a higher level action is merited, but that would not be the focal point. Processes for confirming response validity could become standard practices incorporated into the data that would be a by-product of normal test analyses, much like difficulty and discrimination-type data are done currently.

The other part of considering cheating as a threat to response validity is that it enables the validity argument to be crafted to the situation. As technology capability increases, examinees can find ever more creative ways to fraudulently obtain an item answer. Test developers need to have tools to keep ahead of the game. The ability to have a validity argument in their arsenal will help to level the playing field.

Establishing Baseline Data for Incidents of Misconduct in the Next Generation Assessment Environment

Authors: Deborah J. Harris, Chi-Yu Huang, & Rya Dunnington

Statistical indices, such as fit indices, the number of wrong-to-right erasures, and the amount of time an examinee spends in responding to an item, are part of the totality of evidence used in determining if an incident of misconduct has occurred during a test administration. However, it is often difficult to determine when a value is indicative of aberrance. Having baseline data to compare values to both increases the likelihood of the values not being over or under interpreted, and provides a context in which to present information if the decision is made to pursue an incident.

In the next generation of assessments, with more emphasis on new item types, computer delivery, monitoring growth over time, and using assessment results as a component in teacher evaluation, identifying misconduct is both more challenging and important. This paper discusses the philosophy and policy issues related to the establishment and use of baseline data, provides examples of how to develop baseline data for next generation assessments, including individual, classroom, and school/district level, and both on a single testing event, and longitudinally.

The paper ends with a discussion of how to present results. For example, using innovative graphic tools, such as Tableau, to display an observed value compared to the benchmark data can facilitate the interpretation of the evidence to those without statistical training, such as arbitrators and judges.

Enhancing Security by Reducing Item Exposure: Detecting Low and Unusual Performance during Testing

Authors: Richard Feinberg, Michael G. Jodoin, Carol A. Morrison, & Thomas Fogle

Test security should be a routine component of maintaining any high-stakes examination. Most approaches are forensic in nature and rely on identifying examinees after the testing event for subsequent investigation and action. Unfortunately, at this point, secure and expensive test material has already been exposed. The focus of this presentation is on a proactive, simple, and efficient supplementary approach that limits exposure of test content by identifying examinees during the testing event who are extremely unlikely to pass so that the test may be terminated or alternative content can be administered.

Data were sampled from a high-stakes licensure examination on medical knowledge for examinees testing between May 2008 until May 2012 (N = 122,225). Criteria were developed that incorporate simple scoring procedures related to actual responses, omitted responses, and items not presented (e.g., potential timing or pacing artifacts) both at the section and overall level. Several variations of the above criteria were investigated to span a range of perspectives and tolerances for potential security threats. Given that it is difficult to discern untoward motivations or pre-knowledge, the methodology is targeted at examinees unlikely to pass; low performance regardless of whether the underlying reason is low ability, lack of preparation, or motivations other than meeting the minimum passing standard. Developed criteria were applied to a larger dataset of examinees testing between May 2005 and November 2012 (N = 268,951) to investigate the characteristics of and at what point during the test dispositively low-performing examinees would have been identified.

Results of the simple scoring procedures will be presented. Benefits and limitations of this approach, observations regarding alternative approaches, implications for implementation cost, potential for errors in identifying low performers, effects of minimizing exposure of content for such examinees, and policy consequences for all exam stakeholders will be discussed.

When Numbers are not Enough: Collection and Use of External Information to Assess the Ethics and/or Professionalism of Examinees Suspected of Test Fraud

Author: Marc J. Weinstein

Although reliable statistical evidence of test fraud can, without more, provide a test sponsor with a sufficient basis to cancel an examinee's score, some exam sponsors prefer to obtain and analyze external or collateral information prior to taking any action with respect to an examinee suspected of test fraud. This is especially true where ethics and/or professionalism are an essential component of the core values of the testing organization and a necessary qualification for the examinee to obtain the credential sought. Test sponsors that place a high value on the ethics and/or professionalism of their examinees include, but are not limited to, medical specialty certifying boards, professional licensing agencies and educational organizations that administer examinations for admission to graduate and professional degree programs. Thus, in situations where a test sponsor detects likely test fraud through statistical data analysis, it is critical to such test sponsors to determine not only whether an examinee's score is not a reliable measure of the person's true knowledge of the subject matter tested, but also whether the person engaged in conduct that fails to meet the high standards of ethics and/or professionalism required by the test sponsor. In cases such as this, the test sponsor must identify, preserve and collect reliable external or collateral information and evidence in order to determine whether an examinee has personally engaged in conduct that falls short of the ethics or professionalism required by the sponsoring organization. This paper and presentation will identify and discuss the best methods of identifying, preserving and collecting such external or collateral information, consider the weight that should be given to various types of information gathered, and discuss examples of how such collateral evidence can be used by test sponsors to confidently take action against examinees whose conduct is found to fall short of the sponsor's standards for ethics and/or professionalism.

CONCURRENT SYMPOSIUM 6A: DETECTION OF TEST COMPROMISE

Potential Test Fraud Challenge

Organizer: Dennis Maynes
Team 1: Kim Brunnert, Daniel John Wilson, Brian Bontempo, Cleopatra DeLeon, & Sarah Thomas
Team 2: Yu Zhang & Jiyoon Park
Team 3: Joy Matthews-Lopez & Leslie Rosales
Team 4: Ben Babcock

Braindumps are an especially pernicious problem for many certification and licensure bodies. Techniques are emerging for dealing with and detecting braindump users (see Smith & Davis-Becker, 2011; and Maynes and Burns, 2012). These techniques rely upon security items (Smith & Davis, 2011) or embedded verification items (Maynes & Burns, 2012). More work is needed to understand the false positive rates and the power of these approaches. A few years ago a certification program republished an exam which was known to be compromised with new, non-scored items. The data suggested that the new items were psychometrically unsound. But, this was not true. Contamination in the data by braindump users prevented psychometric analysis of the new items.

For this session, invited participants will analyze this data set. The goals of the analysis will be to identify the individuals who used braindump content, to describe and document the impact of compromise upon the psychometric properties of items (both old and new), and to decontaminate the braindump users from the data set so that a proper psychometric analysis may be performed on the new items. The goal of the session is to provide an opportunity for innovation and sharing of different methods to detect and excise braindump users from data when new items have been added to the exam.

Invited participants will be asked to produce and present a short paper that documents the methods that were used.

CONCURRENT SYMPOSIUM 6B: SECURITY PLANNING

Extending the Role of Psychometrics in Integrated Security Plans

Authors: Lisa O'Leary & Jill Burroughs

Proactive consideration of security throughout the test development process can increase the perceived validity of a testing program by collecting continual evidence that supports the intended use and interpretation of the test scores. An integrated security plan that leverages technology to address program design, legal considerations, content development, and psychometric analyses can be effective in protecting the security of exam content and the integrity of credential decisions. This diversified approach to exam security includes verifying candidate eligibility, protecting intellectual property, establishing parameters for candidate flagging, determining thresholds for differential performance flagging, aligning pilot testing and forms maintenance, and aligning forms maintenance and retake policies. The use of a data repository aids in the timely detection of suspect candidate behavior, routine security tracking and exam maintenance, and in-depth analyses to address specific breeches in security.

This presentation will focus on how psychometrics can be utilized from test design through exam maintenance in a comprehensive security plan. It will present how commonly used techniques to identify suspect candidate and form performance can be incorporated with probabilistic-based analysis methods of security monitoring to enhance the defensibility of a testing program and its candidate decisions. Case studies will be discussed that highlight how technology can be utilized to perform routine security checks in conjunction with more advanced psychometric analyses to develop a systematic, inclusive approach to security. Examples will be given of how differential person and item functioning can supplement other security metrics to aid in candidate enforcement as well as item development and maintenance, forms assembly, and equating (including item banking and anchor sets). The extension of these methods in security analyses enhances a testing program's ability to: support candidate decisions, take action against particular candidates, assess the extent of form/item exposure, evaluate item pools, and determine appropriate next steps for exams.