ETS

Measuring the Power of Learning.®

Research Report
ETS RR–17-23

A Statistical Procedure for Testing Unusually Frequent Exactly Matching Responses and Nearly Matching Responses

Shelby J. Haberman

Yi-Hsuan Lee

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Statistical Procedure for Testing Unusually Frequent Exactly Matching Responses and Nearly Matching Responses

Shelby J. Haberman[1] & Yi-Hsuan Lee[2]

1 Edusoft, Rosh Haayin, Israel
2 Educational Testing Service, Princeton, NJ

In investigations of unusual testing behavior, a common question is whether a specific pattern of responses occurs unusually often within a group of examinees. In many current tests, modern communication techniques can permit quite large numbers of examinees to share keys, or common response patterns, to the entire test. To address this issue, statistical methods are provided to identify examinees in a test with answers that exactly match and to assess whether such exact matches are unusual. In addition, methodology is provided to identify examinees with response patterns unusually similar to circulated keys. Application is made to a testing program.

Keywords Multidimensional item-response theory; multiple comparisons; test security; test collusion; key sharing

doi:10.1002/ets2.12150

In investigations of unusual testing behavior, a common question is whether a specific pattern of responses occurs unusually often within a group of examinees. In the test-security literature, this question has been addressed by statistical methods designed for three types of irregularities: copying, preknowledge, and group collusion (or test collusion; Lee, Lewis, & von Davier, 2014). Traditional investigation into this issue has emphasized answer copying between specific pairs of examinees under investigation (e.g., Angoff, 1974; Holland, 1996; Lewis & Thayer, 1998; van der Linden & Sotaridona, 2006; Wollack, 1997). Residual analysis and person-fit statistics have been proposed to examine aberrant response patterns, or more specially, item preknowledge (e.g., Drasgow, Levine, & Williams, 1985; Karabatsos, 2003; McLeod & Lewis, 1999; McLeod, Lewis, & Thissen, 2003; Meijer & Sijtsma, 2001; Segall, 2002; Shu, Henson, & Luecht, 2013). Residual analysis of both response and timing data has also been considered for the same purpose (e.g., van der Linden & Guo, 2008). More recently, there has been research concerning test collusion, or large-scale sharing of test materials or answers to one or more subsets of items prior to the examination (Belov, 2013, 2014; Wollack & Maynes, 2016; Zhang, Searcy, & Horn, 2011). These studies focused on test collusion due to item preknowledge, which typically occurs when some examinees have access to some items in a test prior to the test administration. Test collusion may involve teachers, school administrators, examinees who collaborate and communicate during the test, or someone sharing test materials on the Internet (Belov, 2013).

In many current tests, modern communication techniques can in some cases permit quite large numbers of examinees to share keys, or common response patterns, to the entire test. This special type of test collusion, termed *key sharing* in this report, has received frequent attention in newspapers (Mytelka, 2010; Richardson, 1996; Strauss, 2014). Generally, *key sharing* is likely to occur in linear tests and multistage tests but not in computerized adaptive tests. The issue of key sharing arises in scenarios that differ from the scenarios involving answer copying and item preknowledge. For instance, as described in Mytelka (2010), Richardson (1996), and Strauss (2014), the keys are usually developed specifically for a test administration and then transmitted to examinees who acquire them. Examinees using or sharing the keys are not limited to the same test location and may have no connection except to the same key sources and keys. It is also possible that the keys are a result of group collusion happening during the test that yields the same response pattern for all items. As a result, statistical approaches to detecting key sharing should analyze responses to the entire test for all examinees in an administration at once, rather than by pairs testing at the same location or subgroups that partition the examinees based on their geographic location (e.g., test center and classroom) or certain relations (e.g., same school). Because the scenario involving key sharing is so different from the scenarios involving answer copying and item preknowledge, existing approaches to examining answer copying and item preknowledge may not be appropriate for detecting key sharing. For

Corresponding author: S. J. Haberman, E-mail: shelbyh@edusoftlearning.com

example, although approaches for detecting answer copying (e.g., Angoff, 1974; Holland, 1996; Lewis & Thayer, 1998; van der Linden & Sotaridona, 2006; Wollack, 1997) have been employed to examine response similarity, they typically rely on specific pairs of examinees. Because detecting groups of any size in the administration involves all possible pairs, which generally leads to a large number of pairwise comparisons, these approaches may not be efficient when applied to an entire administration once the issue of multiple comparisons is taken into account. Methods designed for item preknowledge usually focus on the contrast between compromised items and those that are not in a test, a feature nonexistent in key sharing. The methods considered in Wollack and Maynes (2016) and Zhang et al. (2011) are based on pairs of examinees and therefore may lose power when applied to addressing key sharing. The approaches proposed by Belov (2013, 2014) begin with groups of examinees in an administration, aiming to detect aberrant examinees in each group that may be affected by item preknowledge and to detect the corresponding compromised subset of items. As noted earlier, partitioning examinees into subgroups is not adequate for examining key sharing. Thus alternative statistical methods of identification of shared keys and of examinees who may be using these keys are required.

This report proposes a two-stage procedure for assessing key sharing. Multidimensional item-response theory (MIRT) is employed to aid in both key identification and in identification of examinees who may be employing such a key. More specifically, the proposed procedure begins with identification of examinees in a test with item responses that exactly match. This case is referred to as *exact matching*. A MIRT model that yields probabilities of examinees of different ability levels selecting each of the item choices is employed to assess whether such exact matches on item responses are so unusual that they should be of interest for analysis of test security or for exclusion from equating and linking. Once circulated keys are identified, additional analysis at the second stage is conducted to identify examinees with response patterns unusually similar to the circulated keys. This case is referred to as *near matching*. The proposed procedure is intended to identify examinees in an administration who exhibit unusual similarity in responses to the entire test in groups of any size greater than one. Pairs of examinees exhibiting answer copying can also be identified by the proposed procedure, although it does not require partition of examinees into pairs or groups in advance. In the following section, the MIRT model employed is described. In the next section, computations are provided for the expected number of examinees in a sample or subsample who have identical or nearly identical responses that are not all correct. Next, results are applied to an international test of English proficiency, and analytical results for the power of the procedure are presented. The final section considers policy implications for the results and provides general discussion about sensitivity of the procedure to a number of features studied.

The Multidimensional Item-Response Theory Model

In the analysis of item responses required in this report, it is important to consider not just the item scores for examinee responses but also the specific responses supplied that are incorrect or receive only partial credit. The tests under consideration measure multiple skills, each item is related to a specific skill, and the issue of key sharing is likely to involve multiple skills rather than individual skills. For these purposes, the one-dimensional nominal response model (Bock, 1972) is adapted for use with guessing (Penfield & de la Torre, 2008; Thissen & Steinberg, 1984) with the aid of a type of between-item multidimensional model (Adams, Wilson, & Wang, 1997). Common between-item models can be obtained by use of suitable restrictions on commonly discussed multidimensional item-response models (Reckase, 2007, 2009). Computations required for estimation of item parameters, latent-variable distributions, and examinee probabilities are performed using a stabilized Newton–Raphson algorithm (Haberman, 1988, 2013).

In the model under study, $D \geq 1$ skills are tested, and each item in the test is related to a specific skill. The assessment is taken by I examinees numbered from 1 to I, and J scored items numbered from 1 to J are given to all these examinees. This condition normally applies to tests that are not adaptive. The response for examinee i to item j is denoted by X_{ij}, and $\mathbf{X}_i$ is the J-dimensional vector with elements X_{ij} for $1 \leq j \leq J$. Let $\mathbf{X}_i = \mathbf{x}_T$ if all responses of examinee i are correct (i.e., $\mathbf{x}_T$ is the true key of the test), and let element j of $\mathbf{x}_T$ be x_{jT} for $1 \leq j \leq J$. For each item j is a corresponding skill $d(j)$ between 1 and D. To ensure stable parameter estimation in the model, assume that for each skill there are at least three items. There are $r_j > 1$ possible item responses for item j numbered from 0 to $r_j - 1$, and each response x for item j has a corresponding integer item score $S_j(x)$ between 0 and $S_j(x_{jT})$. For right-scored items, the item score $S_j(x)$ is 1 if $x = x_{jT}$ is the true key of item j and is otherwise 0. If item j is not right-scored, then more than two values of the item score $S_j(x)$ are present. In typical applications, the minimum item score 0 corresponds to a completely wrong response, and the maximum item score $S_j(x_{jT})$ corresponds to a completely satisfactory response. Other item scores reflect partial credit for the answer. In tests in which examinees sometimes fail to answer items, the item response 0 corresponds to an omission. Not distinguishing various reasons for omitted responses (i.e., omitted or not reached) should not be an issue in this context because omissions, especially not-reached items, are unlikely to be observed among the examinees sharing keys and because tests that aid in making high-stakes decisions that do not penalize for guessing typically have relatively few omitted responses.

The model is a latent-structure model with a multivariate normal latent variable $\boldsymbol{\theta}_i$ of dimensional D for each examinee i. The normality assumption is usually robust in real applications (Haberman, 2005; Haberman, von Davier, & Lee, 2008). The pairs $(\mathbf{X}_i, \boldsymbol{\theta}_i)$, $1 \le i \le I$, are assumed independent and identically distributed, and the X_{ij}, $1 \le j \le J$, are conditionally independent given $\boldsymbol{\theta}_i$. The $\boldsymbol{\theta}_i$ are assumed to have a common positive-definite covariance matrix, and to each item response x to item j corresponds an intercept β_{xj}, a slope a_{xj}, and a guessing parameter c_j between 0 and 1, the conditional probability that the item response $X_{ij} = x$ given $\boldsymbol{\theta}_i = \boldsymbol{\omega}$ is

$$p_j\left(x|\,\boldsymbol{\omega}\right) = c_j S_j\left(x\right) + \left(1 - c_j\right) \frac{exp\left(a_{xj}\omega_{d(j)} + \beta_{xj}\right)}{\sum_{x'=0}^{r_j-1} exp\left(a_{x'j}\omega_{d(j)} + \beta_{x'j}\right)}, \tag{1}$$

where $\boldsymbol{\omega}$ is a D-dimensional vector with elements $\omega_{d(j)}$, $1 \le d(j) \le D$. All right-scored items j are assumed to have the same positive guessing parameter $c_j = c < 1$, and $c_j = 0$ for all other items. Introducing a common guessing parameter for the right-scored items permits a bit better fit for high scores, while ensuring stable parameter estimation. Normally, further restrictions are imposed to identify model parameters, but these restrictions are not needed for the estimation of the required probabilities. Nonetheless, in the basic analysis in this report, the simplifying assumption is made that $a_{xj} = a_{x'j}$ if $S_j(x) = S_j(x')$, so that the response X_{ij} and the latent vector $\boldsymbol{\theta}_i$ are conditionally independent given the item score $S_j(X_{ij})$. If $S_j(x) = s$, then the conditional probability $p_{jX|S}(x|s)$ that $X_{ij} = x$ given $S_j(X_{ij}) = s$ has the elementary maximum-likelihood estimate

$$\widehat{p}_{jX|S}\left(x|s\right) = N_{xj}/N^S_{sj}, \tag{2}$$

where N_{xj} is the number of examinees i with $X_{ij} = x$ and $N^S_{sj} > 0$ is the number of examinees i with $S_j(X_{ij}) = s$.

As noted earlier, the model in Equation 1 F is an extension of a number of existing models to accommodate the need to be applicable to multiple skills and the need to consider specific responses supplied that are incorrect or receive only partial credit in addition to item scores. For example, the model reduces to the one-dimensional nominal response model (Bock, 1972) if $c_j = 0$ for all j and $D = 1$. Because each item measures only one skill, the model is called a between-item multidimensional model in the literature and reduces to a type of such models in Adams et al. (1997) if $c_j = 0$ for all items. The unidimensional models proposed in Thissen and Steinberg (1984) and Penfield and de la Torre (2008) adopt different formulations for guessing for nominal responses. Our model is somewhat similar to the formulation in Penfield and de la Torre, but it is applicable to $D > 1$ skills and has a common guessing parameter for different item responses x to all right-scored items. In addition, the model can be used for items that are right-scored and those that are not with different treatments for guessing. It is noteworthy that our MIRT model can be reduced to $D = 1$ for analyses involving only one skill. The standard Rasch model, two-parameter logistic model, and three-parameter logistic model do not serve the purpose because they only consider item scores (correct or incorrect) and cannot differentiate different incorrect choices of an item.

Maximum-likelihood estimates are used in the analyses in this report. The maximum-likelihood estimate of $p_j(x|\boldsymbol{\omega})$ is denoted by $\widehat{p}_j\left(x|\,\boldsymbol{\omega}\right)$. If the J-dimensional $\mathbf{x}$ with elements x_j, $1 \le j \le J$, is in the set $\mathcal{X}$ of possible values of $\mathbf{X}_i$, then the conditional probability that $\mathbf{X}_i = \mathbf{x}$ given that $\boldsymbol{\theta}_i = \boldsymbol{\omega}$ is

$$p\left(\mathbf{x}|\,\boldsymbol{\omega}\right) = \prod_{j=1}^{J} p_j\left(x_j|\,\boldsymbol{\omega}\right). \tag{3}$$

The corresponding estimate is

$$\widehat{p}\left(\mathbf{x}|\,\boldsymbol{\omega}\right) = \prod_{j=1}^{J} \widehat{p}_j\left(x_j|\,\boldsymbol{\omega}\right). \tag{4}$$

If $f(\boldsymbol{\omega})$ denotes the probability density of $\boldsymbol{\theta}_i$ at $\boldsymbol{\omega}$, then the probability $p(\mathbf{x})$ that $\mathbf{X}_i = \mathbf{x}$ is

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\omega}) f(\boldsymbol{\omega})\, d\boldsymbol{\omega}. \tag{5}$$

If $\widehat{f}(\boldsymbol{\omega})$ denotes the estimated probability density of $\boldsymbol{\theta}_i$ at $\boldsymbol{\omega}$, the maximum-likelihood estimate of $p(\mathbf{x})$ is

$$\widehat{p}(\mathbf{x}) = \int \widehat{p}(\mathbf{x}|\boldsymbol{\omega}) \widehat{f}(\boldsymbol{\omega})\, d\boldsymbol{\omega}. \tag{6}$$

The MIRT analysis is conducted with a general program for item-response analysis (Haberman, 2013) that computes $\widehat{p}\left(\mathbf{X}_i\right)$ for each examinee i. These estimated probabilities provide the basis for the procedure discussed in the following section. The MIRT program is freely available from the authors for noncommercial use.

Probabilities of Matching Responses

Analysis of Exact Matching

In a test with a substantial number of items, relatively few examinees will have the exact same response vector $\mathbf{X}_i$, because the set $\mathcal{X}$ of possible values of $\mathbf{X}_i$ is extremely large. This observation can provide a basis for identifying examinees who may be using a common key. For this purpose, a separate investigation is undertaken for each examinee in a set C of $m \le I$ examinees requiring attention. If the entire administration is examined at once, then C is the set of all examinees in the administration and $m = I$. If only a single test center is of interest, then m may be a far smaller positive integer, and C is the set of examinees in the test center. For each examinee i in C, the number M_i of examinees i' in C, $i' \neq i$, is computed for whom $\mathbf{X}_{i'} = \mathbf{X}_i$, so that each item response of examinee i' is the same as the corresponding item response of examinee i. Given the observed examinee response $\mathbf{X}_i$, the probability under the model that M_i or more examinees i' in C, $i' \neq i$, would have $\mathbf{X}_{i'} = \mathbf{X}_i$ is just the binomial probability

$$B_i = \sum_{n=M_i}^{m-1} \binom{m-1}{n} \left[p\left(\mathbf{X}_i\right)\right]^n \left[1 - p\left(\mathbf{X}_i\right)\right]^{m-n-1}, \tag{7}$$

where $p(\mathbf{X}_i)$ is given in Equation 5. It is worth noting that the current application examines the entire administration at once, so that $p(\mathbf{X}_i) = p(\mathbf{X}_i|C)$, where C includes all examinees. In general, Equation 7 is correct if the examinees in C can be regarded as representative of examinees in the administration. Under the model, for any u in the unit interval, u is no less than the probability $P(B_i \le u)$ that $B_i \le u$. By the Bonferroni inequality, the probability that $B_i \le u$ for any examinee i in C does not exceed mu. The estimated value of B_i is

$$\widehat{B}_i = \sum_{n=M_i}^{m-1} \binom{m-1}{n} \left[\widehat{p}\left(\mathbf{X}_i\right)\right]^n \left[1 - \widehat{p}\left(\mathbf{X}_i\right)\right]^{m-n-1}, \tag{8}$$

where $\widehat{p}\left(\mathbf{X}_i\right)$ is given in Equation 6 and estimated based on all examinees in the administration. Thus the Bonferroni significance level for examinee i for the number of matching responses is the minimum of $m\widehat{B}_i$ and 1. Examinees in C with the same response vector $\mathbf{X}_i$ have identical $m\widehat{B}_i$ values.

Finding the number of examinees with the same response vector is readily accomplished by sorting observations by $\mathbf{X}_i$ for examinees i in C. Such sorting is used to complete identification of all distinct response vectors and their frequency of use and can be easily accomplished in SAS. When a number of examinees share the same response vector and the corresponding values of $m\widehat{B}_i$ are very small for these examinees, this common response vector is strong evidence of a key being circulated among a group of examinees.

It should be noted that strong evidence that a key is circulated does not by itself imply that every examinee with that response pattern is using the circulated key. To examine individual connection to that circulated key, additional investigation should be conducted to gather information concerning individual behaviors.

In addition, small values of $m\widehat{B}_i$ may be observed in a large group with a not uncommon response pattern $\mathbf{X}_i$, which is likely to occur for examinees with almost perfect scores. Because those examinees are more likely to match the response patterns by chance, a more conservative approach is taken in our analysis by not identifying such response patterns as circulated keys and not flagging those examinees. For this purpose, a further test is implemented. Let $\mathcal{X}_K$ be a set of response patterns $\mathbf{x}_K$ identified as circulated keys, where $\mathbf{x}_K$ is of dimension J with elements x_{jK} for $1 \le j \le J$. Let $\mathcal{X}_K$ have k elements. For each $\mathbf{x}_K$ in $\mathcal{X}_K$, the probability that an examinee i has a response $\mathbf{X}_i = \mathbf{x}_K$ is estimated to be $\widehat{p}\left(\mathbf{x}_K\right)$, so that the estimated probability that any examinee in C has this response pattern does not exceed $m\widehat{p}\left(\mathbf{x}_K\right)$. If both $m\widehat{p}\left(\mathbf{X}_i\right) \le 0.01$ and $m\widehat{B}_i \le 0.01$ with $\mathbf{X}_i = \mathbf{x}_K$, then examinee i is identified as using the circulated key $\mathbf{x}_K$. If the number k of circulated keys is relatively large, say, greater than five, then one might change the requirement that $m\widehat{p}\left(\mathbf{X}_i\right) \le 0.01$ to the requirement that $mk\widehat{p}\left(\mathbf{X}_i\right) \le 0.01$ to account for the number of circulated keys examined with Bonferroni corrections. In general, in cases in which $\widehat{p}\left(\mathbf{X}_i\right)$ is greater than $\widehat{B}_i$, inferences about individual examinees with a given response pattern may differ from inferences about the existence of a circulated key with that response pattern.

Analysis of Near Matching

So far, analysis has considered identifying a circulated key from the data. Once a circulated key is identified from the data or obtained from material found with an examinee, there is an added issue of examinees who do not always apply the circulated key correctly. This case is referred to as near matching in this report. Such error may occur if (a) the examinees apply the circulated key under high pressure in the testing room or (b) they need to memorize the circulated key rather than bringing a copy of the circulated key. Thus additional analysis is considered to compare a circulated key $\mathbf{x}_K$ with the response pattern of every examinee in C. A simple approach is to score the test of each examinee based on the circulated key $\mathbf{x}_K$ and based on the true key $\mathbf{x}_T$ and then compare if the score on the circulated key is significantly better than the score on the true key. Recall that x_{jK} and x_{jT}, $1 \le j \le J$, are elements of $\mathbf{x}_K$ and $\mathbf{x}_T$, respectively. Let the alternative item score $S_{jK}(x)$ for response x to item j be $S_j(x_{jT})$ if $x = x_{jK}$, and let $S_{jK}(x)$ be $S_j(x_{jK})$ if $x = x_{jT}$. For other values of x, let $S_{jK}(x) = S_j(x)$. One may then compare the total score

$$T_{iK} = \sum_{j=1}^{J} S_{jK}\left(X_{ij}\right) \tag{9}$$

to the conventional total

$$T_i = \sum_{j=1}^{J} S_j\left(X_{ij}\right). \tag{10}$$

A better score on the circulated key compared to the true key may suggest that the circulated key is being used. Examining this issue with a significance test relies on a conditional inference. Let J_K be the set of items j for which $S_{jK}(x)$ and $S_j(x)$ are the same for all responses x. For example, in one circulated key found in one of the test administrations discussed in the section Two Illustrations of the Analysis, there were nine items where the true key and the circulated key apparently being used by a group of examinees were different and 67 items where the true and circulated keys were the same. (Eight of the nine items were right-scored; the other one had possible scores 0, 1, and 2, and the circulated key received partial credit on that item.) The set J_K then consists of these 67 items where the keys coincide. The score T_{iK} is the sum of the 76 item scores obtained with the circulated key, whereas T_i is the sum of item scores using the true key. The difference in this example between T_{iK} and T_i is then an integer between -9 and 9. The estimate $\widehat{P}_{iK}$ is then computed for the conditional probability P_{iK} under the model that, given X_{ij}, j in J_K, $T_{iK} - T_i$ would be at least as large as the observed difference. In the example, if, for examinee i, the difference $T_{iK} - T_i$ is 7, then the probability computed is the estimated conditional probability under the model that $T_{iK} - T_i \ge 7$ given the observed responses X_{ij} for items j in the set J_K of 67 items for which the apparently circulated key and the true key coincide. To complete the analysis, let C' consist of examinees in C not already identified as exactly matching an unusually large number of examinees, and let m' be the number of examinees in C'. As in the case of exact matches, the Bonferroni inequality is then employed, so that for each examinee in C', the significance level used is $m'\widehat{P}_{iK}$ for examinee i. To take into account that the examinees' responses would be compared with multiple circulated keys, one could further adjust the significance level $m'\widehat{P}_{iK}$ by multiplying by the number of circulated keys considered in the analysis (i.e., $m'k\widehat{P}_{iK}$). Because the circulated keys identified for the test studied in the following section are often variants of one or two main circulated keys, this further adjustment to the proposed methodology can be quite conservative.

Two Illustrations of the Analysis

Two administrations of an international test of English proficiency were examined. In each case, only listening and reading sections were used, so that there were $D = 2$ skills. For both administrations, C was the entire set of examinees in the administration. Lists were made of examinees with $m\hat{B}_i \leq 0.01$ and $m\hat{p}\left(\mathbf{X}_i\right) \leq 0.01$. These lists were then used to identify apparent circulated keys. For purposes of illustration, the matching responses for three or more examinees were used as a circulated key to further identify nearly identical matches.

There were 34 listening items and 42 reading items in both administrations. In the first case, counting missing responses, there were 5–15 possible responses to a given listening item. All of the items were right-scored. In addition, for 39 reading items, there were five possible responses, and right-scoring was used. For three reading items, there were either 36, 39, or 42 possible responses, and possible item scores were 0, 1, and 2. In this administration, approximately 19,000 examinees were considered. Only nine examinees were identified. Using the Bonferroni significance levels, a pair of examinees with matching responses had a significance level of about 8×10^{-14}. For this pair, $m\hat{p}\left(\mathbf{X}_i\right)$ was only 4×10^{-18}, so that very strong evidence existed that the match was not coincidental. A second group of seven examinees with the same responses had a significance level of about 3×10^{-8}. For these examinees, $m\hat{p}\left(\mathbf{X}_i\right)$ was about .03, so that evidence about individual examinees in the group was somewhat weaker than evidence for the existence of a circulated key. Thus the list for this case only consisted of the pair of examinees. Had the matching responses for the seven examinees been used as a circulated key, no further examinees would have been identified by comparison of the circulated and true keys with a Bonferroni significance level less than .01. As will be discussed in the concluding section, the presented procedures are very sensitive to the quality of the circulated key. In this case, the circulated key associated with only two examinees had a far lower significance level than the circulated key associated with seven examinees. The principal difference was that the circulated key for the two examinees corresponded to a T_{iK} of 63, whereas the circulated key for the seven examinees corresponded to a T_{iK} of 77, a value not far from the maximum possible value of T_i of 79.

In the second case, approximately 10,000 examinees were studied. For 33 listening items, counting missing responses, there were either 5, 8, or 13 possible responses to a given item, and these items were right-scored. There was one listening item with 149 possible responses, and possible item scores were 0, 1, and 2. In addition, 39 of the 42 reading items had five possible responses, and right-scoring was used. For three reading items, there were either 35, 39, or 42 possible responses, and possible item scores were 0, 1, and 2. In this administration, four examinees were identified. Using the Bonferroni significance levels, one pair of examinees with matching responses had a significance level of approximately 2×10^{-7}. For this pair, $m\hat{p}\left(\mathbf{X}_i\right)$ was only 2×10^{-11}, so that very strong evidence existed that the match was not coincidental. The second pair of examinees with the same responses had a significance level of approximately 2×10^{-21}. For these examinees, $m\hat{p}\left(\mathbf{X}_i\right)$ was about 2×10^{-25}, so that individual connections to the circulated key were clear. The matching responses for these two pairs were not used further to flag examinees with unusually strong performance on a circulated key. Thus the list for the second administration contained these four examinees.

Power Calculation

The most elementary approach to considering the power of the procedure is to consider the probability of detection of a random group of examinees who copy from each other without error (i.e., exact matching). Suppose the group has $G \geq 2$ members, and suppose that the group common responses have the same distribution as for the general population. It is assumed that G is small relative to the total sample size and that all other responses of examinees are obtained by independent work, so that any distortion in parameter estimation caused by the group is negligible. With the significance level .01 used in the example, the probability of detection of this group is then well approximated by the joint probability that

$$m \sum_{n=G-1}^{m-1} \binom{m-1}{n} \left[p\left(\mathbf{X}_i\right)\right]^n \left[1 - p\left(\mathbf{X}_i\right)\right]^{m-n-1} \leq .01 \tag{11}$$

and

$$mp\left(\mathbf{X}_i\right) \leq .01. \tag{12}$$

Estimation of this joint probability is straightforward. If g is a real function on [0, 1], then $E(g(p(\mathbf{X}_i)))$ has the estimate

$$\widehat{E}\left(g\right)=m^{-1}\sum_{i\in C}g\left(\widehat{p}\left(\mathbf{X}_i\right)\right), \tag{13}$$

where $\widehat{p}\left(\mathbf{X}_i\right)$ is computed for each examinee i in the administrations. In this example, the function $g(p(\mathbf{X}_i))$ is 1 if Equations 11 and 12 both hold, and $g(\mathbf{X}_i)$ is 0 otherwise. In other words, the estimated probability of detection $\hat{E}(g)$ is the fraction of examinees in C who satisfy conditions in Equations 11 and 12 when $G=2, 3, \ldots, m-1$. For the first illustrative example, the estimated probability of detection is .972 for $G=2$, .991 for $G=3$, and .994 for any $G\geq 4$. The constant value for larger G is due to the condition in Equation 12. Similar results are observed in the second illustrative example. Estimated probabilities are .951 for $G=2$, .982 for $G=3$, and .987 for $G\geq 4$.

The Bonferroni procedure used to detect exact matches is very unlikely to detect any case when the model holds for all observations (i.e., every examinee works independently). In this case, the probability of detecting a random group of 2 examinees does not exceed the expectation of $g(p(\mathbf{X}_i))$, where $g(p(\mathbf{X}_i))$ is 0 unless Equations 11 and 12 both hold for $G=2$, and $g(p(\mathbf{X}_i))$ is otherwise equal to $m(m-1)p(\mathbf{X}_i)/2$. In this case, $\hat{E}(g)$ is only .00002 in the first illustrative example and .00003 in the second illustrative example. The power of the procedure for detecting a random group of examinees who copy from each other with errors (i.e., near matching) can be similarly evaluated and is not further discussed.

Conclusions

The methodology employed is able to identify examinees who exhibit unusual similarity in responses by a two-stage procedure that involves exact matches followed by a study of unusual similarity to groups of examinees with exact matches. The methodology is quite conservative because of the conservativeness of the Bonferroni adjustment, so that one should distinguish between its use for test security for identification of specific examinees, its use in test security to identify test centers or other units of investigation with unusually frequent patterns of identical and near-identical responses, and its use for test analysis and linking. In many applications involving test security, it is necessary to state that behavior exhibited by a group of examinees is so unusual that it cannot reasonably be expected to have occurred by chance. In these applications, the Bonferroni inequality is vital. Application of these results by a testing organization may vary. Legal considerations are obviously important, and factors such as protection of the public and the rights of examinees should be considered as appropriate. It is desirable, when possible, to gather additional information concerning examinee behavior related to the test to decide on appropriate action.

In the case of test analysis or in the case of a general investigation about the frequency of use of circulated keys, a much less stringent criterion is likely appropriate, especially in terms of better performance on circulated keys than on true keys. For equating, the cost of the resulting reduction in sample size is readily balanced by the desire to have examinees who appear to be taking a test in good faith. To assess the size of the problem with circulated keys, it is not necessary to conclusively demonstrate apparent use of circulated keys by any specific examinee.

The procedures discussed are very sensitive to the quality of the circulated key and the size m of the group in question. It is typically the case that the probability $p(\mathbf{x}_K)$ associated with a circulated key $\mathbf{x}_K$ is smaller if the set of items not in J_K is relatively large. Indeed, for each illustrative administration considered, the R^2 statistic for regression of $\log p(\mathbf{X}_i)$ on T_{iK} is approximately .9. To illustrate this issue, recall that in the first administration discussed in the section Two Illustrations of the Analysis, the circulated key associated with only two examinees had a far lower significance level than the circulated key associated with seven examinees. This result was due to clear differences in key quality; that is, $T_{iK}=63$ for the two examinees, whereas $T_{iK}=77$ for the seven examinees, given that $T_i=79$. It is noteworthy that a case with three examinees with the exact same responses in the first administration and a common T_{iK} of 76 resulted in a significance level of .013, which was above the threshold used. On the other hand, had the set C just been the set of 224 examinees in the country in which the three examinees were tested, then the significance level would have changed to approximately 2×10^{-8}. For each of these three examinees, $m\widehat{p}\left(\mathbf{X}_i\right)$ would have been only approximately .00001.

It is reasonable to consider the sensitivity of results to models used and to subsets of examinees considered. In one examination, the constant guessing parameter c was set to 0 for all items rather than just for items with more than two scores. Changes in status (i.e., flagged or not) involved examinees with estimated Bonferroni significance levels for exact or near matching relatively close to .01. In addition, estimation of the probabilities $p(\mathbf{X}_i)$ is performed here with both examinees who appear to use circulated keys and examinees who do not. It is reasonable to expect that this procedure reduces the number of examinees identified. To assess the impact, one might eliminate the examinees identified by the matching or near-matching criterion during estimation and then apply the results to the entire sample of examinees. Generally, the results are expected to differ more substantially with more examinees identified in the initial run.

Acknowledgments

This work is fully collaborative. Authors are listed in alphabetical order. The authors are grateful to Charlie Lewis and Robert Mislevy for comments on relevant issues. Any opinions expressed in this report are those of the authors and not necessarily of Educational Testing Service.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. https://doi.org/10.1177/0146621697211001

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69*(345), 44–49. https://doi.org/10.1080/01621459.1974.10480126

Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement, 50*, 141–163. https://doi.org/10.1111/jedm.12008

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing, 2*(3), 37–58. https://doi.org/10.7333/1410-0203037

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51. https://doi.org/10.1007/bf02291411

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Haberman, S. J. (1988). A stabilized Newton–Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology, 18*, 193–211. https://doi.org/10.2307/271049

Haberman, S. J. (2005). *Latent-class item response models* (Research Report No. RR-05-28). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02005.x

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (Research Report No. RR-96-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1996.tb01685.x

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298. https://doi.org/10.1207/s15324818ame1604_2

Lee, Y.-H., Lewis, C., & von Davier, A. A. (2014). Monitoring the quality and security of multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 285–300). New York, NY: CRC Press.

Lewis, C., & Thayer, D. (1998). *The power of the K-index (or PMIR) to detect copying* (Research Report No. RR-98-49). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01798.x

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147–160. https://doi.org/10.1177/01466219922031275

McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*, 121–137. https://doi.org/10.1177/0146621602250534

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135. https://doi.org/10.1177/01466210122031957

Mytelka, A. (2010, January 18). In alleged scheme, SAT was sent from Thailand via South Korea to Connecticut. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/blogs/ticker/in-alleged-scheme-sat-was-sent-from-thailand-via-south-korea-to-connecticut/20557

Penfield, R. D., & de la Torre, J. (2008, March). *A new response model for multiple-choice items*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. *26*, pp. 607 – 642). Amsterdam, Netherlands: North-Holland. https://doi.org/10.1016/S0169-7161(06)26018-8

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-89976-3

Richardson, L. (1996, October 29). Time-zone caper: Suspect is arrested in testing scheme. *New York Times*. Retrieved from http://www.nytimes.com/1996/10/29/nyregion/time-zone-caper-suspect-is-arrested-in-testing-scheme.html

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics, 27*, 163 – 179. https://doi.org/10.3102/10769986027002163

Shu, Z., Henson, R. A., & Luecht, R. M. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika, 78*, 481 – 497. https://doi.org/10.1007/s11336-012-9311-3

Strauss, V. (2014, November 16). The six-step SAT cheating operation in Asia and how to stop it. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2014/11/16/the-six-step-sat-cheating-operation-in-asia-and-how-to-stop-it/

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501 – 519. https://doi.org/10.1007/bf02302588

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365 – 384. https://doi.org/10.1007/s11336-007-9046-8

van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*, 283 – 304. https://doi.org/10.3102/10769986031003283

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21*, 307 – 320. https://doi.org/10.1177/01466216970214002

Wollack, J. A., & Maynes, D. M. (2016). Detection of test collusion using cluster analysis. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 124 – 150). New York, NY: Routledge.

Zhang, Y., Searcy, C. A., & Horn, L. (2011, April). *Mapping clusters of aberrant patterns in item responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Suggested citation:

Haberman, S. J., & Lee, Y.-H. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses* (Research Report No. RR-17-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12150

Action Editor: Matthias von Davier

Reviewers: Charles Lewis and Robert Mislevy