Research Article doi: 10.12973/eu-jer.8.4.1307

European Journal of Educational Research
Volume 8, Issue 4, 1307 - 1322.
ISSN: 2165-8714
http://www.eu-jer.com/

Implementation of the Omega (ω) Index to Detect Large-Scale Systematic Cheating

Alvin Vista*
Australian Council for Educational Research,
AUSTRALIA

Received: July 10, 2019 ▪ Revised: October 8, 2019 ▪ Accepted: October 10, 2019

Abstract: Cheating detection is an important issue in standardized testing, especially in large-scale settings. Statistical approaches are often computationally intensive and require specialised software to conduct. We present a two-stage approach that quickly filters suspected groups using statistical testing on an IRT-based answer-copying index. We also present an approach to mitigate data contamination and improve the performance of the index. The computation of the index was implemented through a modified version of an open source R package, thus enabling wider access to the method. Using data from PIRLS 2011 (N=64,232) we conduct a simulation to demonstrate our approach. Type I error was well-controlled and no control group was falsely flagged for cheating, while 16 (combined n=12,569) of the 18 (combined n=14,149) simulated groups were detected. Implications for system-level cheating detection and further improvements of the approach were discussed.

Keywords: *Answer-copying indices, item response theory, PIRLS, cheating detection, standardized testing, test integrity.*

To cite this article: Vista, A. (2019). Implementation of the omega (ω) index to detect large-scale systematic cheating. *European Journal of Educational Research*, *8*(4), 1307-1322. https://doi.org/10.12973/eu-jer.8.4.1307

Introduction

Cheating on standardised assessments can occur at an individual level or systematically (e.g., classroom-, school-, or even district-level). In both situations, cheating has serious consequences and therefore there is a long history of various methods for detecting cheating. Chajewski and colleagues (2014) provide a comprehensive discussion of various methods of fraud detection and macro level screening systems. These methods can be categorised into two main groups: 1) non-statistical techniques (e.g., surveillance, audit and monitoring, field reports) and 2) statistical techniques (e.g., checking for unusual changes in performance over time, and data-mining).

In the context of large-scale testing that is taken only *once* by an individual, non-statistical techniques can be impractical and some statistical methods such as trend-analysis are not applicable. It is logistically impractical to undertake large-scale surveillance and there are limited sources of supplementary performance data to use for statistical techniques based on comparisons over time. Although there are other non-statistical methods that may be useful in large-scale testing, and we recommend that several complementary techniques be used, the scope of this paper is limited to statistical techniques. The method discussed in this paper therefore focuses on data-mining of responses from a single test administration such as end-of-year examination.

At a pairwise level, evidence for cheating is based on observation of matching correct and/or matching incorrect answers from the source to the copier. Several indices for detecting answer copying have been developed over the years with varying degrees of effectiveness. Below are examples of answer-copying indices that have been reported in the literature as among the most effective:

- Omega (ω) index (Wollack, 1997)
- Generalized binomial test (GBT) (van der Linden & Sotaridona, 2006)
- S1 and S2 indices (Sotaridona & Meijer, 2003) and the conceptually similar precursors, the K index (Holland, 1996) and K-variants (Sotaridona & Meijer, 2002)

* **Correspondence:**
Alvin Vista, Teach For All, Perth, WA, Australia. ✉ alvin.vista@teachforall.org

The ω and GBT indices are based on item response theory (IRT) while the S1 and S2 indices are based on classical test theory (CTT). Individual-level statistical approaches often provide more comprehensive evidence than broader group-based statistical approaches such as score distribution analyses. Therefore we are using an approach that was designed to detect cheating at an individual level but adapting it for large scale implementation.

For this study, the ω index was chosen as the most appropriate for the data and because it is well-established as among the most useful indices in the literature. The mathematical details of the ω index are described below but additional details of the other indices can be found in the literature (see Holland, 1996; Sotaridona & Meijer, 2003; van der Linden & Sotaridona, 2006).

The ω index is computed by standardising the difference between the expected number of answer matches and the observed number of matches for a given pair of students. This given pair will be designated as the "copier" (C) and the "source" (S) hereafter. The probability of the expected number of matches – given S's answers, C's ability, and item parameters – is computed via IRT.

The IRT-based indices take into account the estimated latent abilities of the individuals being compared in computing the expected matches, whereas CTT-based indices rely on raw score as the main indicator of ability and compute the probabilities of matching items based on either the binomial (for K and K-variants) or Poisson (for S2) distributions without taking into account the latent abilities of the examinee pair (see Holland, 1996; Sotaridona & Meijer, 2002). The GBT index, while IRT-based, models the probability of an answer match using also the binomial distribution. The IRT-based indices work with dichotomously scored items (i.e., scored correct-incorrect) with the latent ability estimated using 2-3PL models† (Birnbaum, 1968) or with raw responses with the ability estimated using a Nominal Response Model (NRM).

In particular, using the NRM (Bock, 1972), the probability for C responding with S's response for item $i(U_{iS})$ is given as:

$$P_{\mathrm{C}}(U_{iS}) = \frac{\exp(\xi_{iS} + \lambda_{iS}\theta_{\mathrm{C}})}{\sum_{m=1}^{K}\exp(\xi_{im} + \lambda_{im}\theta_{\mathrm{C}})}$$

where ξ and λ represent the item intercept and slope parameters respectively; K and m represent the number of response categories and response category indicator respectively; and θ_{C} represents C's ability.

The ω index is then calculated as:

$\omega = \frac{h_{\mathrm{CS}} - \sum_{i=1}^{n} P_{\mathrm{C}}(U_{iS})}{\sigma_{h_{\mathrm{CS}}}}$, where h_{CS} represents the observed match between C and S across n total number of items. The standard error of the difference between the expected and observed matches is computed as:

$$\sigma_{h_{\mathrm{CS}}} = \sqrt{\sum_{i=1}^{n} P_{\mathrm{C}}(U_{iS})\left(1 - P_{\mathrm{C}}(U_{iS})\right)}$$

The values of ω are asymptotically normally distributed and standard null hypothesis statistical testing can be applied, where the null hypothesis is rejected for values of ω greater than a critical value (Wollack, 2004).

The logic behind all answer-copying indices is that the evidence of cheating can be found in matching response patterns (either matching correct or incorrect items) that are *beyond* the expected number of matches. That is, the indices statistically model if answer-matches have occurred beyond what can be attributed to random chance.

Large-scale implementations of cheating detections method based on IRT-based answer-copying indices have limitations that fall into two main areas: psychometric and logistic limitations. These limitations and proposed ways to mitigate them are discussed below.

Psychometric limitations and proposed mitigation

The main psychometric limitation of this method – indeed all IRT-based models – is that the ability estimates are contaminated by systematic error in the response pattern (Wollack & Cohen, 1998). In the context of cheating analysis, the response pattern of the copier affects the ability estimates and therefore confounds the probability of expected matches.

The impact of contaminated response pattern on the estimation of item and person parameters has long been investigated. Levine and Drasgow (1982) have shown that aberrant-response detection methods‡ are robust even when the aberrant response patterns were included in the IRT parameter estimation. However, this may be true when the

† In 2- and 3PL models, unlike Rasch or 1PL models, the raw score is no longer a sufficient statistic for estimating latent abilities and the complete response pattern needs to be taken into account.

‡ Levine & Drasgow were investigating "appropriateness indices", which can be considered analogous to answer-copying indices, albeit at an individual level rather than between pairs. These indices (see Levine & Rubin, 1979) measure the extent to which a response pattern in a multiple-choice test is not an appropriate measure of ability due to various reasons including, but not limited to, cheating.

context is individual-level cheating. When the proportion of aberrant response is low relative to the population, the impact might not be substantial. The magnitude of contamination might be different if the context is systematic cheating, in which cheating is prevalent across classes, schools, or even districts. Also, even if the item parameter estimation might be robust against aberrant response patterns, ability estimates are directly determined by the response patterns regardless of whether these are honest or not.

To mitigate this psychometric limitation, theta can be estimated using additional information. Estimating theta, either through maximum likelihood or Bayesian estimators, would be as follows:

$\hat{\theta}_j = Y\left(\theta|\boldsymbol{\xi}, \mathbf{U}_j\right)$, where $\boldsymbol{\xi}$ is the matrix of item parameters (e.g., ξ and λ), $\mathbf{U}_j$ = (U_{1i}, U_{2i}, …U_{ji}) is the vector of item responses to *i* items by examinee *j*, and the expectation *Y* refers to either the maximum likelihood or the mean or mode of the posterior distribution of theta (EAP or MAP respectively) depending on the type of estimator.

Theta is latent and therefore it can be conceptualised as missing data that is estimated from a set of observed data (Bock & Aitkin, 1981; Rubin, 1987). Because θ is unobservable, $\mathbf{U}_j$ is our only observed and yet always incomplete data (Dempster, Laird, & Rubin, 1977).

The accuracy of theta can therefore be improved by increasing the number of items used to estimate it. Viewed another way, we can use a fuller set of test response data ($T_{\text{augmented}}$) to estimate theta even if we are only concerned with a subset ($T_{\text{investigated}}$) for our cheating analysis purposes. Thus, the vector of item responses includes *more* items than the set of items used for computing $\sum_{i=1}^{n} P_{\text{C}}(U_{i\text{S}})$, $i\left\{i \in T_{\text{investigated}}\right\}$:

$\mathbf{U}_j$, for items $i\left\{i \in T_{\text{augmented}}\right\}$ where $T_{\text{investigated}} \subseteq T_{\text{augmented}}$; and provided item independence holds, any items that belong to the relative complement of $T_{\text{investigated}}$ in $T_{\text{augmented}}$ can be used to augment the estimation of theta.

Augmenting the information needed to estimate the latent trait mitigates another related limitation. All answer-copying indices suffer from an inherent weakness when applied to groups with ability that is more homogeneous especially if they are on the tails of the distribution – that is, a group composed of uniformly high or low performers (Sunbul & Yormaz, 2018). Because all answer-copying indices are based on matching answers, as the group becomes more homogeneous the chances of matching answers increase even without copying. It becomes more difficult to discriminate between honest and dishonest matches in more homogeneous groups with performance in the tails of the distribution. The extent of homogeneity can be decreased as we include more items in estimating theta.

For the target (or investigated) test itself, having a more diverse set of items (so that the variance in test score is increased) and increasing the test length has been found to improve detection rates (Sotaridona & Meijer, 2002; Wollack, 1997). Answer-copying indices also work better when the test items are more difficult because homogeneous groups of students with lower ability are less likely to be misdetected (Zopluoglu & Davenport, 2012). For those who are uniformly low in ability, there is a higher likelihood of guessing and so it is less likely that the few items they have answered correctly will be the same matching items. For the group with high ability, however, it is more likely that students will be incorrect on the same (difficult) items as well as a lower chance of guessing on easier items.

Logistic limitations and proposed mitigation

The next main limitation, concerning logistics, becomes increasingly important as the scale that this method is applied becomes larger. The answer-copying indices were originally designed to detect cheating on an individual basis and requires the specific designation of the source and cheater. The indices would yield different values for A = source and B = cheater versus B = source and A = cheater. This implies that to use these indices in a large-scale setting and in the context of systematic cheating, the indices would need to be computed for all possible pair permutations. The number of computations required increases very quickly as can be seen by the number of pair permutations given by the formula:

$\frac{g!}{(g-k)!}$, where *g*=sample of students and *k*=2 because we are selecting pairs.

In the context of systematic cheating, such as in a situation where cheating occurs at school-level, testing all possible pairs in a *single* school with 500 students in any grade level would require computation of the indices for $\frac{500!}{(500-2)!}$ = 249,500 pair permutations. Although schools of this size may seem unusually large for developed countries, even *classes* often have more than 100 students in the developing world (Benbow et al., 2007). Even disregarding pair-ordering, it would still require computing $\frac{g!}{k!(g-k)!} = \frac{500!}{2!(500-2)!}$ = 124,750 index values. In systems with thousands of schools, the number of pair combinations can quickly run to the hundreds of millions. Perhaps this is not an insurmountable challenge for developed countries, but many countries in the developing world[§] may have more

[§] As a recent example, large-scale cheating has been reported as common in Bihar, India a state with population larger than many countries in Europe (British Broadcasting Corporation, 2015).

constrained time and computational resources – precisely the countries that lack more extensive capabilities to prevent cheating and where the impact of systematic cheating can be more damaging.

While approaches based on answer-copying indices have been reported in the literature for several decades, computing these indices previously required specialised code. It was relatively recently that packages based on open-source languages have been available to specifically compute some of these indices. We adapted **CopyDetect** (Zopluoglu, 2013), an R language (R Core Team, 2016) package, to compute the ω index that we report here. The computation of the item parameters and estimation of the thetas were done using another R package, **mirt** (Chalmers, 2012). The computations involved are substantial:

- The parameters ξ and λ need to be estimated per response category per item
- The ability estimate $\hat{\theta}$ needs to be calculated per student
- The probability for C responding with S's response, $P_C(U_{iS})$, needs to be calculated per item per student
- The observed match, h_{CS}, needs to be counted per pair
- The ω index and corresponding *p*-value need to be calculated per pair

This is why it is important to make the process as efficient as possible. To implement a method using IRT-based indices on large-scale systematic cheating contexts, we propose that the investigation proceeds in stages, with each stage filtering possible suspects efficiently. The proposed two-stage approach is summarised as follows:

- Stage 1
  - Compute the ω index values for all students paired with a "supersource" – a hypothetical student who has a nearly-complete test key for the multiple choice questions and therefore a "source" for the cheater. For this stage, statistical significance was set at α=.05. Cases with statistically significant ω index values were designated as suspected.
  - Flag groups which have proportion of suspected students greater than the nominal Type I error rate for Stage 1.
- Stage 2
  - Create a set of set of all pair combinations from each flagged group in Stage 1: $\binom{g}{k} = \frac{g!}{k!(g-k)!}$ pairs per group. Where g=number of cases in the group, and k=2.
  - Randomly sample 500 pairs from $\binom{g}{k}$ – the set of all pair combinations per group.
  - Compute ω index values for these sampled pairs from each of flagged group in Stage 1. For Stage 2, statistical significance was set more conservatively at α=.01.
  - Flag groups which have the proportion of suspected students greater than the nominal Type I error rate for Stage 2.

The first stage requires a conceptual shift where instead of attempting to detect all pairs copying from each other, we first only attempt to detect groups of students copying from a single source and assume that cheating is largely successful. Differing from individual copying, we assume that if there is *systematic* cheating, the copied answers are almost always the correct answer (such as system-level cheating through test-form or even answer-key leakage). In the exploratory Stage 1, we base all pair comparisons as having a single source (i.e., everybody is copying from one person, with varying degrees of copying effectiveness). We can think of this source as having the test key, or having a role of a "supersource", where all or most answers are correct. Therefore, instead of computing $\frac{g!}{(g-k)!}$ or even $\frac{g!}{k!(g-k)!}$ index values, we only need to compute g values per group, massively reducing the amount of computations involved by a factor of $\frac{g-1}{2}$ for pair combinations (i.e., k=2).

This approach will yield only a preliminary evidence of cheating given the less stringent alpha level (i.e., a more liberal criterion for statistical evidence of cheating). However at this first stage, we are only concerned with detecting broad patterns at a group level. Groups that are shown to have patterns of unusual index values relative to the general population will be examined in more detail in the second stage confirmatory approach.

In the second stage, a random selection of pairs from each of the flagged groups will be statistically tested using a more stringent alpha level of α = .01. For pairs with index values that fall above the critical values, we reject the null hypothesis that no cheating has occurred between the pair.

The sampled pairs are across all flagged groups, thereby assuming that the cheating pattern is similar although there might be differences in extent and pattern from one group to the other. For example, some groups may have used only a portion of the leaked key while others used the entire key. Unlike Stage 1, which compares individual responses versus

a single response pattern, Stage 2 computes the indices from a sample of pairs within these groups, looking for more detailed evidence of cheating that is specific to the group.

The nominal Type I error rate for cheating detection in both stages is alpha. We should expect to reject the null hypothesis due to random chance alone for the proportion of all pairs equal to or less than alpha. If the empirical rejection rate for the sample in the group exceeds alpha, this can be interpreted as evidence that there is systematic cheating in the group.

Method

Data

The data used for the comparison groups were based on publicly available raw response data from PIRLS 2011 (International Association for the Evaluation of Educational Achievement [IEA], 2012). Due to possible misinterpretation of results, and to emphasise that this paper is not about investigating cheating among the PIRLS participants, a subset has been randomly selected and all identifiers have been removed. This subset (*N*= 64,232) is used as the analysis data throughout this paper. The subset was based on those who were administered a randomly chosen booklet (form 13). The main reason for choosing response data from a single booklet is that answer-copying indices require that the set of items are the same across the compared pairs. This subset is then grouped using IDCNTRY[**] as the arbitrary grouping variable, as IDSCHOOL would have resulted in too many groups that have smaller number of students per group because PIRLS is administered across 13 booklets per school. The main reasons PIRLS data were used are 1) they are publicly available, large-scale and include item-level response data, and 2) the rigorous testing process and multiple levels of security implemented[††] in PIRLS provide credibility to use the data as the comparison (or non-cheating) group[‡‡].

The cheating groups were simulated by forming 18 random groups from the analysis data and simulating correct answers in various patterns (described below). Each group is about 1.2% of the original data, with sample sizes reported in Table 1. About 22% of the total data were therefore simulated to have system-level cheating in varying degrees (from less than 30% to 80% of items copied). The group sizes of the cheating groups are comparable to the remaining original cases in 57 comparison groups.

Table 1: Group sample sizes

Group ID	Frequency	Percent
0001	795	1.2
0002	965	1.5
0003	778	1.2
0004	674	1.0
0005	836	1.3
0006	3708	5.8
0007	650	1.0
0008	618	1.0
0009	736	1.1
0010	739	1.2
0011	731	1.1
0012	728	1.1
0013	698	1.1
0014	760	1.2
0015	623	1.0
0016	599	0.9
0017	618	1.0
0018	823	1.3
0019	749	1.2
0020	922	1.4
0021	719	1.1
0022	665	1.0
0023	668	1.0
0024	522	0.8
0025	762	1.2
0026	581	0.9
0027	1229	1.9
0028	1677	2.6

[**] The identifier for country, but the original PIRLS country codes were anonymised as the Group ID in Table 1.

[††] For more details on the operations and quality assurance of PIRLS, see Martin & Mullis (2012).

[‡‡] At this stage, this is a presumption based only on PIRLS test security and quality control. Later, we will examine if there is psychometric evidence based on our results if this presumption is justified.

0029	621	1.0
0030	900	1.4

Table 1. Continued

Group ID	Frequency	Percent
0031	527	0.8
0032	784	1.2
0033	654	1.0
0034	659	1.0
0035	749	1.2
0036	732	1.1
0037	729	1.1
0038	1070	1.7
0039	939	1.5
0040	719	1.1
0041	545	0.8
0042	1368	2.1
0043	746	1.2
0044	637	1.0
0045	2369	3.7
0046	2000	3.1
0047	622	1.0
0048	573	0.9
0049	607	0.9
0050	1108	1.7
0051	1924	3.0
0052	666	1.0
0053	746	1.2
0054	696	1.1
0055	608	0.9
0056	529	0.8
0057	683	1.1
641	811	1.3
681	743	1.2
742	803	1.3
782	787	1.2
843	786	1.2
883	775	1.2
941	782	1.2
942	806	1.3
943	777	1.2
981	816	1.3
982	799	1.2
983	815	1.3
6121	779	1.2
7122	777	1.2
8123	797	1.2
9121	763	1.2
9122	763	1.2
9123	770	1.2
Total	64232	100.0

Note: The original PIRLS country codes were anonymised. The countries also cannot be deduced from the group sample size because a random number of students were taken from each group to form the cheating groups. The groups simulated to have cheating students are italicised.

Implementation

Form 13 has 35 items in total, with 15 multiple choice items and 20 constructed response items (IEA, 2012). The simulation of cheating was done only for the nominal response data. This means that the test length for the simulation is considerably shorter compared to other studies. For example, Romero and colleagues (2015) used a number of tests with lengths that range from 36 to 54 items while Wollack and Cohen (1998) used 40 and 80-item tests. The nominal response data have 4 categories for each item (multiple choice items with 4 options), while the constructed response were scored ranging from 0-3. Each group were simulated to have different patterns of cheating as shown in Table 2. Because the groups were simulated to be copying from a "supersource" with a perfect response pattern, the copied items are all correct. Responses that were originally coded as "not reached" or "omitted" remained unchanged.

Table 2: Cheating patterns by group

Group ID	Cheating group	Cheating pattern
641	First 4	Copied the first 4 items
742	Middle 4	Copied the middle 4 items
843	Last 4	Copied the last 4 items
941	Random 4 (3 sets)	Copied random blocks totalling 4 items
942		
943		
681	First 8	Copied the first 8 items
782	Middle 8	Copied the middle 8 items
883	Last 8	Copied the last 8 items
981	Random 8 (3 sets)	Copied random blocks totalling 8 items
982		
983		
6121	First 12	Copied the first 12 items
7122	Middle 12	Copied the middle 12 items
8123	Last 12	Copied the last 12 items
9121	Random 12 (3 sets)	Copied random blocks totalling 12 items
9122		
9123		

The additional information to augment the theta estimation came from the constructed response items. Constructed response items are generally more difficult to systematically cheat on. For example, leaking the answer keys would be more efficient in multiple-choice items. For our purposes in this paper, we assume that the constructed responses are honest. There were 3 scenarios (i.e., different sets of $T_{\text{augmented}}$) in terms of the extent of augmentation: 1) all 20 constructed response items were included, 2) only 10 were included, and 3) only 5 were included.

The parameters for the multiple choice and constructed response items were estimated using NRM and generalized partial credit (Muraki, 1992) models respectively. Expected A Posteriori (EAP) estimates were used as person measures (theta). The IRT modelling was implemented using **mirt** (Chalmers, 2012). The item parameters used for computing the ω index were estimated with the cheating groups included. This means that the item parameters are contaminated by aberrant responses, but this reflects real-world scenarios where it is difficult or even impossible to obtain a "clean" calibration sample.

In Stage 1, all students in each group were designated as copiers and paired with the single "supersource". The ω index was then computed for each pair. The statistical test for significance of the computed index values was on the null hypothesis that there is *no cheating* between the pair.

We then computed the proportion of cases with ω values above the critical value at α = .05 within each group. This has the effect of computing the proportion of students that were designated as suspects. Groups which have a proportion of suspected students greater than the nominal Type I error rate (i.e., the proportion of suspects is greater than 5%) were flagged.

Studies on Type I error rates of answer-copying indices have consistently shown that the empirical (or observed) Type I error rate of ω is lower than the nominal Type I error rate (e.g., Romero et al., 2015; Sotaridona & Meijer, 2003; Wollack, 2004). As such, it is expected that the rejection rate for the random pairs in the comparison (non-cheating) groups should not exceed the alpha level (i.e., the "detection rate" should not be more than the nominal Type I error set at alpha), and comparison groups that were flagged were considered as false positives.

In Stage 2, a random sample of 500 pairs for each flagged group were chosen and ω values were computed for each pair. While adequate for our purposes of checking relative proportions, the sample is still a very small proportion of the actual number of pair-combinations for $N = 500$. A balance between making the results more reliable and the amount of computation time needs to be considered when increasing the sample. The pairs were now compared to each other rather than to a single "supersource". We then proceeded with the same statistical test and computation of the proportion of suspected pairs as in Stage 1, except with a more conservative α = .01.

Results

Augmented theta estimation

As discussed previously, there may be situations where the theta estimates are contaminated considerably by widespread cheating. This is reflected in our scenario where approximately 22% of the data were simulated to have varying degrees of systematic cheating. As shown in Figure 1, the distribution of ability estimates is severely skewed when only the "cheatable" items were included (i.e., only $T_{\text{investigated}}$). As more extra items were included in the ability

estimation, the distribution becomes more normal and the ability estimates become closer to what would be expected if there were no widespread cheating.

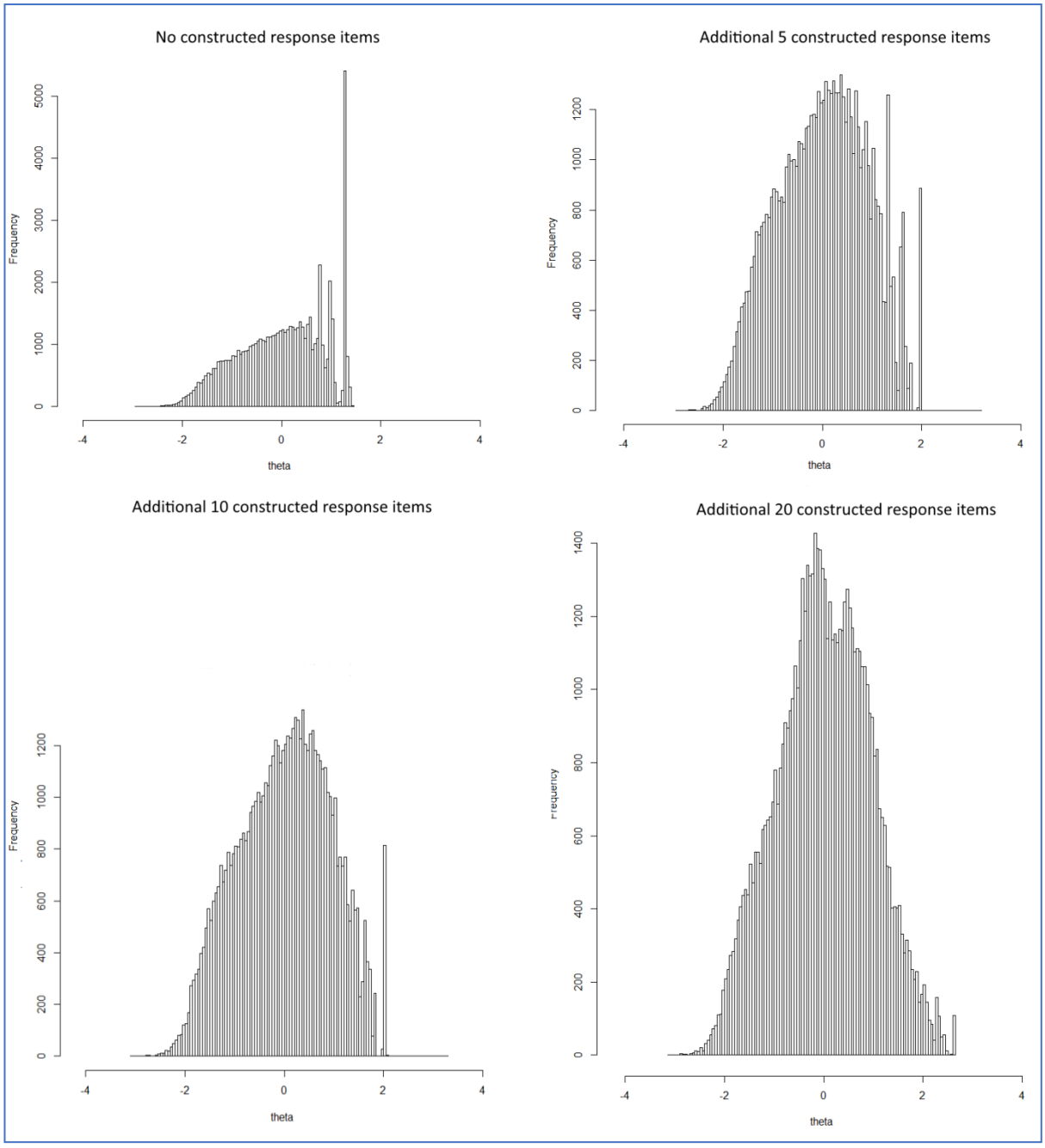

Figure 1: Distribution of ability estimates by set of items used in the theta estimation.

Stage 1

The results for Stage 1 across all three scenarios of theta augmentation are reported in Table 3. As previously discussed, we assume that the comparison groups did not have systematic cheating and therefore those that were flagged can be treated as false positives. Similarly, simulated cheater groups that were not flagged can be treated as false negatives.

Table 3: Summary of results for Stage 1

Group ID	Proportion of suspects		
	Augmented by 20 CR items	Augmented by 10 CR items	Augmented by 5 CR items
0001	3.1%	0.9%	0.4%
0002	2.1%	1.6%	0.9%
0003	3.1%	2.8%	2.3%
0004	0.7%	0.3%	0.3%
0005	3.8%	2.4%	0.5%
0006	2.0%	0.8%	0.5%
0007	1.7%	1.7%	0.3%
0008	1.0%	0.3%	0.0%
0009	3.7%	2.3%	0.8%
0010	2.4%	1.5%	0.7%
0011	4.5%	3.6%	1.8%
0012	**7.0%** [A]	**6.2%** [A]	2.5%
0013	2.9%	2.0%	1.4%
0014	1.3%	1.3%	0.3%
0015	3.0%	3.0%	2.2%
0016	1.3%	0.5%	0.2%
0017	2.9%	1.8%	0.8%
0018	2.4%	1.3%	0.9%
0019	0.3%	0.0%	0.1%
0020	3.3%	1.3%	0.4%
0021	4.3%	2.9%	1.4%
0022	2.4%	1.4%	0.9%
0023	2.5%	2.5%	1.0%
0024	1.9%	1.7%	1.0%
0025	1.6%	1.0%	0.7%
0026	1.4%	0.5%	0.2%
0027	0.5%	0.2%	0.1%
0028	0.4%	0.4%	0.1%
0029	2.7%	3.5%	1.6%
0030	1.1%	0.7%	0.3%
0031	2.3%	1.5%	0.4%
0032	1.8%	1.1%	0.6%
0033	1.5%	0.8%	0.6%
0034	0.3%	0.5%	0.0%
0035	**5.6%** [A]	2.3%	1.1%
0036	4.5%	3.0%	0.5%
0037	0.4%	0.4%	0.0%
0038	2.1%	1.6%	1.3%
0039	2.7%	1.2%	0.9%
0040	4.0%	2.1%	1.5%
0041	1.3%	0.4%	0.4%
0042	3.7%	2.0%	1.5%
0043	2.1%	1.5%	0.3%
0044	0.9%	0.3%	0.2%
0045	0.9%	0.7%	0.3%
0046	2.4%	1.8%	1.2%
0047	1.9%	1.9%	0.8%
0048	3.3%	2.1%	0.5%
0049	1.6%	2.1%	1.3%
0050	1.4%	0.5%	0.3%
0051	1.8%	1.4%	0.7%
0052	0.2%	0.6%	0.2%
0053	2.3%	0.4%	0.3%
0054	2.0%	0.9%	0.3%
0055	2.3%	1.6%	0.7%
0056	3.2%	0.9%	0.6%
0057	4.5%	2.6%	1.6%
641	**6.0%**	1.6% [B]	0.6% [B]
681	**11.6%**	**6.6%**	2.2% [B]
742	3.4% [B]	2.0% [B]	1.0% [B]
782	**9.4%**	**6.1%**	2.3% [B]
843	**6.5%**	3.3% [B]	1.8% [B]

Table 3. Continued

Group ID	Proportion of suspects		
	Augmented by 20 CR items	Augmented by 10 CR items	Augmented by 5 CR items
883	**19.0%**	**9.7%**	3.4% [B]
941	**5.4%**	3.2% [B]	1.3% [B]
942	**7.2%**	3.5% [B]	2.0% [B]
943	3.2% [B]	1.8% [B]	0.9% [B]
981	**14.6%**	**6.9%**	3.4% [B]
982	**18.6%**	**10.6%**	4.3% [B]
983	**20.0%**	**13.1%**	4.8% [B]
6121	**38.3%**	**25.2%**	**5.9%**
7122	**39.0%**	**27.8%**	**8.2%**
8123	**46.0%**	**31.4%**	**9.4%**
9121	**49.1%**	**32.1%**	**9.3%**
9122	**53.6%**	**35.8%**	**11.5%**
9123	**63.4%**	**50.9%**	**16.4%**

Note: Italicised groups were simulated cheaters; proportions that exceed the nominal Type I error rate are bolded.
[A] Stage 1 false positive
[B] Stage 1 false negative

When thetas were estimated using all available constructed response items, reducing the contamination of aberrant response on ability estimation, only two cheating groups with the lowest degree of simulated cheating were not flagged: cheating on the middle 4 items (Group 742) and random cheating on 4 items (Group 943). In addition, two of the comparison (non-cheating) groups were flagged. The number of false positives was reduced as fewer additional constructed response items were included in estimating ability, but the number of false negatives also increased. When only 5 constructed response items were used to augment theta, Stage 1 was only able to detect the groups with 80% copying.

Stage 2

The schools that were flagged in Stage 1 proceeded to Stage 2 analysis and results are reported in Table 4. All cheating groups that were flagged in Stage 1 were also flagged in Stage 2 (i.e., no false negatives). The only scenario where an honest group's empirical Type I error rate exceeds the nominal value was when theta estimation only included 10 constructed response items.

Table 4: Summary of results for Stage 2

Group ID	Proportion of suspects		
	Augmented by 20 CR items	Augmented by 10 CR items	Augmented by 5 CR items
0012	0.2%	**1.2%** [A]	-
0035	0.8%	-	-
641	**2.0%**	-	-
681	**6.0%**	**4.8%**	-
742	-	-	-
782	**2.6%**	**3.4%**	-
843	**3.0%**	-	-
883	**6.4%**	**5.0%**	-
941	**1.8%**	-	-
942	**1.8%**	-	-
943	-	-	-
981	**5.6%**	**1.4%**	-
982	**7.6%**	**7.2%**	-
983	**7.6%**	**5.0%**	-
6121	**15.0%**	**8.6%**	**3.8%**
7122	**12.8%**	**5.0%**	**1.6%**
8123	**14.2%**	**6.4%**	**2.0%**
9121	**17.8%**	**11.4%**	**2.4%**
9122	**19.8%**	**12.0%**	**4.2%**
9123	**28.6%**	**17.8%**	**6.4%**

Note: Italicised groups were simulated cheaters; proportions that exceed the nominal Type I error rate are bolded. Blank cells mean that these groups were not flagged in Stage 1 and therefore were excluded for Stage 2 analysis.
[A] Stage 2 false positive

The distribution of *p*-values for the sampled pairs in Stage 2 is shown in Figure 2. The two false negatives flagged in Stage 1 were no longer flagged in Stage 2, and it shows that ω index *p*-values of the sampled pairs are evenly distributed. It would be expected by chance that no more than 1% of *p*-values would be equal to or less than .01, and indeed groups 0012 and 0035 have 0.2% and 0.8% of sampled pairs with *p*-values < .01, respectively. The skew in the distributions of *p*-values becomes more pronounced as the number of simulated copied items increases.

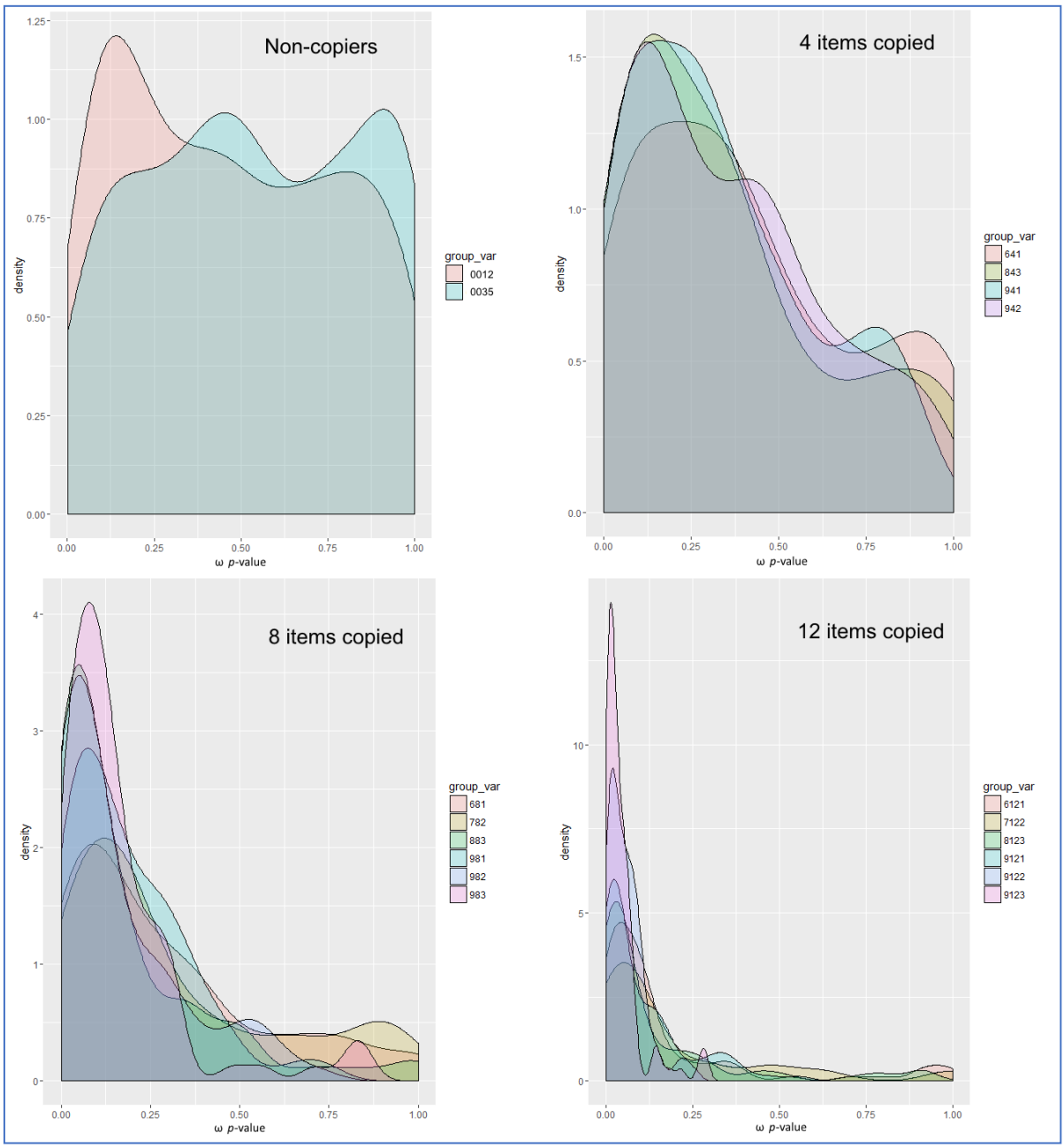

Figure 2: Distribution of p-values among sampled pairs in Stage 2. This shows index values computed using the fully augmented theta estimates only.

For the simulated cheating groups, statistical power of the ω index as applied in this setting corresponds to the proportion of pairs that have ω index values reject the null hypothesis. As shown in Table 4, the power is increasing as the degree of cheating increases. This is not surprising and confirms results from other studies (see Romero et al., 2015; Sotaridona & Meijer, 2002; Sotaridona & Meijer, 2003; Wollack, 1997; Wollack, 2004). The results show comparatively lower levels of power, but this can be expected given that the index was computed for only 15 items and power of the ω index has been found to be affected by test length (Sotaridona & Meijer, 2002; Wollack, 1997).

Precision and specificity analysis

The main purpose of this paper is to demonstrate the viability of this approach and we recommend that other researchers try to replicate this using much richer datasets and preferably with actual (rather than simulated) systematic cheating. To provide some idea as to the group-level precision and specificity of this approach, we

conducted a limited supplemental analysis where we explored a scenario with more restrictive Stage 1 and several incrementally restrictive second stages based on the original and the more restrictive Stage 1. Each scenario is a separate replication, although the corresponding second stages are dependent on the results from their respective first stages. The Stage 2 results for the more restrictive Stage 1 (with $\alpha_1 = .01$) are reported in Table 5.

Table 5: Summary of results from supplementary Stage 2 analyses

Group ID	**Proportion of suspects**			
	$\alpha_1 = .01$			
	$\alpha_2 = .001$	**$\alpha_2 = .0005$**	**$\alpha_2 = .0001$**	**$\alpha_2 = .00005$**
0024	**.002**[A]	<.0005	**.0002**[A]	<.00005
681	**.006**	**0.003**	**.0002**	**.0004**
883	**.012**	**0.007**	**.0016**	**.0012**
981	**.004**	**0.003**	**.0002**	**.0006**
982	**.008**	**0.013**	**.0024**	**.0008**
983	**.014**	**0.004**	**.0016**	**.0006**
6121	**.012**	**0.007**	**.0004**	**.0004**
7122	**.020**	<.0005[B]	**.0002**	<.00005[B]
8123	**.022**	**0.009**	**.0012**	**.0002**
9121	**.030**	**0.016**	**.0034**	**.0034**
9122	**.024**	**0.012**	**.0032**	**.0006**
9123	**.026**	**0.019**	**.0022**	**.0006**

Note: Italicised groups were simulated cheaters; proportions that exceed the nominal Type I error rate are bolded. This supplementary analysis used a more conservative alpha in Stage 1 ($\alpha_1 = .01$), groups that were not flagged in Stage 1 were excluded for Stage 2 analysis.

α_1 = Stage 1 alpha

α_2 = Stage 2 alpha

[A] Stage 2 false positive

[B] Stage 2 false negative

The consolidated results are summarised in Table 6, which shows the number of groups correctly or incorrectly detected. The precision (or positive prediction value) and specificity (or true negative rate) values are computed as follows:

$$PPV = \frac{\sum \text{True positive}}{\sum \text{Predicted positive}}$$

$$TNR = \frac{\sum \text{True negative}}{\sum \text{True negative} + \sum \text{False positive}}$$

Table 6: Group-level precision and specificity based on multiple Stage 1 and Stage 2 scenarios

Number of groups	**$\alpha_1 = .05$**			**$\alpha_1 = .01$**			
	$\alpha_2 = .01$	**$\alpha_2 = .001$**	**$\alpha_2 = .0005$**	**$\alpha_2 = .001$**	**$\alpha_2 = .0005$**	**$\alpha_2 = .0001$**	**$\alpha_2 = .00005$**
Predicted positives	16	15	15	12	10	12	10
Predicted negatives	59	60	60	63	65	63	65
True positives	16	15	15	11	10	11	10
True negatives	57	57	57	56	57	56	57
False positives	0	0	0	1	0	1	0
False negatives	2	3	3	7	8	7	8
Overall total	75	75	75	75	75	75	75
Precision (PPV)	100.00%	100.00%	100.00%	91.67%	100.00%	91.67%	100.00%
Specificity (TNR)	100.00%	100.00%	100.00%	98.25%	100.00%	98.25%	100.00%

α_1 = Stage 1 alpha α_2 = Stage 2 alpha
PPV = positive predictive value TNR = true negative rate

In the context of detecting systematic cheating at group-level, the precision and specificity values show that this approach has a reasonably well-controlled precision at detecting groups, even with a relatively short test. Across all the replications, there has never been more than 1 falsely identified group out of 57 groups that are assumed to be innocent. We focused on and reported specificity rather than sensitivity (or true positive rate) because in this context, false accusations have more dire consequences than undetected cheating, which is discussed in more detail in the next

section. The results show that specificity also appears to be reasonably well-controlled, and relatively stable across the various scenarios.

Discussion

Before discussing the results, we need to keep in mind that cheating detection methods should come second in terms of priority to methods for *deterring* cheating. A quote by Wollack (2004) sums up this principle succinctly:

> *Copying indices are a last resort—they can be used only after the test has been administered and the testing agency suspects that someone's score is spurious. Having available statistical tools to detect suspected copying is important, but exam developers and administrators must continue to proactively address the copying problem by creating a testing environment that will, to as large an extent as large an extent as possible, discourage and prevent copying (p. 44).*

Keeping that overarching principle in mind, the simulation results provide some evidence that the two-stage approach is promising for detecting systematic cheating in a large-scale context. Given that the ω index is largely used for pairwise investigation of cheating, its applications in large scale contexts has considerable logistical limitations.

The exploratory first stage allows for rough but very fast preliminary run through large datasets to flag groups that need to be examined in detail. The primary purpose of Stage 1 is data reduction – to reduce substantially the number of pairs to be examined in the next stage. As such, a more liberal statistical test sufficed for this purpose.

The second stage then allows for a more detailed, but sample-based, statistical testing of random pairs within the suspected groups. Unusually high proportion of flagged pairs within the sampled group can be interpreted as evidence that systematic cheating may have occurred. However, the second stage should not be treated as an end of the process but rather as a stepping stone towards additional and more extensive rounds of investigation where other methods can be brought to bear. These supplementary methods can include school/class-level audit, having a random sample retake the test under close supervision, looking at performance over time using additional data, school-level trend analysis. These methods require more resources but they are feasible on a much smaller and more manageable number of cases (i.e., the suspected group).

In addition, the augmentation process in estimating the thetas improve and mitigate the psychometric limitations (Sunbul & Yormaz, 2018; Zopluoglu & Davenport, 2012) inherent in contaminated data (because of the presence of cheaters). This augmentation process is especially useful in standardized test settings where other item formats are available. Our results have shown that even a small number (5 to 10) "uncontaminated" items can drastically improve the ability estimates, thereby also improving the performance of the detection process. What emerged from the insights on psychometric and logistic challenges is an approach that is both efficient and effective as confirmed by the results of the supplemental precision and specificity analysis.

Limitations and Extensions to Future Research

This study is limited by relying on simulated data on cheaters. Applying the method to actual data is recommended for further validation. In this particular simulation, the ω index appears to have lower statistical power compared to other simulations, but this could be due to *considerably shorter* test length. Only 16 out of the 18 simulated cheater groups have been detected in Stage 1 and further confirmed in Stage 2. Nevertheless, the empirical Type I error rates remain well-controlled even if the item parameters were contaminated by a substantial proportion of cheaters.

Incidentally, our initial presumption that *the comparison groups are non-cheaters* has been supported by the evidence from our results – by Stage 2, the proportion of pairs which we reject the null hypothesis no longer exceeds our nominal alpha level when using fully augmented theta.

Because of the serious consequence in even the *allegation* of cheating, we recommend against a purely statistical set of evidence to accuse schools of systematic cheating. Where integrity is at stake and stigma can be very difficult to remove, a false negative is far more preferable than a false positive. It is also worth noting that statistical analyses only provide evidence and never conclusive determination. As such, we recommend that these statistical procedures be used only for flagging groups. Even for flagging purposes, as a first step to a more comprehensive investigation, we recommend using corroborating evidence from multiple statistical evidence to reduce false positives as much as possible.

It would be useful to apply this method on other large-scale data with different test lengths. This would show a clearer picture of achievable statistical power using this two-stage sample-based approach. It would be ideal if actual data with suspected systematic cheating can be obtained. Where systematic cheating was suspected to have occurred and other sources of evidence are available, statistical methods such as answer-copying indices would have a more robust benchmark for comparison.

Exploring ways to further improve the estimation of both item and person parameters would also be useful. If a clean calibration subset of the data is available or can be identified, the item parameters can be estimated using only that

subset and thereby minimising the contamination due to aberrant responses. This is not always possible. Moreover, even with uncontaminated item parameters, theta estimation will remain contaminated.

We have shown here that including additional items, especially those that are less susceptible to systematic cheating, can make the cheating detection process more effective. This augmentation can be improved further by utilising information from other domains (e.g., through multidimensional IRT models that incorporate items from other tests that were also taken by the student). Even if no additional test items are available, additional information outside of test responses can be utilised using imputation methods (Rubin, 1987) or conditioning models (Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992) that take into account background variables to increase the reliability of the person estimates.

Finally, digital testing environments allow for additional data to be captured to further augment cheating detection methods. For example, applications of statistical approaches can be combined with data analytics to utilise behavioural indicators such as response times. These techniques can broaden the repertoire of cheating detection methods as well as strengthen the current approaches.

References

Benbow, J., Mizrachi, A., Oliver, D., & Said-Moshiro, L. (2007). *Large class sizes in the developing world: What do we know and what can we do*. Washington, DC: American Institute for Research.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29.51.

Bock, R.D., & Aitkin, M. (1981). Marginal Maximum Likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

British Broadcasting Corporation. (2015). *India students caught 'cheating' in exams in Bihar*. Retrieved from http://www.bbc.com/news/world-asia-india-31960557.

Chalmers R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chajewski, M., Kim, Y., Antal, J. & Sweeney, K. (2014). Macro level systems of statistical evidence indicative of cheating. In Kingston, N. M., & Clark, A. K. (Eds.). *Test fraud: Statistical detection and methodology*. New York, NY: Routledge.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B, 39*(1), 1-38.

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (ETS Tech. Rep. No. 96–4). Princeton, NJ: Educational Testing Service.

International Association for the Evaluation of Educational Achievement. (2012). *PIRLS 2011*. Boston, MA: TIMSS & PIRLS International Study Center.

Levine, M. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical Statistical Psychology, 35*, 42-56.

Levine, M. & Rubin, D. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R.J., Johnson, E.G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131–154.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org/

Romero, M., Riascos, A., & Jara, D. (2015). On the Optimality of Answer-Copying Indices: Theory and Practice. *Journal of Educational and Behavioral Statistics*, *40*(5), 435–453.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Sunbul, O., & Yormaz, S. (2018). Effects of Test Level Discrimination and Difficulty on Answer-Copying Indices. *International Journal of Evaluation and Research in Education, 7*(1), 32-38.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement, 39*(2), 115–132.

Sotaridona, L.S., & Meijer, R.R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*(1), 53-69.

van der Linden, W.J., & Sotaridona, L.S.(2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*(3), 283-304.

Wollack, J.A.(1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307–320.

Wollack, J.A.(2004). Detecting answer copying on high-stakes tests. *The Bar Examiner, 73*(2), 35-45.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*(2), 144-152.

Zopluoglu, C., & Davenport Jr, E. C. (2012). The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement, 72*(6), 975-1000.

Zopluoglu, C. (2013). CopyDetect: An R package for computing statistical indices to detect answer copying on multiple-choice examinations. *Applied Psychological Measurement, 37*(1), 93-95.