KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

**Received:** December 9, 2014
**Revision received:** March 17, 2016
**Accepted:** March 28, 2016
**OnlineFirst:** November 15, 2016

www.estp.com.tr
**DOI** 10.12738/estp.2017.1.0105 ♦ February 2017 ♦ 17(1) ♦ 5–26

***Research Article***

# Determination of Type I Error Rates and Power of Answer Copying Indices under Various Conditions[*]

Seha Yormaz[1]
*Mersin University*

Önder Sünbül[2]
*Mersin University*

## Abstract

This study aims to determine the Type I error rates and power of $S_1$, $S_2$ indices and kappa statistic at detecting copying on multiple-choice tests under various conditions. It also aims to determine how copying groups are created in order to calculate how kappa statistics affect Type I error rates and power. In this study, option-matrices were obtained by using five choices of multiple choice test data generation. R-3.0 software was used for data generation and analysis. The effects of sample size, test length, and source ability level on Type I error rates and also amount of copying on power according to nominal α-levels were investigated. The findings indicated that Type I error rates for $S_1$ and $S_2$ indices were lower in all conditions, while the same error rates were relatively higher for copy detection using kappa statistics. Kappa statistics were seen to be more powerful than $S_1$ and $S_2$ indices under all conditions. As sample size, test length, and amount of copying increased, the $S_2$ index was also revealed to be more sensitive than the $S_1$ index. Type I error rates and power results of the kappa statistics were similar to each other.

**Keywords**

Copy detection methods • $S_1$ and $S_2$ indices • Kappa statistic

[*] This study is derived from the master's thesis conducted by Seha Yormaz at the Institute of Educational Sciences of Mersin University under the supervision of Assistant Professor Önder Sünbül.

[1] **Correspondence to:** Seha Yormaz, Department of Measurement and Evaluation in Education, Faculty of Education, Mersin University, Mersin Turkey. Email: sehayormaz@gmail.com

[2] Department of Measurement and Evaluation in Education, Faculty of Education, Mersin University, Mersin Turkey. Email: ondersunbul@gmail.com

**Citation:** Yormaz, S., & Sünbül, Ö. (2017). Determination of type I error rates and power of answer copying indices under various conditions. *Educational Sciences: Theory & Practice*, *17*, 5–26. http://dx.doi.org/10.12738/estp.2017.1.0105

Cheating, a factor which threatens tests' psychometrics properties, is a serious problem within the processes of measurement and assessment. Examinees' test scores do not reflect their true scores in the presence of cheating. So, this situation results in giving wrong decisions about the cheaters in terms of the features to be tested. From another point of view, if cheating systematically increases the number of correct answers, this gets evaluated as a systematic error, and systematic errors jeopardize test validity. If cheating does not systematically increase the number of correct answers, it can be evaluated as a random error, and this type of error threatens reliability. As a result, cheating damages the psychometric properties of tests. Nowadays, research on cheating and copy detection methods have increased in order to deal with the problem of cheating. Cheating methods were categorized by Cizek (1999) into three main headings: giving/taking/receiving information, using forbidden materials, and circumventing or taking advantage of the testing process. Researches show that more than 50% of students have admitted to cheating on their exams, and the rate of cheating has increased rapidly (Cizek, 1999, p. 35). For this reason, researchers have developed various copy detection methods to detect examinees' cheating levels. One of these methods depends on statistical evidence.

Saupe (1960) placed methods that depend on statistical evidence into two categories: empirical methods and chance methods. In empirical methods, the answer distribution of examinees suspected of cheating is compared with that of a group known not to have cheated. In methods that depend on chance probability, the answer distribution of examinees suspected of cheating is compared with a theoretical distribution, such as binomial, Poisson, or standard normal distribution. Over time, methods of detection have moved away from empirical approaches and toward chance methods (Cizek, 1999, p. 139). Numerous indices since 1927 have been developed for detecting cheating. Most of those indices are based upon classical test theory, while some others on item response theory. Most of these methods focus on the probable similarity of answers (chance probability) that independent examinees give without cheating or collaborating. Bird (1927; 1929), Crawford (1930), Anikeef (1954), and Saupe (1960) are pioneers of these indices.

However, because calculating these indices is complicated, satisfactory results of the indices' sensitivity for detecting copying cannot be precisely obtained (Argenal, Co, Cruz, & Patungan, 2004). Angoff (1974) developed eight indices (*A* through *H*) that use the same procedure as the index developed by Saupe (1960) by using different parameters for identifying cheaters. Angoff (1974) showed parameters that conclude that the number of correct and wrong answers (B and H indices) were more successful at detecting cheating. A number of indices were developed after Angoff's, and numerous statistical methods have been used for detecting cheating. When reviewing the related literature, the following indices and methods for detecting cheating have been discovered: $g_1$ and $g_2$ indices (Frary, Tideman, & Watts, 1977, 1993), PAIR1 and PAIR2 indices (Hanson, Harris, & Brennan, 1987), ESA index (Bellezza & Bellezza, 1989), $B_m$ index (Bay, 1994), Harpp-Hogan index (Harpp, Hogan, & Jennings, 1996), $K$ index (Holland, 1996), $\omega$ index (Wollack, 1997), $\bar{K}_1$ and $\bar{K}_2$ indices (Sotaridona & Meijer, 2002), $S_1$ and $S_2$ indices (Sotaridona & Meijer, 2003), Differential Item Functioning (DIF; Giardano, 2005), kappa statistic (Sotaridona, van der Linden, & Meijer, 2006), response time (RT) models (van der Linden, 2006), generalized binomial test (GBT; van der Linden & Sotaridona, 2006), $\tau$ statistics (van der Ark, Emons, & Sijtsma, 2008), CTAD (Deng, 2008), cumulative sum statistic (CUSUM$_{LR}$; Armstrong & Shi, 2009), factor analysis (Clark, 2010), Kullback-Leibler (K-L) divergence and $K$-index (Belov & Armstrong, 2010), Final Log Odds Ratio (FLOR; Hui, 2010), Variable Match Index (VM-index; Belov, 2011), and Deterministic, Gated Item Response Theory Model (DGIRTM; Shu, 2011).

This study is limited to $S_1$ and $S_2$ indices and kappa statistics. In order to understand the $S_1$ and $S_2$ indices, one must first understand the $K$-index.

## K Index

The K index, is used to determine irregular matching of wrong answers for two examinees who took the same multiple-choice exam. The K index is an estimator of the chance probability of two examinees having wrong answers that match.

This index, first developed by Frederic Kling in 1979, was reintroduced by Holland (1996) under the following assumptions:

1. Matching wrong answers from examinees who took the same exam is an indicator of copying.
2. If two examinees answer an item wrong and select the same option, then there is strong evidence for the violation of independence regarding the answers obtained from a single item.
3. If two examinees answered an item correctly, there is weak evidence for cheating because both examinees could know the correct answer.
4. Omitted answers are not seen as evidence of copying. For example, examinees are encouraged to not mark items for which they are not sure of the answer. Thus the rate of leaving difficult items unmarked increases. Another reason for leaving items unanswered is that students were unable to use the exam time appropriately.

The K index, which is produced under these assumptions, is easy to calculate. However, its limitations are that only information related to examinees' matching wrong answers is used and the reliability is low for small samples. K index with lower values at various significance levels as a result of calculations is considered evidence that the cheater had copied from a source (Holland, 1996). In order to clear up some sets of problems observed with the K index, Sotaridona and Meijer (2002) put forward the $\bar{K}_1$ and $\bar{K}_2$ indices, where whose estimates are obtained from linear and quadratic regressions.

### $\bar{K}_1$ and $\bar{K}_2$ Indices

The difference between these indices, which were developed by Sotaridona and Meijer (2002), and the K-index is the parameter estimation of the *p*-binominal probability. While the data set of the copying subgroup is only used for estimating the *p* for the K index, the data set of all examinees with wrong responses matching the source's wrong answers is used for $\bar{K}_1$ and $\bar{K}_2$ indices. In this way, Sotaridona and Meijer (2002) claimed that $\bar{K}_1$ and $\bar{K}_2$ indices are more powerful indices, and they presented in their study that these indices are superior to the K index for detecting answer copying. Smaller values of $\bar{K}_1$ and $\bar{K}_2$ indices from several nominal alpha levels are evaluated as evidence of copying.

### $S_1$ Index

The $S_1$ index, which was developed by Sotaridona and Meijer (2003), is similar to the $\bar{K}_1$ index in that it is constructed from the *M* (distribution of matched wrong answers) variable. However, binomial distribution is used for $\bar{K}_2$ whereas Poisson distribution is used for the $S_1$ index.

### $S_2$ Index

*K*, $\bar{K}_1$ and $\bar{K}_2$ indices are built on matched incorrect answers, and the $S_1$ index does not use information regarding correct answers that match the source. Sotaridona and Meijer (2003) considered this as a limitation and thus developed the $S_2$ index. According to Sotaridona and Meijer (2003), when matching correct answers are ignored, copiers' correct answers are accepted as if they had marked the correct answer on their own, yet there is the possibility that the copier might have marked the correct answer by cheating or guessing. The $S_2$ index is an extension of $S_1$ and was developed for these limitations. $S_2$ differs from $S_1$ in that $S_2$ accounts in its calculations for the information obtained from examinees' same correct answers. Lower values of the $S_2$ index from several nominal alpha levels are evaluated as evidence of cheating. Like the $S_1$ index, the model's adequateness is assessed using the $g^2$ index.

## Kappa Statistic

Sotaridona, van der Linden, and Meijer (2006) suggested using Cohen's kappa statistic to detect cheating for multiple-choice answers. In contrast with other statistics, there is no assumption that investigates the answering process of the copier or the source. The only assumption is the existence of several probabilities for answers generated by examinees (Sotaridona et al., 2006). Sotaridona et al. aimed to determine the Type I error rate of the probability of the copier and source obtaining the same answer according to their ability levels. Their study found that, except from the situations where the ability levels were in opposite direction (low ability level copier - high ability level source or high ability level copier – low ability level source), there was an increase in the Type I error rate. They suggested recoding the answers to decrease the Type I error rate. A decrease in Type I error rates was observed at the end of the study.

Steps for recoding answers as followed in the study are given below:

1. Calculate *a* values, which are defined as the ratio of test takers to preferential options.
2. Construct the coding template (schema).
3. Recode answers according to constructed template (Sotaridona et al., 2006, p. 421).

In calculating the *a* values, Sotaridona et al. (2006) suggested that just as the subgroup with the same ability level as the copier can be taken, so can examinees with the same number of correct responses as the copier (or when the ability level is hard to estimate, the percentage of correct responses from the copier's group as formed according to examinees' correct responses) also be taken.

## Type I Error Rate and Power of Indices to Detect Answer Copying

*Type I error rate* and *power to detect answer copying* are two important concepts needed to decide how to apply indices in terms of usefulness and trustability. What is expected from the indices is for them to identify copiers correctly in cases of cheating. For this reason, power to detect cheating means identifying copiers correctly, and Type I error means deciding an examinee is a copier when they are not. High index power and low Type I error rates are desired for indices that detect copying.

Table 1
*Type I Error and Copy Detection Power*

| | | Decision Made for Hypothesis Test | |
|---|---|---|---|
| | | $H_0$ denied<br>"Copy" | $H_0$ accepted<br>"No Copy" |
| Reality | $H_0$ true "No Copy" | Type I Error | True Decision |
| | $H_0$ wrong "Copy" | (Copy Detection Power)<br>True Decision | Type II Error |

A limited number of studies in the literature have focused on Type I error rates and copy detection power for $S_1$, $S_2$, and kappa statistics under various conditions. To clarify the powerful parts of the related indices, a literature review has been given for previously developed indices:

Hanson, Harris, and Brennan (1987) compared $g_2$, $B$, $H$, $P$, $CP$, Pair I, Pair II indices for types of copying and showed that indices generated very similar detection rates. Additionally, indices for "random cheating" are more powerful than indices developed for "serial cheating" when the percentage of cheating is higher.

Bay (1995) compared $B_{m,}$ $g_2$, and ESA indices in her study. The study showed the three indices were insufficient at detecting answer copying when copying was identified in 25% or less of the total number of items.

Frary and Tideman (1997) compared $B$ and $g_2$ indices. The study showed that the $B$ index is more efficient at identifying high-scoring copiers and $g_2$ is more efficient at identifying low-scoring copiers. The researchers also highlighted that when the $g_{2w}$ index (derived from the $g_2$ index) was used with $g_{2,}$ greater copy detection rates were obtained.

Wollack (1997) compared $g_2$ and $\omega$ indices for various types of copying in terms of sample size, test length, and amount of copying. The study underlined that the $\omega$ index is more powerful than the $g_2$ index at copy detection under all simulated conditions, and the $g_2$ index is insufficient at controlling Type I error rates.

Sotaridona and Meijer (2002) compared K, $\bar{K}_1$, $\bar{K}_2$, and $\omega$ indices in terms of sample size, test length, and amount of copying. The study showed that since all indices were capable of controlling Type I errors, they were better than the $g_2$ index (Wollack, 1997), which is weak at controlling Type I errors. In addition, the researchers showed that the $\omega$ index is relatively more powerful than other indices; however, for situations where $\omega$ is not applicable, $\bar{K}_2$ is more powerful than K and $\bar{K}_1$.

Sotaridona and Meijer (2003) compared their $S_1$ and $S_{2,}$ indices with $K$, $\bar{K}_2$, and $\omega$ in terms of sample size, test length, and amount of copying. The study showed that $S_1$ is better than $\bar{K}_2$, and that $S_2$ and $\omega$ are better than the others in terms of power. Another result of the study was that for item parameters estimated accurately from nominal response model (NRM) of item response theory, the $\omega$ index generates more sensitive results for copier ability levels and is applicable for small sample sizes. However, researchers noted that $S_1$ and $S_2$ have less power for detecting answer copying in small sample sizes.

Wollack (2003) compared *Scrutiny!*, $K$, $g_2$, and $\omega$ indices, showing that the $\omega$ index has the greatest power under all conditions.

Wollack (2006) later proposed the simultaneous use of $\bar{K}_2$, $S_1$, $S_2$, $\omega$, $H^*$, and $B$ indices. Comparison for single use of indices showed that $\bar{K}_1$, $S_1$, $S_2$, $\omega$, and $B$ indices had low Type I error rates for all nominal α-levels (0.01, 0.005, 0.001, and 0.0005), while the $H^*$ index had high rates. In addition, $\omega$ had the highest power among the compared indices; $S_2$ was the second most powerful index. The study showed that simultaneous use of $\omega$ and $H^*$ was more powerful than other combinations. The researcher recommended using $S_2$ for cases where $\omega$, which depends on item response theory, cannot be used.

Sotaridona et al. (2006) showed the kappa statistic, which they suggested for copy detection, had powerful results for five option multiple-choice tests of 30 or 60 items. However, the recoding study for the statistic's sensitivity toward source and copiers' ability levels showed a remarkable increase for Type I errors at the 0.05 nominal alpha level when the ability levels of copier and source were similar.

The literature review of prior studies shows that, apart from $\omega$ (which depends on item response theory), the most powerful indices for detecting answer copying are $S_1$ and $S_2$. No study was found to have compared $S_1$ and $S_2$ with the kappa statistic for Type I error rates or power. For this reason, this study's purpose is to compare $S_1$ and $S_2$ with the kappa statistic under various conditions.

Table 2

*Literature Review for Studies Comparing Related Indices*

| *Research* | *Indices* | *Sample Size* | *Test Length* | *Amount of Copying* | *α-level* | *Data Type* |
|---|---|---|---|---|---|---|
| Sotaridona & Meijer (2002) | $K$, $\bar{K}_1$, $\bar{K}_2$, $\omega$ | 100, 500, 2000 | 40, 80 | 10%, 20%, 30%, 40% | 0.0001<br>0.0005<br>0.001<br>0.005<br>0.01 | Simulation |
| Sotaridona & Meijer (2003) | $\bar{K}_1$, $\omega$, $S_1$, $S_2$ | 100, 500 | 40, 80 | 10%, 20%, 30%, 40% | 0.0001<br>0.0005<br>0.001<br>0.005<br>0.01 | Simulation |
| Wollack (2006) | $\bar{K}_2$, $S_1$, $S_2$, ω, H, B | 25, 50, 100, 250, 500, 1000, 2000, 5000, 10000 | 40, 80 | 10%, 20%, 30%, 40% | 0.0005<br>0.001<br>0.005<br>0.01 | Simulation |

## Purpose of the Study

Educational and psychological testing for decision making is frequently performed for individuals. The appropriateness of the decisions made for individuals depends on reliability and validity of tests. However, many factors decrease the reliability and validity of these tests. Copying is an individual attempt at pretending to have knowledge by cheating (Cizek, 1999). Because of this decision making as "master" for a copier who is not capable for the intended trait to measure will be a fault. As is the case with copying during a testing process, being able to determine answer copying is very important for a test's validity.

A literature review shows that numerous methods have paved the way for detecting answer copying; $S_1$ and $S_2$ have the greatest power of copy detection, apart from the $\omega$ index, which depends on item response theory. At the same time, no study could be found that compares $S_1$ and $S_2$ with the kappa statistic, which was purposed by Sotaridona et al. (2006) for detecting answer copying in terms of power and Type I error rates. This study aims to compare the kappa statistic with $S_1$ and $S_2$ indices, which are used for detecting answer copying under various conditions, to determine which one has lower Type I error rates and higher detection rates. Another purpose of this study is to explain how the copier subgroup used for the kappa statistic calculation affects the Type I error rate and power. Additionally, no study could be reached for copy detection indices in Turkey. This study also supposes to contribute to measurement and evaluation studies in Turkey by being the first study on copy detection methods here.

### Research Question

What are the Type I error rates and power of $S_1$ and $S_2$ indices and kappa statistics as used for detecting answer copying under various conditions?

**Sub-questions**. Answers to the questions given below were attempted:

1. What are the Type I error rates of the indices under various conditions?
    1. What are the interaction effects of conditions for the specified nominal alpha levels?
2. What are the copy detection powers of indices for various conditions?
    1. What are the interaction effects of conditions for the specified nominal alpha levels?

This study is limited to data sets that were generated by a computer program, the indices included in the study, predetermined conditions (sample size, test length, source ability level, and amount of copying), and their levels.

# Methodology

## Type of the Study

This study aims to provide information about Type I error rates and power for $S_1$ and $S_2$ indices and kappa statistics by using simulated datasets under various conditions. From this framework, the study can be evaluated as a fundamental research.

## Data Generation and Data Analysis Conditions

The data sets of the study were generated using the R-3.0 program for four conditions (sample size, test length, amount of copying, and source ability level over the copier) with over 100 replications.

In this study, the mcIRT package (Reif, 2013) of the R-3.0 program was used to generate data according to NRM. For data generation, examinees ability parameters were obtained from a standard normal distribution with a mean of 0.00 and a standard deviation of 1.00. This was done separately for three sample sizes: 100, 500, and 2,000.

By using the R-3.0 program, multiple-choice test items with five options and test lengths of 40 and 80 items were generated for chosen sources with five different ability levels (40-49%, 50-59%, 60-69%, 70-79%, and 80-90%). In the study, 30 different condition levels were created for Type I error rates (3 [sample size] x 2 [test length] x 5 [source ability level]). For detecting answer copying, copying ratios of 10%, 20%, 30%, and 40% of the total test items were used in addition to the previous conditions. To investigate power, 120 different condition levels were generated (3 [sample size] x 2 [test length] x 5 [source ability level] x 4 [amount of copying]). For each condition level, 100 replications were conducted for data generation. As a result, 3,000 datasets were generated for studying Type I errors (30 x 100) and 12,000 datasets were generated for copy detection power (120 x 100).

The sample size used in the study were chosen to compare test length, amount of copying conditions, and ability levels by using 100 replications with the studies of Sotaridona and Meijer (2002; 2003). To determine the upper and lower bounds of the source ability levels' percentages, the limits of 40% and 90% from Sotaridona and Meijer's (2003) study were taken into consideration. Manipulated conditions and their levels are given in Table 3.

Table 3
*Manipulated Conditions and Levels*

| *Condition* | *Number of Levels* | *Level Values* |
|---|---|---|
| Sample Size (*N*) | 3 | 100 (*N1*)<br>500 (*N2*)<br>2000 (*N3*) |
| Test Length (*T*) | 2 | 40 (*T1*)<br>80 (*T2*) |
| Source Ability Level (*Y*) | 5 | 40 - 49% (*Y1*)<br>50 - 59% (*Y2*)<br>60 - 69% (*Y3*)<br>70 - 79% (*Y4*)<br>80 - 90% (*Y5*) |
| Amount of Copying (*C*) | 4 | 10% (*C1*)<br>20% (*C2*)<br>30% (*C3*)<br>40% (*C4*) |
| Indices | 3 | $S_1$, $S_2$ and Kappa |
| Number of Replications | 100 | |

## Process

Studies given below were conducted respectively with the program written in R by the researchers:

**Study 1: Determining Type I error rates of the indices.** No cheating data was generated according to sample size (100, 500, and 2,000) and test length (40, 80). In the study, sampled individuals were assigned to 4 row 20 individual class designs. In order to choose source and copier, ability-percentage intervals were first determined and five groups were created for choosing the source according to the ability levels mentioned in Table 3. For each replication, a source was chosen from the whole sample according to ability-level restrictions, and then without replacement, a copier with a lower ability level than the source was chosen at random from the nearby source.

After determining the copier and source, 3,000 (3 x 2 x 5 x 100) data sets were generated with 100 replications for the conditions of sample size, test length, and source ability level.

**Study 2: Determining power of the indices.** For Study 2, datasets were first generated according to sample size (100, 500, and 2,000) and test length (40 and 80). For each replication, source and copier were selected by the same procedure as in Study 1. From the responses of the chosen source, 10%, 20%, 30% and 40% of the test length (for 40 items: 4, 8, 12, and 16 responses; and for 80 items: 8, 16, 24, and 32 responses respectively) were chosen randomly, and the copier's responses were changed in order to match the source's chosen responses; thus, the cheating condition was created. For every level of sample size, 12,000 (3 x 2 x 5 x 4 x 100) data units were generated through 100 replications for the conditions of test length, source ability level, and amount of copying.

To investigate the effect of group, which is required for calculating the kappa statistic in Studies 1 and 2 in order to detect answer copying, three different groups were constructed. For this purpose, the procedures below were conducted by determining the copier's number of correct responses:

1. A group with the same number of correct responses as copier was constructed for the first group.
2. For the second group, five groups were constructed according to the examinees' number of correct responses, and the group that included the copier was determined.
3. For the third group, 10 groups were constructed according to examinees' number of correct responses, and the group that included the copier was determined.

In Studies 1 and 2, the $S_1$ and $S_2$ indices and Kappa statistic for each copier subgroup were calculated for each data set. The calculated kappa statistics were named *Kappa1* for the first group, *Kappa2* for the second group, and *Kappa3* for the third group.

The values after the calculations were compared with the determined nominal alpha levels (0.0005, 0.001, 0.005, 0.01, and 0.05) in order to identify the Type I error rates and powers. Nominal alpha levels were determined according to the literature review.

## Data Analysis

In the study, calculations were conducted using a program which was written in R-3.0 programming language for the index equations. The obtained results were analyzed by comparing them with the determined nominal alpha levels.

**Determining the Type I error rates of the indices.** To determine the Type I error rates of the indices, the probability values for the selected pair of copier and source from each replication for the non-cheating data were calculated.

The calculated probabilities were compared with nominal alpha levels (0.0005, 0.001, 0.005, 0.01, and 0.05). If the probability value was equal to or less than the determined nominal alpha level, then a "1" was assigned to that pair (who had not cheated but had been mistakenly marked as "cheating"), otherwise a "0" was assigned.

**Determining the copy detection power of the indices.** As mentioned in Study 2, different amounts of copying were created, and probability values were calculated for the selected pair of source and copier for each replication.

As with determining Type I errors, "1" was assigned for probability values equal to or less than each specified nominal alpha level (0.0005, 0.001, 0.005, 0.01, and 0.05), and "0" for others; binary matrices (0 - 1) were created. For each matrix, averages were obtained by calculating the ratio of the sum of "1" values to the total selected pairs, thus attempting to present the probability of a true decision of "cheating" for selected pairs (copy detection power).

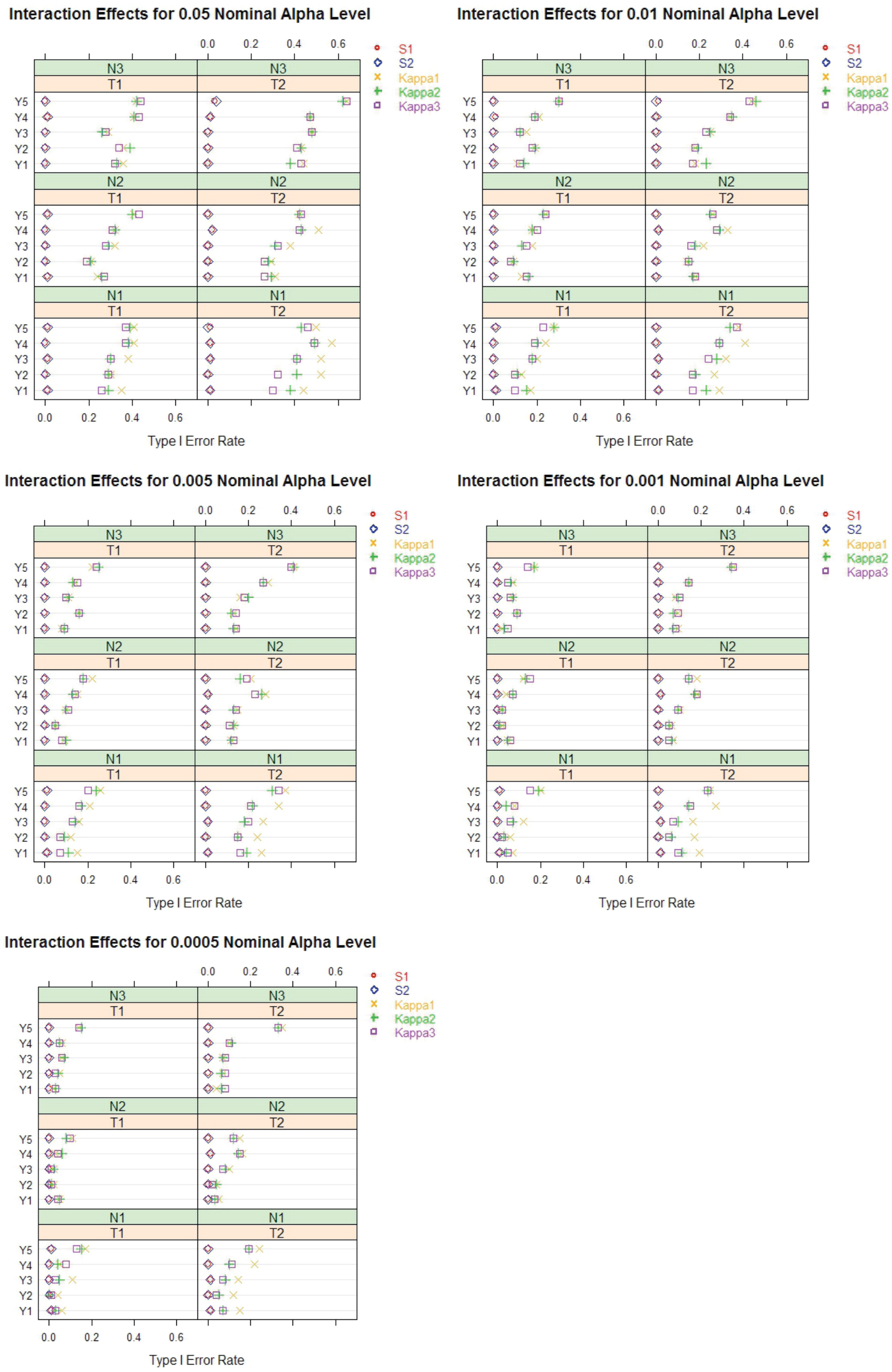
## Findings

In this section, results are given for Type I error rates and powers of $S_1$ and $S_2$ with different group sizes used for calculating the kappa statistic to determine cheating. Multivariate graphs were used to present the Type I error rates and powers of indices under various conditions.

### Type I Error Rates of Indices

The interaction effects of sample size, test length, and source ability level for Type I error rates of copy detection were given using graphs at the nominal alpha levels of 0.05, 0.01, 0.005, 0.001, and 0.0005.

**Interaction effects of conditions on Type I error rate.** Multivariate graphs were used to investigate the interaction effects of sample size, test length, and source ability level for Type I error rates.



Interaction Effects for 0.05 Nominal Alpha Level

Interaction Effects for 0.01 Nominal Alpha Level

Interaction Effects for 0.005 Nominal Alpha Level

Interaction Effects for 0.001 Nominal Alpha Level

Interaction Effects for 0.0005 Nominal Alpha Level

*Graph 1*. Type I error rates of conditions for specific nominal alpha levels.

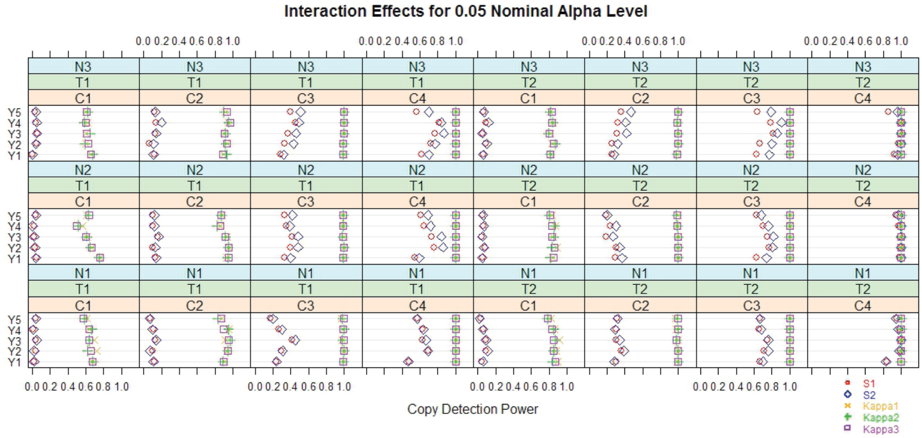
As can be seen in Graph 1, Type I error rates of $S_1$ and $S_2$ indices for all conditions and condition levels can be said to be lower than the specified nominal alpha levels. Additionally, $S_1$ and $S_2$ indices for sample sizes of 100 showed some errors for test length of 40 items and at the upper intervals of the source ability levels. Also, some minor errors were observed for test length of 80 items and source ability levels at the 40-49% and 60-69% ranges.

Type I error rates of kappa statistics were also observed to be high for all conditions. When investigating the graphs for the interaction effects of conditions, the Type I error rate of kappa statistics for detecting answer copying increased when test length and source ability level increased. For cases where the sample size was 500, the error rate decreased with respect to the other sample sizes. When investigating the error rates among kappa statistics, Kappa1 statistic was observed to have higher error rates than other kappa statistics, especially for the sample sizes of 100; for other sample sizes, Kappa1 had values closer to Kappa2 and Kappa3 statistics. For sample sizes of 2,000, test length of 80 items, and source ability level between 80-90%, Type I error rates of kappa statistics were seen to be quite high. The Type I error rate of each kappa statistic had the smallest value for sample sizes of 500, test length of 40 items, and source ability level between 50-59%. Distinct from the others, at the 0.05 nominal alpha level for conditions with a sample size of 2,000, test length of 80 items, and source ability level between 80-90%, a small increase was observed in the Type I error rates of $S_1$ and $S_2$ indices. When investigating the graphs for 0.05 and 0.01 nominal alpha levels, the error rate of the Kappa2 was observed to be slightly higher than that of Kappa3 for sample sizes of 100.

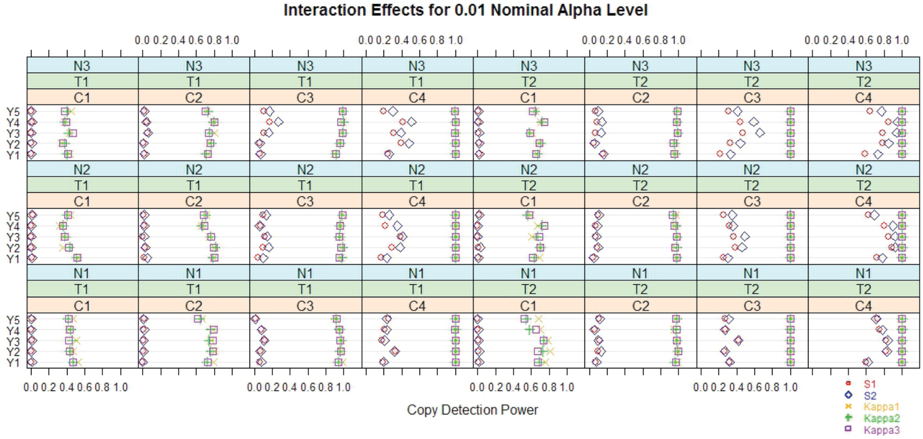
As the source ability level increased, the difference between the two statistics decreased, especially for the 0.05 nominal alpha level. For the 0.05 and 0.01 nominal alpha level graphs, divergence of the rate of copy detection Type I errors of Kappa2 and Kappa3 statistics was observed to decrease at the 0.005 nominal alpha level. When investigating the graphs of 0.001 and 0.0005 nominal alpha levels, distinct from the others, one can see cases where Type I error rates were closer to 0 for all kappa statistics regarding sample sizes of 100 and 500, test length of 40 items, and source ability level between 40-59%. However, neither of these values was less than the specified 0.001 and 0.0005 nominal alpha levels. Because of this, it would be wrong to claim that kappa statistics had fewer Type I error rates for these specific condition levels.

## Copy Detection Power of Indices

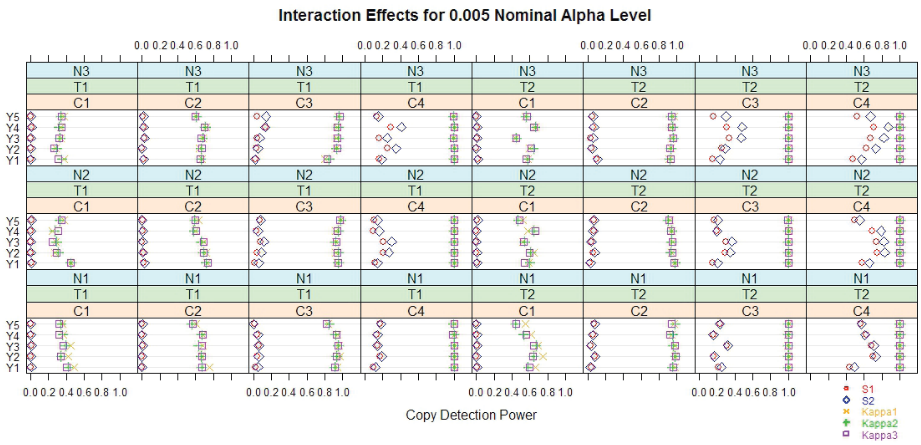
**Interaction effects of conditions on power.** Interaction effects of conditions on power for the specific nominal alpha levels of 0.05, 0.01, 0.005, 0.001, and 0.0005 are shown in Graphs 2,3,4,5, and 6.



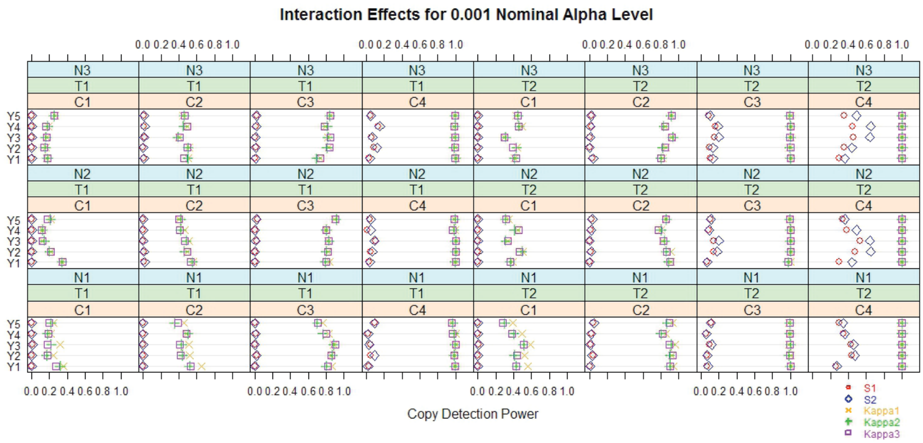
*Graph 2*. Copy detection powers of indices for 0.05 nominal alpha level.



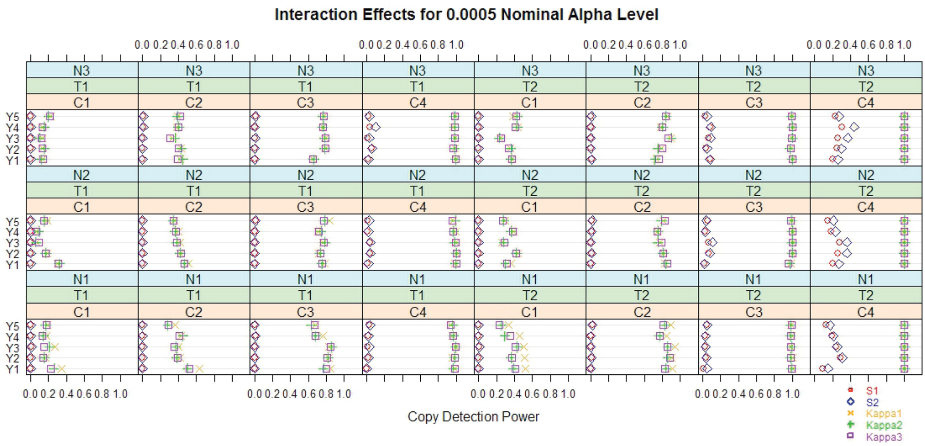
*Graph 3.* Copy detection powers of indices for 0.01 nominal alpha level.



*Graph 4.* Copy detection powers of indices for 0.005 nominal alpha level.



*Graph 5*. Copy detection powers of indices for 0.001 nominal alpha level.



*Graph 6.* Copy detection powers of indices for 0.0005 nominal alpha level.

When investigating the graphs, power of indices increased with the simultaneous increase of sample size, test length, and amount of copying for all specified nominal alpha levels. In addition, powers of kappa statistics were observed to be higher than the $S_1$ and $S_2$ indices for all conditional levels.

$S_1$ and $S_2$ had more similar values when the amount of copying was between 10% and 20% for sample sizes of 100. The power of $S_2$ index was observed to increase when the amount of copying was between 30-40%, especially with larger sample sizes. In the interaction effects of conditions, $S_1$ and $S_2$ had the highest values when the amount of copying was between 30-40%. Additionally, an increase of power was observed when the source ability level was between 50-79% for the specified levels of the condition, and the lowest power value was observed for the 40-49% and 80-90% intervals.

According to the graphs for cases with a 30-40% copying ratio, kappa statistics' power approached 1.00. For cases with a 10-20% copying ratio, the difference between levels became more explicit. For this low amount of copying, small increases were observed for power of kappa statistics when the source ability level decreased. For sample sizes of 100, Kappa1 had slightly more power than the other kappa statistics.

When observing Graph 2, which includes interaction effects of conditions on powers for 0.05 nominal alpha level, distinct from the other nominal alpha levels, all Kappa statistics had values closer to each other for all conditions. In cases with amount of copying between 10-20% and test length of 40 items, little decrease in the power of kappa statistics was observed with increases in the source ability level.

Again, distinct from the other nominal alpha levels, the power of $S_1$ and $S_2$ increased with increases in the source ability level for cases having sample sizes of 2,000, a test length of 40 items, and a 30% copying ratio, as well as those having sample sizes of 2,000, a test length of 80 items, and a 20% copying ratio.

When investigating Graphs 3, 4, 5, and 6, distinct from Graph 2, kappa statistics were observed to differ under certain copy detection conditions. Kappa1 statistic was observed to have slightly more power for sample sizes of 100 with the subset of 10% copying ratio for the 0.01 and 0.005 nominal alpha levels, and with the subsets of both 10% and 20% copying ratios for the 0.001 and 0.0005 nominal alpha levels.

The power of kappa statistics was observed to decrease with increases in source ability level. However, when investigating Graphs 5 and 6, small increases in the power of statistics were observed as the source ability level increased for sample sizes of 2,000.

## Discussion

In this study, Type I error rates and copy detection powers of $S_1$ and $S_2$ indices and kappa statistics were determined and compared with each other under various conditions. When comparing the indices' powers, kappa statistics were found to be more powerful than the other indices in all three copying groups under all conditions. However, the Type I error rate of kappa statistics for detecting answer copying was found to be quite high at all specified nominal alpha levels.

In line with other studies in the literature (Sotaridona & Meijer, 2003; Wollack, 2006), Type I error rates of $S_1$ and $S_2$ indices were found to be quite low at certain nominal levels. Sotaridona and Meijer (2003) and Wollack (2006) concluded in their studies that the powers of the indices also increased as sample size, test length, and amount of copying increased, and that the $S_1$ index is weaker than $S_2$ index at determining answer copying. The results of their studies also indicate that both $S_1$ and $S_2$ indices are found to be insufficient at determining answer copying with amount of copying between 10-20%. Similar results have also been obtained in the current study; this shows that the $S_2$ index, which not only matched source and copier's incorrect answers but also their correct answers, has more power than the $S_1$ index at determining answer copying.

According to the results, Kappa1, Kappa2, and Kappa3 statistics, which were calculated using copying groups of three different sizes, had similar Type I error rates and power at determining answer copying. However, the Kappa1 statistic, which was calculated by the group created according to the copier's correct answers, was found to have a little more power than Kappa2 and Kappa3 statistics in sample sizes of 100, but also having more Type I errors than Kappa2 and Kappa3.

The Type I error rate of kappa statistics at determining answer copying was lowest with sample sizes of 500, test length of 40 items, and source ability level between 50-59%. However, due to the fact that Type I error rates at these levels were higher than the specified nominal alpha levels, the error rates can also be concluded as high at copy detection. Likewise, the research results show that error rates increase as the source ability level and test length increased.

As expected, research results indicate that the more the test length and amount of copying increased, the more powerful the indices became at determining answer copying because the number of matching incorrect answers of the source and the copier increased. However, the power of kappa statistics was observed to slightly decline at determining answer copying as sample size increased.

The power of $S_1$ and $S_2$ indices at determining answer copying increases as sample size increases. In small samples, the power of the two indices at determining answer copying is low and very similar. As sample size increases, the variations also increase; the $S_2$ index was found to have a lower error rate and greater power, especially when the amount of copying was high. However, $S_1$ and $S_2$ indices were found to be weak at determining answer copying when the amount of copying was set between 10-20%, as the studies in the literature (Sotaridona & Meijer, 2003; Wollack, 2006) have also shown.

When the power of indices at determining answer copying is examined with regard to the source ability level, the power of $S_1$ and $S_2$ indices were found to increase when the source ability level was between 50-79%, while decreasing at the upper end of the intervals. This is considered to result particularly from the power of $S_2$ index increasing more at this level than the other intervals of ability level; hence, the increase in the number of matching correct and incorrect answers of the source and copier was high. A very small decline appeared in the power of kappa statistic at determining answer copying as the source ability level increased in the three groups.

In conclusion, the kappa statistic can be said to be more powerful than the $S_1$ and $S_2$ indices under all conditions. However, the Type I error rate of kappa statistic being high under all the circumstances should not be overlooked. Additionally, the fact that variance can be negative while calculating kappa statistics, albeit rather improbable, is one of the limitations of this statistic in determining copies. In this study, data with negative variances were not included when calculating the power and Type I error rates.

The results of this study were found to parallel other study results in the literature (Sotaridona & Meijer, 2003; Wollack, 2006). Type I error rates of $S_1$ and $S_2$ indices were observed to be low, and their power at determining answer copying increased as the sample size, test length, and amount of copying also increased. However, the $S_2$ index was found to be more powerful than the $S_1$ index, especially at greater amounts of copying and larger samples.

Unlike the other studies in the literature, this study investigated the effect of the source ability level on Type I error rates and power of indices at determining answer copying. The results show that when the source has a medium (50-79%) level of ability, the $S_1$ and $S_2$ indices had more power. However, this weakened a little at the upper end of the intervals. Results also show that when the source has a high level of ability, kappa statistics had an increase in Type I errors and a small decrease in power.

Although the copying groups formed in the calculation of the kappa statistic used to determine answer copying had a small effect on the power of this statistic, using it to calculate the statistics of a group formed according to the copier's correct answers can be suggested if necessary, especially in small samples. However, Kappa1 statistic's Type I error rates being higher than the others should not be ignored.

The overall conclusion is that kappa statistics were observed to be more powerful than $S_1$ and $S_2$ indices, while also having higher Type I error rates. This shows the high possibility of an erroneous "Copy" decision in the process of copy detection using kappa statistics. Therefore, $S_2$ is considered appropriate for use in copy detection, especially in large samples and in cases of high amounts of copying.

## Suggestions

1. In this study, $S_1$ and $S_2$ indices and kappa statistics were used as copying detection methods. The same study could be conducted using different methods of determining answer copying.
2. Another study with different sample sizes, test lengths, various source ability levels, and amount of copying can be conducted to examine their impacts on power and Type I error rates of the copy detection methods used in this study.
3. Similar studies can be conducted using copying groups of different sizes for calculating kappa statistics. Researchers can offer different recoding methods.
4. This study was conducted using simulated data. The same study could be done using real data.

## References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of American Statistical Association, 69*, 44–49.

Anikeef, A. M. (1954). Index of collaboration for test administrators. *Journal of Applied Psychology, 38*, 174–177.

Argenal, R. N., Co, F. F., Cruz, E., & Patungan, W. R. (2004, October). *A new index for detecting collusion and its statistical properties.* Paper presented at the 9th National Convention on Statistics, Mandaluyong City, Phillipines.

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*(5), 391–410.

Assessment Systems Corporation. (1993). *Scrutiny!: Software to identify test misconduct* [Computer software]. San Antonio, TX: Advanced Psychometrics.

Bay, M. L. G. (1994). Detection of copying on multiple-choice examinations. *Dissertation Abstracts International*: *Section A, Humanities and Social Sciences*, *56*(3-A), 899.

Belleza, F. S., & Belleza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16*(3), 151–155.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and *K*-index. *Applied Psychological Measurement, 34*(6), 379–392.

Belov, D. I., & Armstrong, R. D. (2011). Distributions of the Kullback–Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology, 64*(2), 291–309.

Bird, C. (1927). The detection of cheating in objective examinations. *School and Society, 25*, 261–262.

Bird, C. (1929). An improved method of detecting cheating in objective examinations. *Journal of Educational Research, 25*, 261–262.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

Deng, W. (2008). *An innovative use of the standardized log-likelihood statistic to evaluate person fit.* Dissertation Abstracts International Section A: Humanities and Social Sciences. Rutgers State University, NJ.

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education, 6*(2), 153–65.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 6*, 152–165.

Giordano, C., Subhiyah, R., & Hess, B. (2005, April). An analysis of item exposure and item parameter drift on a take-home recertification exam. Paper presented at the *Annual Meeting of the American Educational Research Association*. Montreal, Canada.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying* (ACT Research Report Series No. 87–15). Iowa City, IA: American College Testing.

Harpp, D. N., & Hogan, J. J. (1993). Crime in the classroom - Detection and prevention of cheating on multiple-choice exams. *Journal of Chemical Education, 70*, 306–311.

Harpp, D. N., Hogan, J. J., & Jennings, J. S. (1996). Crime in the classroom - Part II, an update. *Journal of Chemical Education, 73*(4), 349–351.

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 RR-94-4). Princeton, NJ: Educational Testing Service.

Hui, H. F. (2010). Stability and sensitivity of a model-based person-fit index in detecting item pre-knowledge in computerized adaptive test. *Dissertation Abstracts International. Section A: Humanities and Social Sciences*. University of Hong Kong.

Kadane, J. B. (1999). An allegation of examination copying. *Chance, 12*(3), 32–36.

Lewis, C., & Thayer, D. T. (1998). *The power of the K-index (or PMIR) to detect copying* (Research Report RR-98-49). Princeton, NJ: Educational Testing Service.

Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement, 20*, 475–489.

Shu, Z. (2010). *Detecting test cheating using a deterministic, gated item response theory model* (Doctoral dissertation, University of North Carolina, Greensboro, NC). Retrieved from https://libres.uncg.edu/ir/uncg/f/Shu_uncg_0154D_10504.pdf

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement, 30*(5), 412–431.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the *K*-index for detecting answer copying. *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. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204.

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*(4), 307–320.

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indices to ımprove detection rates. *Applied Measurement in Education*, *19*(4)*,* 265–288.

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