*Article*

*Journal of Educational and Behavioral Statistics*
*2015, Vol. 40, No. 5, pp. 435–453*
*DOI: 10.3102/1076998615595628*


# On the Optimality of Answer-Copying Indices: Theory and Practice

**Mauricio Romero**
*University of California-San Diego and Quantil, Matemáticas Aplicadas*

**Álvaro Riascos**
*Universidad de los Andes and Quantil, Matemáticas Aplicadas*

**Diego Jara**
*Quantil, Matemáticas Aplicadas*

*Multiple-choice exams are frequently used as an efficient and objective method to assess learning, but they are more vulnerable to answer copying than tests based on open questions. Several statistical tests (known as indices in the literature) have been proposed to detect cheating; however, to the best of our knowledge, they all lack mathematical support that guarantees optimality in any sense. We partially fill this void by deriving the uniformly most powerful (UMP) test under the assumption that the response distribution is known. In practice, however, we must estimate a behavioral model that yields a response distribution for each question. As an application, we calculate the empirical type I and type II error rates for several indices that assume different behavioral models using simulations based on real data from 12 nationwide multiple-choice exams taken by fifth and ninth graders in Colombia. We find that the most powerful index among those studied, subject to the restriction of preserving the type I error, is one based on the work of Wollack and is superior to the index developed by Wesolowsky.*

Keywords: ω *index; answer copying; false discovery rate; Neyman–Pearson lemma*

## 1. Introduction

Multiple-choice exams are frequently used, as they are considered by many to be an efficient and objective way of evaluating knowledge. Nevertheless, they are more vulnerable to answer copying than tests based on open questions. Answer-copy indices provide a statistical tool for detecting cheating by examining suspiciously similar response patterns between two students. However, these indices have three problems. First, similar answer patterns between a pair of

students may occur without answer copying. For example, two individuals with very similar educational backgrounds are likely to provide similar answers. The second problem is that a statistical test (an index) is by no means a conclusive basis for accusing someone of copying, since it is impossible to completely eliminate type I errors. In other words, it is possible that two individuals may share the same response pattern by chance. Finally, every index assumes responses are stochastic. If the assumed probability distribution is incorrect, the index can lead to incorrect conclusions. Furthermore, all the indices in the literature are ad hoc and there are no theoretical results that support the use of one index over the other.

Wollack (2003) compares several indices using real data and finds that among those that preserve size (i.e., indices that have an empirical type I error rate below the theoretical one. That is, in practice they are less than or equally likely to erroneously reject the null hypothesis than suggested by the size of the test), the ω index, based on the work of Wollack (1997), is the most powerful one. However, the set of indices studied is not comprehensive and in particular does not include the index developed by Wesolowsky (2000).

Thus, there are two gaps in the literature that this article seeks to fill. First, it provides theoretical foundations for validating the use of indices that reject the null hypothesis of no cheating for a large number of identical answers under the assumption that student responses are stochastic.

Second, we calculate the empirical type I and type II error rates of two refinements of the indices first developed by Frary, Tideman, and Watts (1977), the ω and γ indices based on the work of Wollack (1997) and Wesolowsky (2000), respectively. Using Monte Carlo simulations and data from the SABER tests taken by fifth and ninth graders in Colombia in May and October 2009, we find that the conditional version of the standardized index first developed by Wollack (1997) is the most powerful among those that preserve size.

The article is organized as follows. The second section derives an optimal statistical test (index) to detect answer copying using the Neyman–Pearson lemma (NPL). The third section presents two of the most widely used indices, which are based on the work of Wollack (1997), Frary et al. (1977), Wesolowsky (2000), and van der Linden and Sotaridona (2006). The fourth section presents a brief summary of the data used and is followed by a section that presents the methodology of the Monte Carlo simulations used to find the empirical type I and type II error rates (to test which behavioral model gives the best results) and its results. Finally, the sixth section concludes.

## 2. Applying the NPL to Answer Copying

It is normal for two answer patterns to have similarities by chance. Answer-copying indices are used to detect similarities that are so unlikely to happen naturally that answer copying becomes a more natural explanation

than chance. Most answer-copy indices are calculated by counting the number of identical answers between the test taker suspected of copying and the test taker suspected of providing answers. For examples, see van der Linden and Sotaridona (2004, 2006); Sotaridona and Meijer (2003, 2002); Sotaridona, van der Linden, and Meijer (2006); Holland (1996); Frary et al. (1977); Cohen (1960); Bellezza and Bellezza (1989); Angoff (1974); Wesolowsky (2000); and Wollack (1997). In all these indices, the null hypothesis is the same: There is no cheating.

All these indices are ad hoc since they are not derived to be optimal in any sense. To the authors' knowledge, this article presents the first effort to rationalize the use of these indices to detect answer copying using the NPL (Neyman & Pearson, 1933), resulting in the uniformly most powerful (UMP) test (index), assuming we know the underlying probability that two individuals have the same answer in each question. However, we must turn to empirical data to find the performance of each index since different behavioral models result in different response probability distributions.

First, we state the problem formally. Let us assume that there are $N$ questions and $n$ alternatives for each question. We are interested in testing whether the individual who cheated (denoted by $c$) copied from the individual who supposedly provided the answers (denoted by $s$). Let $\gamma_{cs}$ be the number of questions that $c$ copied from $s$. The objective is to test the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 &: \gamma_{cs} = 0 \\ \mathrm{H}_1 &: \gamma_{cs} > 0 \end{aligned}$$

Let $I_{csi}$ be equal to 1 when individuals $c$ and $s$ have the same answer to question $i$ and 0 otherwise. Then, the number of common answers between $c$ and $s$ can be expressed as:

$$M_{cs} = \sum_{i=1}^{N} I_{csi}. \tag{1}$$

Under the null hypothesis $M_{cs}$ is the sum of $N$ independent Bernoulli random variables, each with a different probability of success $\pi_i$, equal to the probability that individual $c$ has the same answer as individual $s$ in question $i$. The distribution of $M_{cs}$ is known as a Poisson binomial distribution. Let $B(\pi_1, ..., \pi_N)$ be that distribution and $f_N(x; \pi_1, ..., \pi_N)$ be the probability mass function (pmf) at $x$. Note that $f_N(x; \pi_1, ..., \pi_N) = \sum_{A \in F_x} \left( \prod_{i \in A} \pi_i \right) \left( \prod_{j \in A^c} (1 - \pi_j) \right)$, where $F_x = \{A : \mathrm{A} \subset \{1, ..., N\}, |A| = x\}$. If $\pi_1 = \pi_2 = \cdots = \pi_N = \pi$, then the Poisson binomial distribution reduces to a standard binomial distribution. Although computing $f_N$ can be computationally intensive, efficient algorithms have been derived by Hong (2013).

Now, let $A$ denote the set of questions that student $c$ copied from $s$. Then if $|A| = k$, it means that $\gamma_{cs} = k$, and $M_{cs}$ has the pmf $\hat{f}_N(x; \pi_1, ..., \pi_N, A)$, where we define $\hat{f}_N(x; \pi_1, ..., \pi_N, A) \doteq f_N(x; \pi_1', .., \pi_N')$ such that

$$\pi_i' = \begin{cases} 1 & \text{if } i \in A \\ \pi_i & \text{if } i \notin A \end{cases}$$

For example, say that there are 50 questions and that the students copied questions 1, 10, and 50 (i.e., $A = \{1, 10, 50\}$), then

$$\hat{f}_N(x; \pi_1, ..., \pi_N, A) = f_N(x; 1, \pi_2, ..., \pi_9, 1, \pi_{11}, ..., \pi_{49}, 1).$$

Before we continue, let us state the NPL:

**Theorem 1:** NPL (Casella & Berger, 2002)

Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, where the pmf is $f(x|\theta_i)$, $i = 0, 1$, using a statistical test (index) with rejection region $R$ (and therefore its complement, $R^c$, is the nonrejection region) that satisfies

$$\begin{array}{ll} x \in R & \text{if } f(x|\theta_1) > f(x|\theta_0)\delta \\ x \in R^c & \text{if } f(x|\theta_1) < f(x|\theta_0)\delta \end{array} \tag{2}$$

for some $\delta \geq 0$, and

$$\alpha = P_{H_0}(\boldsymbol{X} \in R) \tag{3}$$

where $P_{H_i}(\boldsymbol{X} \in A) := P(\boldsymbol{X} \in A)$ if $\theta = \theta_i$. Then

1. (Sufficiency) Any test (index) that satisfies Equations 2 and 3 is a UMP level $\alpha$ test (index).
2. (Necessity) If there exists a test (index) satisfying Equations 2 and 3 with $\delta > 0$, then every UMP level $\alpha$ test (index) is a size $\alpha$ test (index)—satisfying 3—and every UMP level $\alpha$ test (index) satisfies 2 except perhaps on a set $A$ such that $P_{H_o}(\boldsymbol{X} \in A) = P_{H_1}(\boldsymbol{X} \in A) = 0$.

Notice that NPL implies that a likelihood ratio test is the UMP test for simple hypothesis testing. Let us apply the NPL to the simple hypothesis test $H_0 : A = A_0$ and $H_1 : A = A_1$, where $A_0 = \emptyset$ (i.e., there is no cheating) and $A_1$ is a set of questions. If in the data we observe $x$ questions answered equally by individuals $c$ and $s$, then the likelihood ratio test would be:

$$\lambda^A(x) = \frac{\hat{f}_N(x; \pi_1, ..., \pi_N, A_1)}{\hat{f}_N(x; \pi_1, ..., \pi_N, A_0)} = \frac{\hat{f}_N(x; \pi_1, ..., \pi_N, A_1)}{f_N(x; \pi_1, ..., \pi_N)}$$

Now we must find the critical value of the test. In other words, we need the greatest value $c$ such that under the null we have:

$$1 - P_{H_o}\left(\frac{\hat{f}_N(x;\pi_1,...,\pi_N,A)}{f(x;\pi_1,...,\pi_N)} < c\right) = P_{H_o}\left(\frac{\hat{f}_N(x;\pi_1,...,\pi_N,A)}{f_N(x;\pi_1,...,\pi_N)} > c\right) \leq \alpha$$

For any given pair of simple hypotheses ($H_0 : A = A_0$, $H_1 : A = A_1$), we know how to find the UMP (by using the NPL) test. The following lemma will allow us to find the UMP test for more complex alternative hypothesis (e.g., $H_1 : \{A : |A| \geq 1\}$) as it lets us exploit the fact that distribution families with the monotone likelihood ratio property have a UMP that does not depend on the alternative hypothesis (see section 3.4 in Lehmann & Romano, 2005).

**Lemma 1:** $\lambda^A(x) = \frac{\hat{f}_N(x;\pi_1,...,\pi_N,A)}{f_N(x;\pi_1,...,\pi_N)}$ is increasing in $x \in \{0,...,N\}$ for all $A$.

Before we present the proof, we must first recall some useful results proved by Wang (1993).

**Theorem 2:** Theorem 2 in Wang (1993). The pmf of a Poisson binomial satisfies the following inequality:

$$f_N(x;\pi_1,\pi_2,...,\pi_N)^2 > C(x) f_N(x+1;\pi_1,\pi_2,...,\pi_N)\, f_N(x-1;\pi_1,\pi_2,...,\pi_N)$$

where $C(x) = \max(\frac{x+1}{x}, \frac{N-x+1}{N-x})$

which has as an immediate corollary:

**Corollary 1:** The pmf of a Poisson binomial satisfies the following inequality:

$$f_N(x;\pi_1,\pi_2,...,\pi_N)^2 > f_N(x+1;\pi_1,\pi_2,...,\pi_N)\, f_N(x-1;\pi_1,\pi_2,...,\pi_N)$$

Now we are ready to prove the lemma:

*Proof of Lemma 1.* The proof will be done by induction on the size of $A$.

*Base Case*

First, consider the case $|A| = 1$. Without loss of generality, as the pmf is invariant to permutations of the $\pi_i$'s (Wang, 1993), assume $A = \{1\}$. The numerator in the lemma's quotient is 0 for $x = 0$, so we proceed to prove monotonicity $\lambda^A(x)$ in $x$ for $x \geq 1$. Likewise, the case $N = 1$ follows trivially, so we assume $N > 1$.

For simplicity, we call $g(x) = f_{N-1}(x;\pi_2,\ldots,\pi_N)$. First, note that

$$\hat{f}_N(x;\pi_1,\ldots,\pi_N;A) = g(x-1).$$

Second, Corollary 1 states that $g(x-1)g(x+1) < g(x)^2$. Third, we can write $f_N(x;\pi_1,\pi_2,\ldots,\pi_N) = \pi_1 g(x-1) + (1-\pi_1)g(x)$. With these observations, we have

$$\begin{aligned}\frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N;A)}{f_N(x;\pi_1,\ldots,\pi_N)} &= \frac{g(x-1)}{\pi_1 g(x-1)+(1-\pi_1)g(x)} \times \frac{\pi_1 g(x)+(1-\pi_1)g(x+1)}{\pi_1 g(x)+(1-\pi_1)g(x+1)} \\ &< \frac{\pi_1 g(x)g(x-1)+(1-\pi_1)g(x)^2}{[\pi_1 g(x-1)+(1-\pi_1)g(x)][\pi_1 g(x)+(1-\pi_1)g(x+1)]} \\ &= \frac{g(x)}{\pi_1 g(x)+(1-\pi_1)g(x+1)} \\ &= \frac{\hat{f}_N(x+1;\pi_1,\ldots,\pi_N;A)}{f_N(x+1;\pi_1,\ldots,\pi_N)}.\end{aligned}$$

*Inductive Step*

Suppose $\lambda^A(x) = \frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N,A)}{f_N(x;\pi_1,\ldots,\pi_N)}$ is increasing in $x \in \{0,\ldots,N\}$ for all $A$, such that $|A| = k$. Without loss of generality, consider a set $A$ such that $1 \notin A$ and $|A| = k$. Let $\hat{A} = A \cup \{1\}$ (so $|\hat{A}| = k+1$). Then,

$$\begin{aligned}\lambda^{\hat{A}}(x) &= \frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N,\hat{A})}{f_N(x;\pi_1,\ldots,\pi_N)} = \frac{\hat{f}_N(x;1,..,\pi_N,A)}{f_N(x;\pi_1,\ldots,\pi_N)} = \frac{\hat{f}_N(x;1,..,\pi_N,A)}{f_N(x;1,\ldots,\pi_N)} \times \frac{f_N(x;1,\ldots,\pi_N)}{f_N(x;\pi_1,\ldots,\pi_N)} \\ &= \frac{\hat{f}_N(x;1,..,\pi_N,A)}{f_N(x;1,\ldots,\pi_N)} \times \frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N,\{1\})}{f_N(x;\pi_1,\ldots,\pi_N)} < \frac{\hat{f}_N(x+1;1,..,\pi_N,A)}{f_N(x+1;1,\ldots,\pi_N)} \\ &\times \frac{\hat{f}_N(x+1;\pi_1,\ldots,\pi_N,\{1\})}{f_N(x+1;\pi_1,\ldots,\pi_N)} = \frac{\hat{f}_N(x+1;\pi_1,..,\pi_N,\hat{A})}{f_N(x+1;\pi_1,\ldots,\pi_N)} = \lambda^{\hat{A}}(x+1)\end{aligned}$$

Given that $\frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N;A)}{f_N(x;\pi_1,\ldots,\pi_N)}$ is increasing in $x$ for all $A$, then we have that for every $c$ there exists a $k^*$ such that $P_{H_o}\left(\frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N,A)}{f_N(x;\pi_1,\ldots,\pi_N)} < c\right) = \sum_{w=0}^{k^*} f_N(w;\pi_1,\ldots,\pi_N)$. The last equality also comes from the fact that $P_{H_o}\left(\frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N;A)}{f_N(x;\pi_1,\ldots,\pi_N)} = a\right) = P_{H_o}(M_{cs} = b) = f_N(b;\pi_1,\ldots,\pi_N)$, where $b$ is such that $\frac{\hat{f}_N(b;\pi_1,\ldots,\pi_N;A)}{f_N(b;\pi_1,\ldots,\pi_N)} = a$. Notice that $b$ is unique due to the strict monotonicity of $\frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N;A)}{f_N(x;\pi_1,\ldots,\pi_N)}$.

In particular for a given size $\alpha$ of the test, we can find $k^*$ such that

$$1 - P_{H_o}\left(\frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N,A)}{f(x;\pi_1,\ldots,\pi_N)} < c\right) = 1 - \sum_{w=0}^{k^*} f(w;\pi_1,\ldots,\pi_N) \leq \alpha$$

Then, if we reject the null hypothesis when $M_{cs} > k^*$, we get the UMP for a particular set $A$. However, the rejection region is the same for all $A$. Thus, if we reject the null hypothesis when $M_{cs} > k^*$, we get the UMP for all $A$ such that $|A| \geq 1$.

The previous derivation is the first, to the best of our knowledge, that guarantees optimality of indices that reject the null hypothesis for large values of $M_{cs}$. In other words, we have derived the most powerful index among those with size $\alpha$ and shown that this index is one that rejects the null hypothesis for large values of $M_{cs}$. As many existing indices count the number of matches and compare them to a critical value, this implies that they have the same functional form as the UMP. Indices that reject the null hypothesis for large values of identical incorrect answers (such as the K-index; Holland, 1996) can only be UMP if we assume that correct answers are never the result of answer copying.

However, an underlying assumption we have used so far is that we observe the value of $\pi_i$ for all $i$. Instead, we observe the actual answers that individuals provided to the questions in the exam and must infer the value of $\pi_i$ for all $i$ from these observations. Therefore, we cannot actually achieve the UMP. The closer we are to correctly estimating the $\pi_i$'s, the closer our index will be to the UMP.

Additionally, our theoretical result does not cover all possible answer-copying indices. For example, in this article, we only consider blind copying events and not shift-copy events in which $c$ copies answers to the next or previous (instead of current) question by mistake. Notice that in the presence of shift-copy events, the number of common answers between $c$ and $s$ might be smaller than $\gamma_{cs}$. Additionally, Belov (2011) showed that in the presence of variable sections (tests often have an ''operational'' section with identical questions for all students and a ''variable'' section with different questions for adjacent students), the model of Poisson trials does not work because each $\pi_i$ becomes a discrete random variable. Belov (2011) develops an index (the VM index) that considers shift-copy events and takes into account the presence of variable sections. Because of its nature, our theoretical results do not speak to the optimality of the VM index.

In a seminal article, Frary et al. (1977) developed the first indices, known as $g_1$ and $g_2$, that reject the null hypothesis for large values of $M_{cs}$. Wollack (1997), van der Linden and Sotaridona (2006), and Wesolowsky (2000) have proposed further refinements of the methods of Frary et al. (1977) methods. The main difference between these indices is how they estimate the $\pi_i$'s. The next section outlines a methodology to compare indices, in terms of their type I and type II error rate, using real data from multiple-choice exams. We present the result of comparing the two widely used indices developed by Wesolowsky (2000) and Wollack (1997), as they have never been compared in the literature before and they both reject the null hypothesis for large values of $M_{cs}$ (and therefore their use is justified by the results from this section).

## 3. Copy Indices

Let us assume that student $j$ has a probability $\pi^j_{iv}$ of answering option $v$ on question $i$. The probability that two students have the same answer on question

$i(\pi_i)$ can be calculated in two ways. First, assuming independent answers, the probability of obtaining the same answer is $\pi_i = \sum_{v=1}^{n} \pi_{iv}^{c} \pi_{iv}^{s}$.

Second, we could think of the answers of individual $s$ as being fixed, as if he or she was the source of the answers and $c$ the student who copies. In the absence of cheating, conditional on the answers of $s$, the probability that individual $c$ has the same answer as individual $s$ in question $i$ is $\pi_i = \pi_{iv_s}^{c}$, where $\pi_{iv_s}^{c}$ is the probability that individual $c$ answered option $v_s$ which was chosen by $s$ in question $i$.

A discussion of these two approaches is given in Frary et al. (1977) and van der Linden and Sotaridona (2006). The first is known as the unconditional index and is symmetric in the sense that the choice of who is $s$ and who is $c$ is irrelevant since $\pi_i$ is the same either way. The second is known as the conditional index and it is not symmetric, opening the possibility that the index rejects the null hypothesis that student $a$ copied from student $b$ but not rejecting the null hypothesis that $b$ copied from $a$. The details of each situation determine which approach is appropriate. If we believe students copied from each other or answered the test jointly then a conditional index is undesirable, but if we believe that a student is the source (for whatever reason) of answers but did not collaborate with the cheater, then a conditional index might be more appropriate. We study both conditional and unconditional indices.

Indices vary along three dimensions. The first dimension is how they estimate $\pi_{iv}^{j}$. The second is whether they are a conditional or an unconditional index. Finally, they vary in how critical values are calculated. They either use the exact distribution (a Poisson binomial distribution) or a normal distribution, by applying some version of the central limit theorem. This is a common practice, as computing the pmf of a Poisson binomial is an NP-hard problem, which can be computationally intensive and often requires summing a large number of small quantities, which can lead to numerical errors.

In order to use the central limit theorem in this context, recall $M_{cs}$ is the sum of $N$ Bernoulli variables and has mean $\sum_{i=1}^{N} \pi_i$ and variance $\sum_{i=1}^{N} \pi_i(1-\pi_i)$. Thus, $\frac{M_{cs} - \sum_{i=1}^{N} \pi_i}{\sqrt{\sum_{i=1}^{N} \pi_i(1-\pi_i)}}$ converges in distribution to a standard normal distribution as $N$ goes to infinity, as long as $\pi_i \in (0,1)$ for all $i$. In practice, this means there is no question with an option that no student will choose (see section 2.7 of Lehmann, 1999, for more details). There are two advantages to the normal approximation. First, critical values are easier to calculate and more precise (computationally) and second, it allows for a finer choice of critical values.

As mentioned before, Frary et al. (1977) developed the first indices that reject the null hypothesis for large values of $M_{cs}$. However, both Wesolowsky (2000) and Wollack (2003) show that variations of the original method proposed by Frary et al. (1977) yield superior results, and in this article, we study the indices

they developed. The first variation is the ω index developed by Wollack (1997) that assumes there is an underlying nominal response model. The second variation is the γ index developed by Wesolowsky (2000).

### 3.1. ω Index

The ω index (Wollack 1997) assumes a nominal response model that allows the probability of answering a given option to vary across questions and individuals. As before, let $N$ be the number of questions and $n$ be the number of alternatives for answering each question. Suppose that an individual with skill $\theta_j$, who does not copy, has a probability $\pi_{iv}$ of choosing option $v$ in response to question $i$. In other words,

$$\pi^j_{iv} \equiv \pi_{iv}(\theta_j) = \frac{e^{\xi_{iv}+\lambda_{iv}\theta_j}}{\sum_{h=1}^{n} e^{\xi_{ih}+\lambda_{ih}\theta_j}}, \tag{4}$$

where $\xi_{iv}$ and $\lambda_{iv}$ are model parameters and are known as the intercept and slope, respectively. The intercept and slope can vary across questions. The parameters of the questions ($\xi_{iv}$ and $\lambda_{iv}$) are estimated using marginal maximum likelihood, while ability is estimated using the Expected A Posteriori (EAP) method. The estimation is performed using the *rirt* package in R (Germain, Abdous, & Valois, 2014). The ability is estimated taking into account that a correct answer to a "difficult" question indicates a higher ability than a correct answer to a "simple" question. More information on marginal maximum likelihood and EAP can be found in van der Linden and Hambleton (1997) and Hambleton, Swaminathan, and Rogers (1991).

Let $\omega_1$ and $\omega_2$ be the unconditional and conditional (exact) versions of this index, following somewhat the $g_1$ and $g_2$ notation of Frary et al. (1977), and let $\omega_1^s$ and $\omega_2^s$ be the standardized versions (i.e., they use the normal distribution to find the critical values of the index). Specifically,

$$\begin{aligned}
\omega_1 &= M_{cs} \sim B(\pi_1, ..., \pi_N) \\
\omega_2 &= M_{cs} \sim B(\pi'_1, ..., \pi'_N) \\
\omega_1^s &= \frac{M_{cs} - \sum_{i=1}^{N} \pi_i}{\sqrt{\sum_{i=1}^{N} \pi_i(1-\pi_i)}} \sim N(0,1) \\
\omega_2^s &= \frac{M_{cs} - \sum_{i=1}^{N} \pi'_i}{\sqrt{\sum_{i=1}^{N} \pi'_i(1-\pi'_i)}} \sim N(0,1),
\end{aligned}$$

where $\pi_i = \sum_{v=1}^{n} \pi^c_{iv}\pi^s_{iv}$ and $\pi^j_{iv}$ is calculated using Equation 4. Similarly, $\pi'_i = \pi^c_{iv_s} = \frac{e^{\xi_{iv_s}+\lambda_{iv_s}\theta_c}}{\sum_{h=1}^{n} e^{\xi_{ih}+\lambda_{ih}\theta_c}}$, where $v_s$ is the answer of individual $s$ to question $i$.

### 3.2. γ Index

The indices developed by Wesolowsky (2000) assume that the probability that student $j$ has the correct answer (option $\hat{v}$) in question $i$ is given by:

$$\pi^{j}_{i\hat{v}} = \left(1 - (1 - r)^{a_j}\right)^{1/a_j}, \tag{5}$$

where $r_i$ is the proportion of students who had the right answer to question $i$. The parameter $a_j$ is estimated by solving the equation

$$\frac{\sum_{i=1}^{N} \pi^{j}_{i\hat{v}}}{N} = c_j,$$

where $c_j$ is the proportion of questions answered correctly by individual $j$. Finally, we need the probability that student $j$ chooses option $v$ among those that are incorrect, which is estimated as the proportion of students with an incorrect answer who chose each incorrect option. Thus, we have an estimate $\pi^{j}_{iv}$ for every individual $j$, every question $i$, and every option $v$. Let us denote by $\gamma_1$ and $\gamma_2$ the unconditional and conditional version of this index and by $\gamma^{s}_{1}$ and $\gamma^{s}_{2}$ their standardized version, respectively. Specifically,

$$\begin{aligned}
\gamma_1 &= M_{cs} \sim B(\pi_1, ..., \pi_N) \\
\gamma_2 &= M_{cs} \sim B(\pi'_i, ..., \pi'_N) \\
\gamma^{s}_{1} &= \frac{M_{cs} - \sum_{i=1}^{N} \pi_i}{\sqrt{\sum_{i=1}^{N} \pi_i(1 - \pi_i)}} \sim N(0, 1) \\
\gamma^{s}_{2} &= \frac{M_{cs} - \sum_{i=1}^{N} \pi'_i}{\sqrt{\sum_{i=1}^{N} \pi'_i(1 - \pi'_i)}} \sim N(0, 1),
\end{aligned}$$

where $\pi_i = \sum_{v=1}^{n} \pi^{c}_{iv} \pi^{s}_{iv}$ and $\pi^{j}_{iv}$ is calculated using Equation 5 if $v$ is the correct answer and if not, as the proportion of students with answer $v$ among those who chose any incorrect option. Finally, $\pi'_i = \pi^{c}_{iv_s} = \left(1 - (1 - r_i)^{a_c}\right)^{1/a_c}$, if individual $s$ chose the correct option; if individual $s$ chose an incorrect answer, then $\pi'_i$ is equal to the proportion of students with answer $v_s$ among those who chose any incorrect option.

Before we compare how the different versions of the ω and the γ index fare in practice, the following section presents the data used.

## 4. Data

### 4.1. Standardized testing in Colombia

In Colombia, all students enrolled in 5th, 9th, and 11th grades, whether attending a private or public school, are required to take a standardized,

TABLE 1.
*Summary Statistics*

| Test | Subject | Grade | Month | Questions | Students | Examination Rooms |
|---|---|---|---|---|---|---|
| 5041F1 | Math | Fifth | May | 48 | 60,099 | 3,421 |
| 5041F2 | Math | Fifth | October | 48 | 403,624 | 31,827 |
| 5042F1 | Language | Fifth | May | 36 | 60,455 | 3,441 |
| 5042F2 | Language | Fifth | October | 36 | 402,508 | 31,642 |
| 5043F1 | Science | Fifth | May | 48 | 60,404 | 3,432 |
| 5043F2 | Science | Fifth | October | 48 | 405,537 | 31,833 |
| 9041F1 | Math | Ninth | May | 54 | 44,577 | 1,110 |
| 9041F2 | Math | Ninth | October | 54 | 303,233 | 9,059 |
| 9042F1 | Language | Ninth | May | 54 | 44,876 | 1,110 |
| 9042F2 | Language | Ninth | October | 54 | 302,781 | 9,044 |
| 9043F1 | Science | Ninth | May | 54 | 44,820 | 1,107 |
| 9043F2 | Science | Ninth | October | 54 | 30,3723 | 9,053 |

*Source:* Instituto Colombiano para la Evaluación de la Educación. Calculations: Authors.

multiple-choice test known as SABER. These exams are intended to measure the performance of students and schools across several areas. The Instituto Colombiano para la Evaluación de la Educación (ICFES), a government institution, is in charge of developing, distributing, and applying these exams. Scores on the test taken in 11th grade are used by most universities in Colombia as an admission criterion, but there are no consequences for fifth and ninth graders based on their test performance. The ICFES also evaluates all university students during their senior year.

### *4.2. Data Used*

In this article, we analyze all the fifth and ninth grade tests for 2009. Each grade (fifth and ninth) takes three tests: science, mathematics, and language. Students at schools whose academic year ends in December (both private and public) take the exam in September, while students at schools whose academic year ends in June (mainly private schools) take the exam in May. In total, there are two dates, two grades, and three subjects, for a total of 12 exams. The following codes are assigned by the ICFES to each exam: per grade, 5 for fifth and 9 for ninth. Per area, 041 for mathematics, 042 for language, and 043 for science. Per date, F1 for May and F2 for October. For example, exam 9041F2 is taken by ninth graders for mathematics in October. A brief overview of each test is presented in Table 1.

The database contains the answers chosen by each individual to every question on all of the tests, as well as the examination room where the exam was taken. The correct answers for each exam are also available.

## 5. Index Comparison

In this section, we compare the different versions of the ω and the γ indices. In order to do this, we evaluate the type I and type II error rates by creating synthetic samples in which we control the level of cheating between individuals.

### *5.1. Methods*

To find the empirical type I error rate, individuals who could not have possibly copied from one another are paired together and tested for cheating using a particular index. This is done by pairing individuals who took the exam in different rooms, thus eliminating the possibility of answer copying to some extent. We cannot rule out the possibility that proctors give out the answers to students, but as these are low-stakes exams for teachers and schools, we do not believe this is a first-order concern. Additionally, as the exam takes place at the same date and time nationwide, we do not believe that students are able to share their answers with students in other examination rooms.

The empirical type I error rate is calculated as the proportion of pairs for which the index rejects the null hypothesis. To find the empirical type II error rate, we take these answer-copy free pairs and simulate copying by forcing specific answers to be the same. The proportion of pairs for which the index rejects the null hypothesis is the power of the index (recall that the power of the test is the complement of the type II error rate, i.e., Power = 100% − Type II Error).

To make things clearer, let $c$ denote the test taker suspected of cheating, $s$ the test taker believed to have served as the source of answers. The steps taken to find the type I error rate and the power of each index are as follows:

1. One hundred thousand pairs are chosen in such a way that for each couple the individuals performed the exam in different examination rooms. Each pair is constructed by randomly selecting two examination rooms and then randomly selecting one student from each examination room. Then within each pair, the students are randomly ordered. The first student is labeled $s$ (the source) and the second student is labeled $c$ (the copier). This distinction is only important for the conditional (subscript 2) version of the indices. The selection process is done with replacement.
2. The answer-copy methodology is applied to these pairs, and the proportion of pairs for which the index rejects the null hypothesis is the empirical type I error rate estimator.
3. To calculate the power of the index, the answer pattern for individual $c$ is changed by replacing $k$ of his answer to match those of individual $s$. For example, let us assume the answer pattern for $s$ is *ACBCDADCDAB*, which means that there were 11 questions and that he or she answered A to the first question, C to the second question, and so on. Also assume that the original answer pattern of $c$ without copying is *DCABCDAABCB*. Let $k$ be 5 and let us assume that the randomly

selected questions were 1, 4, 5, 10, 11. This means that the modified (with copying) answer patterns for *c* will be *ACACDDAABAB*. Specifically,

a. The level of copying *k* (the number of answers transferred from *s* to *c*) is set.
b. *k* questions are selected randomly.
c. Individual c's answers for the *k* questions are changed to replicate those of individual *s*. Answers that were originally identical count as part of the *k* questions being changed.

4. We apply the answer-copy methodology to the pairs whose exams have been altered. The proportion of pairs accused of cheating is the power of the index for a copying level of *k*.

### 5.2. Results

Throughout the analysis, a size ($\alpha$) value of 0.1% is used and the power of the index is calculated at copying levels ($k$) of $1, 5, 10, 15, 20, ..., N$, where $N$ is the number of questions in the exam. To make the results as comparable as possible and reduce the noise generated by using different random draws, the 100,000 pairs are picked first and then the different indices are applied to the same set of randomly generated pairs.

*5.2.1. Type I error rate.* As can be seen in Tables 2 and 3, the $\gamma_2$, $\gamma_2^s$, and $\omega_2$ indices have an empirical type I error rate that is consistently above the theoretical type I error rate of 1 in 1,000 and is also statistically significant. The $\gamma_1$ index (which is the exact index developed by Wesolowsky, 2000) empirical error rate is above the theoretical one in several cases.

Based on these results, we discard the $\gamma_2$, $\gamma_2^s$, and $\omega_2$ indices and restrict the search for the most powerful index to $\gamma_1$, $\gamma_1^s$, $\omega_1$, $\omega_1^s$, and $\omega_2^s$.

*5.2.2. Power of the indices.* Figure 1 shows the power of the $\gamma_1$, $\gamma_1^s$, $\omega_1$, $\omega_1^s$, and $\omega_2^s$ indices in relation to the fifth grade mathematics test taken in May. Notice that the $\omega_2^s$ index is the most powerful for all levels of answer copying. This is true for all exams as shown in Figures A1–A11 in Online Appendix A. Based on the results of the previous section and this section, we believe this favors the use of the $\omega_2^s$ index over all other versions of the $\omega$ index and all versions of the $\gamma$ index.

In other words, the most powerful index among those studied, subject to the restriction of preserving the type I error, uses a nominal response model for item answering, conditions the probability of identical answers on the answer pattern of the individual who provides the answers, and calculates critical values via a normal approximation.

An important caveat is that the $\omega_2^s$ is superior in this data set (across all grades, subjects, and dates), but in other settings different indices could yield better results as they might give better estimates for the $\pi_i$'s. Additionally, the conditional index might work better simply because of our simulation design, where

TABLE 2.
*Type I Error for the γ Indices*

| Exam | Subject | Grade | Month | $\gamma_1$ | $\gamma_2$ | $\gamma_1^s$ | $\gamma_2^s$ |
|---|---|---|---|---|---|---|---|
| 5041F1 | Math | Fifth | May | 0.67 | 2.81*** | 0.43 | 0.74 |
| | | | | (0.08) | (0.17) | (0.07) | (0.09) |
| 5041F2 | Math | Fifth | October | 1.02 | 3.17*** | 0.71 | 1.1 |
| | | | | (0.1) | (0.18) | (0.08) | (0.1) |
| 5042F1 | Language | Fifth | May | 1.01 | 2.09*** | 0.63 | 1.04 |
| | | | | (0.1) | (0.14) | (0.08) | (0.1) |
| 5042F2 | Language | Fifth | October | 1.4*** | 2.33*** | 1.02 | 1.45*** |
| | | | | (0.12) | (0.15) | (0.1) | (0.12) |
| 5043F1 | Science | Fifth | May | 1.01 | 2.33*** | 0.71 | 1.2** |
| | | | | (0.1) | (0.15) | (0.08) | (0.11) |
| 5043F2 | Science | Fifth | October | 0.9 | 2.07*** | 0.74 | 1.38*** |
| | | | | (0.09) | (0.14) | (0.09) | (0.12) |
| 9041F1 | Math | Ninth | May | 1.68*** | 2.38*** | 1.29*** | 1.3*** |
| | | | | (0.13) | (0.15) | (0.11) | (0.11) |
| 9041F2 | Math | Ninth | October | 2.55*** | 2.33*** | 1.93*** | 1.59*** |
| | | | | (0.16) | (0.15) | (0.14) | (0.13) |
| 9042F1 | Language | Ninth | May | 0.69 | 1.86*** | 0.4 | 0.95 |
| | | | | (0.08) | (0.14) | (0.06) | (0.1) |
| 9042F2 | Language | Ninth | October | 0.89 | 1.97*** | 0.54 | 1.2** |
| | | | | (0.09) | (0.14) | (0.07) | (0.11) |
| 9043F1 | Science | Ninth | May | 1.67*** | 2.25*** | 1.27*** | 1.57*** |
| | | | | (0.13) | (0.15) | (0.11) | (0.13) |
| 9043F2 | Science | Ninth | October | 1.41*** | 2.11*** | 1.16* | 1.72*** |
| | | | | (0.12) | (0.15) | (0.11) | (0.13) |

*Source:* Instituto Colombiano para la Evaluación de la Educación. Calculations: Authors.
*Note.* Number of innocent pairs accused of copying (for every 1,000 pairs) at $\alpha = 0.1\%$. Standard errors in parentheses. For each exam-index combination, we test whether the empirical type I error rate ($\hat{\alpha}$) is greater than the theoretical one $\alpha = 0.1\%$ (i.e., $H_0 : \hat{\alpha} \leq 0.1\%$ vs. $H_1 = \hat{\alpha} > 0.1\%$). The asterisks denote the level at which the null hypothesis can be rejected.
*$p < .10$. **$p < .05$ ***$p < .01$.

one student's answers ($c$) were changed to another student's answers ($s$), which resembles a setting where the cheater copied some answers from the source instead of the source and the cheater collaborating to come up with answers together.

Since different tests have different quantities of questions, this could potentially lead to different results since the standardized indices converge to a normal distribution as the number of questions goes to infinity. In Online Appendix B, we randomly sample 36 questions (the minimum number of questions across all 12 exams) from each exam and repeat the exercise outlined in this section. The overall qualitative results do not change as the same indices ($\gamma_2$, $\gamma_2^s$, and the $\omega_2$)

TABLE 3.
*Type I Error for the ω Indices*

| Exam | Subject | Grade | Month | $\omega_1$ | $\omega_2$ | $\omega_1^s$ | $\omega_2^s$ |
|---|---|---|---|---|---|---|---|
| 5041F1 | Math | Fifth | May | 0.29 | 1.15* | 0.15 | 0.52 |
| | | | | (0.05) | (0.11) | (0.04) | (0.07) |
| 5041F2 | Math | Fifth | October | 0.68 | 1.27*** | 0.46 | 0.7 |
| | | | | (0.08) | (0.11) | (0.07) | (0.08) |
| 5042F1 | Language | Fifth | May | 0.66 | 1.52*** | 0.44 | 0.65 |
| | | | | (0.08) | (0.12) | (0.07) | (0.08) |
| 5042F2 | Language | Fifth | October | 0.89 | 1.66*** | 0.59 | 1.09 |
| | | | | (0.09) | (0.13) | (0.08) | (0.1) |
| 5043F1 | Science | Fifth | May | 0.71 | 1.37*** | 0.5 | 0.82 |
| | | | | (0.08) | (0.12) | (0.07) | (0.09) |
| 5043F2 | Science | Fifth | October | 0.79 | 1.69*** | 0.61 | 1.11 |
| | | | | (0.09) | (0.13) | (0.08) | (0.11) |
| 9041F1 | Math | Ninth | May | 0.96 | 1.37*** | 0.8 | 0.94 |
| | | | | (0.1) | (0.12) | (0.09) | (0.1) |
| 9041F2 | Math | Ninth | October | 1.26** | 1.56*** | 0.95 | 1.03 |
| | | | | (0.11) | (0.12) | (0.1) | (0.1) |
| 9042F1 | Language | Ninth | May | 0.48 | 1.08 | 0.26 | 0.64 |
| | | | | (0.07) | (0.1) | (0.05) | (0.08) |
| 9042F2 | Language | Ninth | October | 0.76 | 1.42*** | 0.56 | 1.03 |
| | | | | (0.09) | (0.12) | (0.07) | (0.1) |
| 9043F1 | Science | Ninth | May | 1.03 | 1.58*** | 0.8 | 1.1 |
| | | | | (0.1) | (0.13) | (0.09) | (0.1) |
| 9043F2 | Science | Ninth | October | 1.06 | 1.68*** | 0.99 | 1.24** |
| | | | | (0.1) | (0.13) | (0.1) | (0.11) |

*Source:* Instituto Colombiano para la Evaluación de la Educación. Calculations: Authors.
*Note.* Number of innocent pairs accused of copying (for every 1,000 pairs) at $\alpha = 0.1\%$. Standard errors in parentheses. For each exam-index combination, we test whether the empirical type I error rate ($\hat{\alpha}$) is greater than the theoretical one $\alpha = 0.1\%$ (i.e., $H_0 : \hat{\alpha} \leq 0.1\%$ vs. $H_1 = \hat{\alpha} > 0.1\%$). The asterisks denote the level at which the null hypothesis can be rejected.
*$p < .10$. **$p < .05$. ***$p < .01$.

have an empirical type I error rate that is consistently above the theoretical type I error rate, and the $\omega_2^s$ index is the most powerful for all levels of answer copying.

## 6. Conclusions

In this article, we justify the use of a variety of statistical tests (known as indices) found in the literature to detect answer copying in standardized tests. In particular, we give grounds to the use of all indices that reject the null hypothesis for large values of the number of answers that pairs of students have in common. We do this by deriving the UMP test (index) using the NPL under the

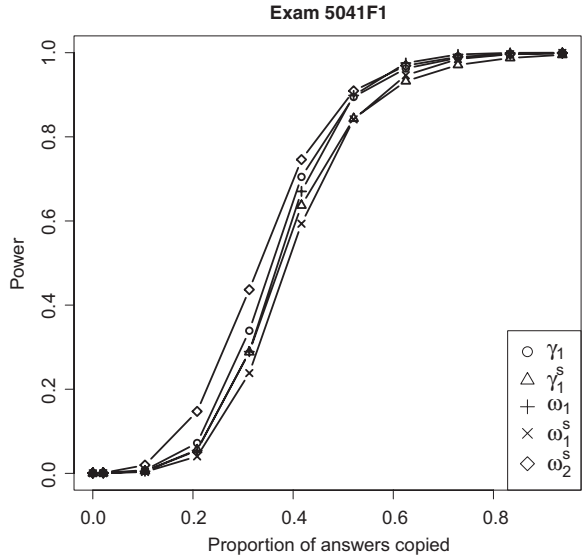


FIGURE 1. *Power in terms of the proportion of answers copied, for all the indices, in the fifth grade mathematics test taken in May.* Source: *Instituto Colombiano para la Evaluación de la Educación. Calculations: Authors.*

assumption that the response distribution is known. We find that the UMP test rejects the null hypothesis for large values of the number of common answers ($M_{cs}$). As many existing indices count the number of matches and compare them to a critical value, this implies that they have the same functional form as the UMP. Indices that reject the null hypothesis for large values of identical incorrect answers (such as the K-index; Holland, 1996) can only be UMP if we assume that correct answers are never the result of answer copying.

In practice, we do not observe the response distribution; instead, we observe the actual answers that individuals provided to the questions in the exam and must infer the response distribution from these observations. The closer we are to correctly estimating the distribution, the closer our index will be to the UMP test. The main difference between indices (that reject $H_0$ for large values of $M_{cs}$) is how they estimate this distribution.

Using data from the SABER fifth and ninth grade tests taken in May and October 2009, in Colombia, we compare eight widely used indices that reject the null hypothesis for large values of the number of common answers ($M_{cs}$) and that are based on the work of Frary et al. (1977), Wollack (1997), Wesolowsky (2000), and van der Linden and Sotaridona (2006). Since all these indices estimate the response distribution differently, in practice they will have different type I and type II error rates. We first filter out the indices that do not meet the theoretical type I error rate

and then select most powerful index among them. We find that the most powerful index, of those that respect the type I error rate, is a conditional index that models student behavior using a nominal response model, conditions the probability of identical answers on the answer pattern of the individual that provides answers, and relies on the central limit theorem to find critical values (which we denote as $\omega_2^s$).

An important caveat is that the $\omega_2^s$ is superior in this data set (across all grades, subjects, and dates), but in other settings different indices could yield better results as they might give better estimates for the $\pi_i$'s. Additionally, the conditional index might work better simply because of our simulation design which resembles a setting where the cheater copied some answers from the source instead of the source and the cheater collaborating to come up with answers together.

These results should have an impact on the academic development and application of these indices. First, it is our hope that future work will provide theoretical proof of the optimality of existing indices that our theoretical result does not cover (e.g., indices that exploit the structure of the test, that consider shift-copy events, that exploit the seating arrangement of the students, among others). Second, we hope that whenever indices are developed in the future, they are accompanied by theoretical support for their optimality. Finally, since many existing indices count the number of matches and compare them to a critical value (which we have proven is the UMP test under our assumptions), empirical simulations such as ours must be conducted in order to determine which behavioral model best approximates the true underlying response pattern, which in turn will indicate which index is best suited for each application.

## Acknowledgments

The authors would like to thank the Instituto Colombiano para la Evaluación de la Educación (ICFES) for financial support during the early stages of this project; the editor, Dan McCaffrey; two anonymous reviewers; Nicola Persico; Decio Coviello; and Julian Mariñno and his group of statisticians for their valuable comments and suggestions on previous versions of this article.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Supplemental Material

The online appendices are available at http://jeb.sagepub.com/supplemental

## References


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

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

Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement*, *35*, 495–517.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Thomson Learning.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *2*, 235–256.

Germain, S., Abdous, B., & Valois, P. (2014). *rirt: Item response theory simulation and estimation* [Computer software manual]. (R package version 1.3.0).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.

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

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, *59*, 41–51.

Lehmann, E. (1999). *Elements of large-sample theory*. New York, NY: Springer.

Lehmann, E., & Romano, J. (2005). *Testing statistical hypotheses*. New York, NY: Springer.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 289–337.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the k-index for detecting answer copying. *Journal of Educational Measurement*, *39*, 115–132.

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

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

van der Linden, W. J., & Hambleton, R. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

van der Linden, W. J., & Sotaridona, L. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement*, *41*, 361–377.

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

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, *3*, 295–312.

Wesolowsky, G. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, *27*, 909–921.

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

Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, *40*, 189–205.

## Authors

MAURICIO ROMERO is a PhD candidate at the University of California, San Diego, 9500 Gilman drive, La Jolla, CA 92093; e-mail: mtromero@ucsd.edu. His research interests are development economics, public finance, and applied econometrics.

ÁLVARO RIASCOS holds a PhD in applied mathematics from Institute of Pure and Applied Mathematics of Rio de Janeiro (IMPA) in Brazil. In 2005 he joined Faculty of Economics at the Universidad de los Andes (Bogotá, Colombia) where he is currently an associate professor. He is founder and co-director of Quantil (www.quantil.com.co), a company founded in 2008 dedicated to the application of mathematics to industry problems. E-mail: alvaro.riascos@quantil.com.co.

DIEGO JARA holds a PhD in Mathematical Finance from Carnegie Mellon University (Pittsburgh, PA). He has over eight years of experience in derivatives markets in investment banks in New York. He is founder and co-director of Quantil (www.quantil.com.co), a company founded in 2008 dedicated to the application of mathematics to industry problems. E-mail: diego.jara@quantil.com.co.

Manuscript received October 06, 2014
First revision received January 19, 2015
Second revision received March 26, 2015
Third revision received May 13, 2015
Accepted May 14, 2015