Detecting Answer Copying Using the Kappa Statistic

Leonardo Sotaridona, Wim J. van der Linden, and Rob R. Meijer,
University of Twente, the Netherlands

A statistical test for detecting answer copying on multiple-choice tests based on Cohen's kappa is proposed. The test is free of any assumptions on the response processes of the examinees suspected of copying and having served as the source, except for the usual assumption that these processes are probabilistic. Because the asymptotic null and alternative distributions of the kappa statistic are derived under the assumption of common marginal probabilities for all items, a recoding of the item alternatives is proposed to approximate this case. The results from a simulation study in this article show that under this recoding, the test approximates its nominal Type I error rates and has promising power functions. *Index terms: answer copying; Kappa statistic; power analysis; Type I error*

In educational testing, the multiple-choice format is often used because it provides an efficient and reliable way of scoring tests for large numbers of test takers. A serious problem with this format, however, is that, unless effective precautions are taken, copying of answers among test takers might occur.

To detect answer copying on multiple-choice tests, both observational and statistical methods can be used (Cizek, 1999). Observational methods use a human observer to establish if answer copying has occurred. The evidence that an observer can collect is observations of certain types of test taker behavior (e.g., one test taker talking to another during the test) or physical evidence (e.g., confiscated cheat sheets exchanged between two test takers). Statistical methods address cheating by modeling the response probabilities of test takers under the assumption of no cheating and looking for patterns of similar answers between test takers that are unlikely under the model.

Several copying statistics have been proposed to detect or back up allegations of answer copying. All these statistics are defined on the response vectors of the test taker suspected of copying and the test taker believed to have served as a source. For simplicity, these test takers are called the copier and the source, respectively. Examples of copying indices are the K-index (Holland, 1996; Lewis & Thayer, 1998) and its variants $\overline{K}_2$ (Sotaridona & Meijer, 2002), S_1, and S_2 (Sotaridona & Meijer, 2003); the B_m-index (Bay, 1995); the g_2-index (Frary, Tideman, & Watts, 1977); and the ω-index (Wollack, 1997; Wollack & Cohen, 1998). For a comprehensive review of copying indices, see Cizek (1999).

Most of the statistics for detecting copying above are based on the number of similar responses between the source and the copier or on a standardization thereof. Under the null hypothesis of no copying, these indices are assumed to follow a distribution, for example, the (generalized)

Applied Psychological Measurement, Vol. 30 No. 5, September 2006, 412–431
DOI: 10.1177/0146621606288891

binomial or Poisson distribution. To evaluate these statistics under the null distribution, parameter estimation may be required, for example, estimation of the item and person parameters in a response model.

A general problem with these statistics is that when the sample size is small, their parameters cannot be estimated reliably. A more vexing problem, however, is that the definitions of the parameters often involve a population of test takers. This population dependence means that pairs of test takers suspected of copying would get a different value for the copying statistic, and hence a different likelihood of being suspected, if they had produced the same pair of response vectors but were included in a different population.

The aim of this study is to investigate a copying statistic with a null distribution that is independent of the behavior of any other test takers than the source and the copier. An advantage of such a statistic is not only that it never penalizes test takers for being "a member of a population" but also that it can be used for a single pair of test takers without requiring any information on other test takers. For two other statistical tests that are not based on population statistics but that do have stronger assumptions on the response processes as the test in the current article, see van der Linden and Sotaridona (2004, in press).

It is emphasized that the meaning of statistical tests for detecting answer copying should not be overrated. These tests can be used as part of a routine screening of large data sets for unlikely agreement between response vectors or as an additional check on observational evidence brought to the attention of the test organization. Some of the statistical tests above are actually used in large-scale testing programs for these purposes. But, like any other statistical test, tests on answer copying entail the possibility of a Type I error. For instance, when used as a routine check of level $\alpha = .01$, these tests are bound to flag 1% of the population as "copiers." Also, the logic of hypothesis testing does not exclude other explanations than answer copying when the null hypothesis is rejected. For example, some test takers may have produced similar patterns of wrong answers because of similar misinformation through common instruction.

Assumptions on Response Process

Consider a test consisting of items $i = 1, \ldots, N$, each with response options $v = 1, \ldots, m$. Index j takes the value c for the test taker suspected of copying and s for the test taker believed to have served as his or her source. In addition, U_{ji} denotes the response of test taker j to item i.

Nothing specific is assumed about the multiple-choice format of the items, except that one alternative is correct and the test takers are instructed to choose the alternative that they believe is true. In addition, it is assumed that the response behavior of c and s is probabilistic (i.e., can be characterized by a probability distribution over the alternatives of each item). But these probabilities are left unspecified and no additional assumption is made that they follow a specific polytomous response model, as has been done, for example, in van der Linden and Sotaridona (in press) and Wollack (1997). In fact, the research reported in this article was just motivated by the question of how much could be inferred about copying behavior of test takers *without* assuming any response model.

Independence and Agreement

The key observation on which the method in this article rests is that if the responses by c and s are probabilistic and the two test takers did not interact in any way, their responses are statistically

independent. If *c* did have access to some of these answers by *s* and copied them, the responses of *c* and *s* on these items would not only be dependent but even in perfect agreement. However, even if the responses by *c* and *s* are independent, some of them agree by chance. The statistical problem we therefore face is to decide how much agreement should be accepted before rejecting the null hypothesis that *c* did not copy any answer. It is again emphasized that, when the null hypothesis is rejected, the logic of statistical testing does not allow us to accept the hypothesis of copying with certainty. Other alternative hypotheses could be possible, such as that of a common history of learning by the two test takers already referred to above.

The notion of agreement between the responses of *c* and *s* can be formalized as follows. Suppose all responses by *c* and *s* on the items in the test are collected in an $m \times m$ table. Responses that are in agreement are classified in the cells on the main diagonal of the table, with responses that do not agree in the off-diagonal cells. This type of table therefore seems appropriate for a statistical analysis of agreement between responses of pairs of test takers to items with a polytomous response format.

It is important to observe that the number of responses in the diagonal is invariant under permutations of the response alternatives over the rows/columns. This property will be used to approximate the conditions under which the statistical test proposed in this article holds.

Conditioning

Some of the statistics for detecting answer copying in the literature are based on the idea to condition on some of the information in the response vector of the source. The reasons for this choice are as follows: (a) Analyzing agreement between incorrect responses only may provide intuitively more powerful evidence of copying (Holland, 1996); (b) the model for the response process implies a conditional statistic, as is the case, for example, with the knowledge-or-copying-or-random-guessing model in van der Linden and Sotaridona (2004); and (c) the statistic was offered as an improvement on a predecessor that used conditioning (Wollack, 1997).

The method presented in this article is based only on the assumption that the responses are probabilistic. Under the null hypothesis of no copying, the only fixed quantity is the number of responses. For the $m \times m$ table introduced above, the hypothesis leads to a multinomial model with a fixed total number of observations but counts in the cells and margins of the table that are random. The only type of conditioning that may seem reasonable is on the incorrect responses by *s*. However, one of the consequences of this choice would be the loss of the row and column in the table for the correct answer and, hence, loss of information (for an example of the loss of a row and a column, see the unconditional coding scheme in Table 3). This study therefore refrains from this type of conditioning.

Kappa

The hypothesis testing problem above exists also in other fields (e.g., psychiatry) when two raters are asked to classify a sample of subjects independently on a scale representing some construct in, for example, a theory of mental health. In these fields, a standard approach is also to tabulate the joint responses of the raters into a two-way table, with one rater represented by the rows and the other by the columns of the table. The statistical test usually conducted to test the hypothesis of no agreement in the table is based on the kappa statistic (Cohen, 1960). The statistical theory of the sampling distribution of kappa began with the derivation of its large-sample standard error

in Fleiss, Cohen, and Everitt (1969). For a review of the theory and some applications of the kappa statistic in medical research, see Agresti (1990, pp. 367-368).

Let π_{vv} denote the probability of a classification in cell (v, v) and π_{v+} and π_{+v} of a classification in row and column v, respectively. These probabilities allow us to calculate a parameter for the agreement between the two raters, which corrects for agreement by chance:

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e}, \tag{1}$$

where

$$\pi_o = \sum_v \pi_{vv} \tag{2}$$

is the probability of an observed agreement between the two raters and

$$\pi_e = \sum_v \pi_{v+}\pi_{+v} \tag{3}$$

the probability of agreement due to chance if the raters operate independently.

Hypotheses

Using parameter κ, the hypothesis to be tested for agreement between raters beyond mere chance can be formulated as

$$\mathrm{H}_0 : \kappa = 0 \tag{4}$$

against

$$\mathrm{H}_1 : \kappa > 0. \tag{5}$$

Null Distribution of $\widehat{\kappa}$

Let $\widehat{\kappa}$ denote the statistics that is obtained by replacing π_o and π_e by the sample statistics $\widehat{\pi}_o = p_o$ and $\widehat{\pi}_e = p_e$, respectively, where

$$p_o = \sum_v p_{vv}, \tag{6}$$

$$p_e = \sum_v p_{v+}\, p_{+v}, \tag{7}$$

and p_{vv}, p_{v+}, and p_{+v} are the empirical proportions in cell (v, v) and row and column v, respectively.

Statistic $\widehat{\kappa}$ is asymptotically normally distributed (Agresti, 1990, p. 366) with mean

$$\mu_{\widehat{\kappa}} = \kappa \tag{8}$$

and variance

$$\sigma^2_{\widehat{\kappa}} = \frac{1}{N}\left\{\frac{\pi_o(1-\pi_o)}{(1-\pi_e)^2} + a + b\right\}, \tag{9}$$

with

$$a = \frac{2(1-\pi_o)\left(2\pi_o\pi_e - \sum_v \pi_{vv}(\pi_{v+} + \pi_{+v})\right)}{(1-\pi_e)^3},$$

$$b = \frac{(1-\pi_o)^2\left(\sum_v \sum_v \pi_{vv}(\pi_{v+} + \pi_{+v})^2 - 4\pi_e^2\right)}{(1-\pi_e)^4},$$

and where N is the number of ratings.

Statistical Test

The following standardization of $\widehat{\kappa}$ is defined as follows:

$$Z_{\widehat{\kappa}} = \frac{\widehat{\kappa} - \mu_{\widehat{\kappa}}}{\sigma_{\widehat{\kappa}}}, \tag{10}$$

with $\mu_{\widehat{\kappa}}$ given in (8) and $\sigma_{\widehat{\kappa}}$ the square root of the variance in (9). Under the null hypothesis in (4), it holds that

$$\mu_{\widehat{\kappa}} = 0,$$

whereas the variance of $\widehat{\kappa}$ in (9) becomes equal to

$$\sigma^2_{\widehat{\kappa}} = \frac{1}{N(1-\pi_e)^2}\left\{\pi_e(1-\pi_e) + \sum_v \sum_v (\pi_{v+}\pi_{+v})(\pi_{v+} + \pi_{+v})^2 - 2\sum_v (\pi_{v+}\pi_{+v})(\pi_{v+} + \pi_{+v})\right\}. \tag{11}$$

To obtain a test statistic for the hypotheses in (4) and (5), it is customary to replace $\sigma^2_{\widehat{\kappa}}$ in (11) by its sample equivalent. The resulting statistic

$$Z_{\widehat{\kappa}} = \frac{\widehat{\kappa}}{\widehat{\sigma}_{\widehat{\kappa}}}, \tag{12}$$

with $\widehat{\sigma}_{\widehat{\kappa}}$ denoting the square root of the sample equivalent of (11), has asymptotic null distribution $Z_{\widehat{\kappa}} \sim N(0,1)$. The test of the null hypothesis under this distribution is right sided with critical value z^* defined by

$$\Pr(Z_{\widehat{\kappa}} \geq z^*) = \alpha. \tag{13}$$

Application to Detection of Copying

The analogy between the problem of detecting answer copying and agreement between ratings seems to suggest that statistic $Z_{\widehat{\kappa}}$ can also be used to test if the response vectors of c and s agree beyond chance. This study will do so while being aware of a potential problem in this application due to the fact that the probabilities that c and s choose an alternative generally differ across items.

As a consequence, the joint response by c and s follows a different multinomial distribution for each item, whereas the sampling model of $Z_{\widehat{\kappa}}$ used to obtain the asymptotic results in Fleiss et al. (1969) is based on the same multinomial distribution for each observation.

The presence of nonidentically distributed observations does not challenge the use of the central limit theorem in the derivation of asymptotic normality of $Z_{\widehat{\kappa}}$ in Fleiss et al. (1969). If the sum of the variances for the cells in the table does not tend to a finite limit, an assumption that is reasonable for our application, a version of the theorem for nonidentically distributed observations holds (Lehmann, 1999, Corollary 2.7.1). But there is a problem related to the probabilities π_{v+} and π_{+v} in (3) and (11). These probabilities are only valid when the assumption of common response probabilities by c and s across all items is met, whereas in fact they do vary.

This article studies the impact of this variation on the Type I error in the simulation study below. Also, it presents the results of a solution to this problem, which consists of permuting the alternatives of the items before pooling their information in the table. As noted earlier, these permutations do not change the number of agreements between the responses in the diagonal of the table but do have an impact on the marginal probabilities for the table. This study capitalizes on this fact by choosing a recoding of the alternative that minimizes the differences between the response probabilities across the items. Computational examples of the types of recoding that is used in this article are given below.

Discussion

Although never discussed in the literature, the same problem is likely to exist for statistical tests of rater agreement based on $\widehat{\kappa}$ in other fields. For example, in studies of clinical diagnosis, it seems hard to believe that clinicians maintain their a priori probabilities of classifying cases in the categories on the scale. Rather, it is expected these probabilities will change during the rating process as a function of the actual cases they have already rated.

It is important to distinguish between the differences in the response probabilities between the items discussed above and the differences between the probabilities for c and s due to their different levels of ability. The result of the latter is different marginal probability distributions for the table and, hence, a maximum possible value of κ smaller than 1 (Cohen, 1960). But this fact does not constrain the conclusions from the statistical test based on statistic $\widehat{\kappa}$ in any way. Because N is the only fixed quantity in the test, lack of agreement between classifications does manifest itself not only as a smaller number of joint responses in the diagonal of the table but also in the form of different marginal distributions. The lower maximum for κ is thus a consequence of the lack of agreement, as it should be.

Power Analysis

The asymptotic distribution of $\widehat{\kappa}$ under the alternative hypothesis is $N(\kappa, \sigma^2_{\widehat{\kappa}})$, with $\sigma^2_{\widehat{\kappa}}$ given by (9). It is thus possible to estimate the (asymptotic) power function of the test, which is given by the probabilities

$$\Pr\left(\frac{\widehat{\kappa} - \kappa}{\sigma_{\widehat{\kappa}}} \geq z^*\right). \tag{14}$$

The only thing needed for this estimate is to calculate an estimate of $\sigma^2_{\widehat{\kappa}}$ in (9) from the sample proportions. Because the power depends on the probabilities π_{vv}, π_{v+}, π_{+v}, and π_o, the power functions for the test can be approximated by generating tables with joint responses of c and s under

various levels of copying and plotting (14) as a function of these levels. In practice, this type of power analysis is thus only possible if the response probabilities of the two test takers are known. An example of this type of power analysis under the extra assumption of a polytomous response model will be given in the simulation studies in the next section.

Simulation Studies

The purpose of these studies is threefold: (a) to study the impact of the differences between the response probabilities of the items on the Type I error of the statistical test based on the kappa statistic, (b) to show the effects of recoding the alternatives of the items before pooling their information in the table, and (c) to explore the power of the test using the estimate proposed in the previous section. This study used multiple-choice tests consisting of 30 and 60 items, each with five response alternatives per item. The ability levels of s and c were varied over the whole range of values that can be met in an application. The significance level used was $\alpha = .05$, and the Type I error was thus evaluated using a critical value for $Z_{\hat{\kappa}}$ equal to 1.645. This choice of significance level was conventional and is not necessarily the right choice in a practical application, where it should be chosen carefully balancing between the Type I and expected Type II errors acceptable to the testing organization and the test takers.

Response Model

To conduct a simulation study, one has to pretend to know the probabilities with which the responses by c and s are generated. This study used the nominal response model for this purpose. Under this model, the probability of test taker j with ability level θ_j responding to option v of item i is given by

$$\pi_{iv}(\theta_j) = \frac{\exp(\zeta_{iv} + \lambda_{iv}\theta_j)}{\sum_{h=1}^{m} \exp(\zeta_{ih} + \lambda_{ih}\theta_j)}, \tag{15}$$

where ζ_{ih} and λ_{ih} are the intercept and slope parameters for item i. Further details of the model can be found in Bock (1972, 1997).

It is emphasized that the assumption of the model in (15) is only an auxiliary assumption that is made to generate the data in this study of the statistical properties of the kappa statistic. If the statistic is applied to actual response data, no assumption of a response model is necessary.

Parameter Values

The impact of the abilities of the source and the copier on κ was controlled by using a grid of possible (θ_c, θ_s)values and then fixing the number of test taker pairs for each grid. The grid was defined on the interval $[-2, 2]$ with an increment of .5 in each component. This interval covers nearly all test takers in a standard, normally distributed population. The number of response vectors generated for each grid point was 2,000. One condition had identical parameter values for all items. Under this condition, the item was chosen to have a probability of .50 for the correct alternative at $(\theta_c, \theta_s) = (0, 0)$. In a second condition, different parameter values were used for the items. In this condition, the slope and intercept parameters of each item were drawn from $U(-1, 1)$ and $U(-1.5, 1.5)$, respectively. (The actual parameter values are available from the authors upon request.)

Table 1

Empirical Type I Error Rates for the Case of Items With Identical Response Probabilities ($\alpha = .05$)

		θ_{source}								
Test Length		–2.0	–1.5	–1.0	–0.5	0	0.5	1.0	1.5	2.0
$N = 30$										
	–2.0	0.02	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00
	–1.5	0.03	0.04	0.03	0.02	0.01	0.00	0.00	0.00	0.00
	–1.0	0.02	0.03	0.04	0.03	0.02	0.01	0.00	0.00	0.00
	–0.5	0.02	0.02	0.03	0.04	0.04	0.03	0.02	0.00	0.00
θ_{copier}	0.0	0.00	0.01	0.02	0.03	0.05	0.03	0.02	0.02	0.01
	0.5	0.00	0.00	0.01	0.02	0.04	0.04	0.03	0.03	0.02
	1.0	0.00	0.00	0.00	0.02	0.02	0.04	0.04	0.03	0.03
	1.5	0.00	0.00	0.00	0.00	0.01	0.04	0.03	0.04	0.04
	2.0	0.00	0.00	0.00	0.00	0.01	0.03	0.02	0.04	0.04
$N = 60$										
	–2.0	0.04	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00
	–1.5	0.03	0.03	0.03	0.02	0.01	0.00	0.00	0.00	0.00
	–1.0	0.02	0.03	0.04	0.02	0.02	0.02	0.00	0.00	0.00
	–0.5	0.01	0.02	0.05	0.04	0.04	0.02	0.01	0.00	0.00
θ_{copier}	0.0	0.00	0.01	0.02	0.04	0.04	0.04	0.03	0.02	0.01
	0.5	0.00	0.00	0.01	0.02	0.03	0.04	0.04	0.03	0.02
	1.0	0.00	0.00	0.01	0.01	0.03	0.04	0.04	0.04	0.04
	1.5	0.00	0.00	0.00	0.00	0.01	0.03	0.04	0.05	0.02
	2.0	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.04	0.04

Generation of Response Vectors

The observed response of test taker j to item i was obtained by drawing a sample from the set $v = \{1, \ldots, 5\}$, where each element of v has a probability of being drawn equal to $\pi_{i1}(\theta_j)$, $\pi_{i2}(\theta_j), \ldots, \pi_{i5}(\theta_j)$ respectively, with $\pi_{iv}(\theta_j)$ computed using (15). As is customary (Sotaridona & Meijer, 2002; Thissen & Steinberg, 1997), the response alternatives with the largest value for the slope parameter were chosen as the correct alternatives of the items.

Study 1: Equal Probabilities for the Items

Table 1 shows the empirical Type I error rates for the case of identical response probability for all items in the test. None of the rates was larger than their nominal value of $\alpha = .05$. The smallest rates were for the combinations of θ values at the opposite ends of the scale. One explanation is that, for these combinations, there are probability distributions for the response alternatives for s and c that are extremely skewed in opposite directions. It is a well-known fact that for such probabilities, the (asymptotic) normal distribution is approximated more slowly in the number of observations than for probabilities close to .50 (e.g., Casella & Berger, 2002, sect. 3.3). In fact, the lengthening of the test from $N = 30$ to $N = 60$ items had only a minor effect for the most extreme combinations of θ values.

Thus, if all items have response probabilities close to each other, the statistical test can be used safely for all test takers. The test then tends to be somewhat conservative—that is, lead to actual probabilities of Type I errors that are smaller than the nominal values of α. The personal bias is that

Table 2

Empirical Type I Error Rates for the Case of ItemsWith Nonidentical Response Probabilities ($\alpha = .05$)

		θ_{source}								
Test Length		–2.0	–1.5	–1.0	–0.5	0	0.5	1.0	1.5	2.0
$N = 30$										
	–2.0	0.90	0.83	0.70	0.48	0.26	0.12	0.04	0.01	0.00
	–1.5	0.81	0.77	0.66	0.49	0.28	0.15	0.06	0.02	0.01
	–1.0	0.70	0.64	0.58	0.44	0.32	0.20	0.12	0.05	0.03
	–0.5	0.47	0.47	0.43	0.38	0.33	0.26	0.18	0.15	0.11
θ_{copier}	0.0	0.26	0.30	0.32	0.32	0.33	0.31	0.31	0.31	0.28
	0.5	0.11	0.13	0.18	0.24	0.35	0.39	0.43	0.48	0.51
	1.0	0.02	0.06	0.10	0.19	0.29	0.44	0.56	0.64	0.72
	1.5	0.01	0.02	0.05	0.13	0.30	0.48	0.62	0.77	0.83
	2.0	0.00	0.01	0.03	0.11	0.28	0.49	0.70	0.87	0.92
$N = 60$										
	–2.0	0.99	0.98	0.89	0.68	0.36	0.11	0.02	0.00	0.00
	–1.5	0.97	0.93	0.83	0.64	0.39	0.17	0.06	0.02	0.01
	–1.0	0.89	0.85	0.72	0.58	0.40	0.26	0.14	0.07	0.03
	–0.5	0.65	0.63	0.58	0.54	0.44	0.34	0.26	0.19	0.13
θ_{copier}	0.0	0.34	0.37	0.42	0.43	0.45	0.45	0.43	0.36	0.36
	0.5	0.12	0.18	0.23	0.37	0.47	0.55	0.60	0.62	0.66
	1.0	0.02	0.05	0.13	0.26	0.42	0.59	0.71	0.79	0.82
	1.5	0.01	0.01	0.07	0.18	0.39	0.59	0.78	0.88	0.93
	2.0	0.00	0.00	0.03	0.14	0.34	0.65	0.86	0.93	0.97

a conservative statistical test of answer copying is to be preferred over one that is liberal. The same point is made for the K index in Holland (1996). A conservative test has a lower likelihood of a decision that is unfavorable to the test taker. The price for this lower likelihood is paid by the testing organization, which loses some of the power of the test to detect answer copying among test takers at the opposite ends of the ability scale.

Study 2: The Impact of Different Probabilities for the Items

The results for the case with different response probabilities for the items are shown in Table 2. These results indicate that, unless an appropriate measure is taken, variation in the probabilities would create a huge problem for a test based on the kappa statistic. All rates were much higher than required, except for the combinations of θ values at the opposite ends of the scale. Thus, if the items differ largely in their response probabilities for the test takers, and nothing is done to decrease this variation, using a test based on statistic κ cannot be recommended for most of the combinations of θ values, except for a few combinations of extreme values in opposite directions. To decrease this variation, response alternatives should be recoded as described in the next study.

Study 3: Unconditional Recoding of the Responses

As already explained, there is no unique way of assigning the response alternatives to the rows/columns in the table for κ. Also, each possible assignment results in the same counts in the diagonal. The freedom can be used to recode the response alternatives to have more acceptable

Table 3

Example of Recoding of Response Alternatives

		a-Value			Recoding			Original Response		Recoded Response	
Item	Answer Key	A	B	C	A	B	C	Source	Copier	Source	Copier
1	B	.15	.60	.25	3	1	2	B	B	1	1
2	A	.80	.15	.05	1	2	3	A	A	1	1
3	A	.70	.12	.18	1	3	2	B	C	3	2
4	C	.30	.17	.53	2	3	1	C	C	1	1
5	B	.40	.55	.05	2	1	3	B	A	1	2
6	A	.45	.25	.30	1	3	2	B	B	3	3

differences in probabilities between the items. In this study, a simple method of unconditional recoding is explored.

The *a*-value of a response alternative is defined as a proportion of test takers who chose the alternative. The method of unconditional recoding assigns the correct alternative for each item to the first row/column of the $m \times m$ table and the other alternatives in decreasing order of their *a*-value to the remaining rows/columns.

More precisely, the method involves the following three steps: (a) calculating the *a*-values, (b) setting up the recoding scheme, and (c) converting the original responses to the codes defined by the scheme. The data in Table 3 illustrate the procedure. In this table, a six-item multiple-choice test consists of three options labeled as *A*, *B*, and *C*. The *a*-values for the alternatives are shown in columns 3, 4, and 5. For each item, the recoded values in columns 6, 7, and 8 reflect the ranks of the *a*-values for the items, with the keyed alternative always assigned the first rank. For example, alternative *B* of Item 1 is the keyed alternative. Therefore, code 1 is assigned to alternative *B*. Because the *a*-value for alternative *C* is larger than for alternative *A*, code 2 is assigned to *C* and code 3 to *A*. The same procedure is applied to the other items. In Table 3, the original response of the source and the copier is shown in columns 9 and 10, whereas the recoded responses are shown in columns 11 and 12. For example, the responses of the source and the copier to Item 5 are *B* and *A*. These responses receive the codes 1 and 2, respectively. A 3×3 table can be constructed based on the recoded responses in Table 3, and the kappa statistic can be computed from this 3×3 table.

For the same data set as in Table 2, the empirical Type I error rates after the use of this recoding method are given in Table 4. Recall that the results in Table 4 were based on the second condition, for which the items had the different parameter values.

The impact of this simple method was already substantial, particularly for the θ values for *c* and *s* in the middle of the scale. For these cases, all error rates were near their nominal values, and a statistical test of answer copying based on statistic κ became possible. Also, the normal approximation seems already to be stable after $N = 30$ items; the increase to $N = 60$ did not introduce much difference in results.

Study 4: Recoding Conditional on θ

If the test is scored using an item response theory (IRT) model, it becomes possible to condition on the ability level of *c*, that is, use only the responses of test takers with approximately the same ability level as *c* to calculate the *a*-values of the items.

Table 4

Empirical Type I Error Rates for the Case of Items With Nonidentical Response Probabilities After Recoding of the Alternatives ($\alpha = .05$)

		θ_{source}								
Test Length		–2.0	–1.5	–1.0	–0.5	0	0.5	1.0	1.5	2.0
$N = 30$										
	–2.0	0.24	0.17	0.11	0.05	0.02	0.00	0.00	0.00	0.00
	–1.5	0.18	0.15	0.11	0.06	0.02	0.01	0.00	0.00	0.00
	–1.0	0.11	0.08	0.09	0.06	0.02	0.02	0.02	0.00	0.00
	–0.5	0.04	0.07	0.05	0.04	0.05	0.04	0.02	0.01	0.01
θ_{copier}	0.0	0.03	0.03	0.03	0.05	0.06	0.06	0.06	0.05	0.03
	0.5	0.00	0.01	0.01	0.04	0.07	0.10	0.12	0.11	0.12
	1.0	0.00	0.00	0.01	0.02	0.05	0.12	0.14	0.18	0.17
	1.5	0.00	0.00	0.00	0.02	0.05	0.11	0.17	0.27	0.31
	2.0	0.00	0.00	0.00	0.01	0.03	0.10	0.21	0.33	0.38
$N = 60$										
	–2.0	0.53	0.33	0.16	0.05	0.01	0.00	0.00	0.00	0.00
	–1.5	0.35	0.23	0.13	0.05	0.02	0.01	0.00	0.00	0.00
	–1.0	0.14	0.14	0.07	0.06	0.04	0.01	0.01	0.00	0.00
	–0.5	0.05	0.05	0.06	0.04	0.05	0.04	0.03	0.02	0.02
θ_{copier}	0.0	0.01	0.02	0.03	0.04	0.06	0.08	0.10	0.07	0.06
	0.5	0.00	0.00	0.02	0.04	0.08	0.15	0.20	0.20	0.20
	1.0	0.00	0.00	0.00	0.02	0.08	0.18	0.26	0.34	0.35
	1.5	0.00	0.00	0.00	0.02	0.08	0.18	0.32	0.44	0.47
	2.0	0.00	0.00	0.00	0.01	0.05	0.18	0.36	0.49	0.57

This study estimated the a-values of the items for c from $1,000$ response patterns simulated for each ability level of c. Table 5 shows the empirical Type I error rates after the recoding of the items conditional on the θ value of c. Again, there were substantial improvements of the Type I error rates of the test. For all practical purposes, the rates were generally close to their nominal level across the whole range of the θ values of c and s. For $N = 60$, the error rates were somewhat higher than for $N = 30$, whereas for high positive θ values for both the source and the copier, they were somewhat inflated. An explanation has not been found for this inflation, which made the statistical test slightly liberal.

Study 5: Recoding Conditional on the Number-Correct Score

In practice, when an estimate of the ability of c in an IRT model may not be available, the number-correct score can be used as an estimate. In this type of conditioning, the response patterns of test takers with similar number-correct scores as the copier can be used to estimate the conditional a-values for the items; otherwise, the recoding remains identical. A computational example of recoding conditional on the number-correct score is given in the appendix.

To calculate the results in this study, 1,000 θ values from $N(0, 1)$ were sampled and item score patterns were generated using the nominal response model in equation (15) with $\lambda \sim U(-1, 1)$ and $\xi \sim U(-1.5, 1.5)$. Then, the score patterns were grouped into six number-correct score categories, and the conditional a-values conditional on these grouped scores were calculated. The range of

Table 5

Empirical Type I Error Rates for the Case of Items With Nonidentical Response Probabilities After Recoding of the Alternatives Conditional on the Ability of the Copier ($\alpha = .05$)

		θ_{source}								
Test Length		–2.0	–1.5	–1.0	–0.5	0	0.5	1.0	1.5	2.0
$N = 30$										
	–2.0	0.06	0.06	0.06	0.06	0.04	0.04	0.03	0.02	0.02
	–1.5	0.05	0.05	0.06	0.05	0.04	0.04	0.04	0.03	0.02
	–1.0	0.06	0.06	0.06	0.06	0.06	0.04	0.04	0.04	0.04
	–0.5	0.05	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.04
θ_{copier}	0.0	0.05	0.05	0.05	0.06	0.05	0.08	0.05	0.04	0.04
	0.5	0.04	0.04	0.06	0.07	0.06	0.06	0.05	0.06	0.07
	1.0	0.04	0.04	0.04	0.04	0.06	0.07	0.07	0.06	0.04
	1.5	0.02	0.03	0.04	0.06	0.05	0.04	0.06	0.06	0.07
	2.0	0.02	0.03	0.03	0.04	0.04	0.06	0.06	0.07	0.06
$N = 60$										
	–2.0	0.08	0.10	0.07	0.07	0.06	0.06	0.04	0.03	0.02
	–1.5	0.07	0.06	0.09	0.06	0.07	0.06	0.04	0.03	0.04
	–1.0	0.08	0.07	0.08	0.07	0.06	0.08	0.06	0.05	0.04
	–0.5	0.08	0.08	0.08	0.07	0.07	0.06	0.07	0.07	0.05
θ_{copier}	0.0	0.08	0.07	0.08	0.08	0.10	0.08	0.09	0.07	0.05
	0.5	0.07	0.08	0.07	0.08	0.08	0.09	0.09	0.10	0.08
	1.0	0.05	0.05	0.08	0.06	0.08	0.08	0.10	0.10	0.11
	1.5	0.04	0.06	0.06	0.06	0.08	0.10	0.12	0.10	0.11
	2.0	0.04	0.04	0.06	0.08	0.09	0.08	0.12	0.11	0.12

scores for each group is shown in Table 6. The minimum number of simulees in each group was 145. To avoid the impact of sampling error, the Type I errors were computed for each cell using 2,000 pairs of response vectors randomly sampled without replacement. These computations may look somewhat complicated, but this is only because several pairs of s and c are dealt with at the same time. In an actual application, the a-values are calculated directly from the responses of the test takers with the same number-correct score (or the same θ value) as c, and once the a-values are available, the recoding and the computation of the test statistic is straightforward.

Table 6 shows that the empirical Type I errors for the recoding conditional on the number-correct scores are close to those for conditioning on θ. For extreme values in the same direction, the test remained somewhat liberal; for values in the opposite direction, the test became more conservative.

Study 6: Power of the Statistical Test

This study conducted a power analysis under the nominal response model, with recoding conditional on the true ability of c. The items had different slope and intercept parameters, which were drawn from $U(-1, 1)$ and $U(-1.5, 1.5)$, respectively.

Let $\gamma = 1, 2, \ldots, N$ denote the number of items copied by c from s. Recall that N is the number of items in a test. Let (U_c, U_s) denote the pair of response patterns of c and s. Copying was simulated by generating a response pattern $U_{c\gamma}$ from U_c by replacing the responses of c with those of s in U_c on

Table 6

Empirical Type Error Rates for the Case of Items With Nonidentical Response Probabilities After Recoding of the Alternatives Conditional on the Number-Correct Score of the Copier ($\alpha = .05$)

		Score Group					
Test Length		0-6	7-10	11-14	15-18	19-21	22-30
$N = 30$							
	0-6	0.10	0.06	0.03	0.08	0.00	0.00
	7-10	0.06	0.08	0.04	0.02	0.00	0.00
	11-14	0.03	0.05	0.07	0.06	0.02	0.00
	15-18	0.01	0.04	0.06	0.08	0.05	0.02
	19-21	0.00	0.01	0.03	0.05	0.07	0.07
	22-30	0.00	0.00	0.01	0.02	0.06	0.08
		0-13	14-21	22-27	28-34	35-40	41-60
$N = 60$	0-13	0.12	0.10	0.03	0.01	0.00	0.00
	14-21	0.10	0.11	0.09	0.04	0.00	0.00
	22-27	0.06	0.10	0.08	0.08	0.04	0.01
	28-34	0.02	0.05	0.09	0.08	0.08	0.04
	35-40	0.00	0.01	0.04	0.07	0.09	0.07
	41-56	0.00	0.00	0.02	0.03	0.05	0.08

γ randomly selected items. The pairs of response patterns $(U_{c\gamma}, U_s)$ were used to calculate the power in (14) as a function of γ. The power was calculated for the significance level $\alpha = .05$. The process was repeated for all pairs of selected θ values of c and s. These θ values are –1.5, –0.5, 0.5, and 1.5.

The results for $N = 30$ and $N = 60$ are shown in Figures 1 and 2. These figures show that the test had high power (at least .80) when the test taker copied a minimum of 40% of the items for $N = 30$ (= 12 items) or at least 30% of the items for $N = 60$ (= 18 items). The power is somewhat higher when the ability of the source is near the middle of the scale (-0.5 and 0.5) than near its extremes (-1.5 and 1.5).

Discussion

A statistical test based on statistic κ to detect answer copying in a multiple-choice test was proposed. The basic idea of this study was that answer copying results in more agreement between the response of two test takers than can be expected by chance. The main advantage of choosing κ as a test statistic is that it is a measure of agreement with known asymptotic properties and a proven record in numerous other types of applications. Another feature of κ is that its calculation does not involve any statistics for an (arbitrary) population.

The first empirical study was to show that κ is sensitive to the differences between the response probabilities of the test takers c and s between the items. However, in the subsequent studies, it was demonstrated how this sensitivity can be removed by recoding the responses. The results for the method with the conditional a-values of the alternatives given the observed number-correct score yielded a satisfactory test, which was slightly liberal for extreme scores in the same direction and conservative for extreme scores in the opposite direction. As discussed earlier, for a statistical test on answer copying to be somewhat conservative is no problem. Also, a slight tendency to

Figure 1
Power Functions for $\theta_s = -1.5,\ -0.5, 0.5, 1.5$ and
$\theta_c = -1.5,\ -0.5, 0.5, 1.5$ at Significance Level $\alpha = .05 (N = 30)$

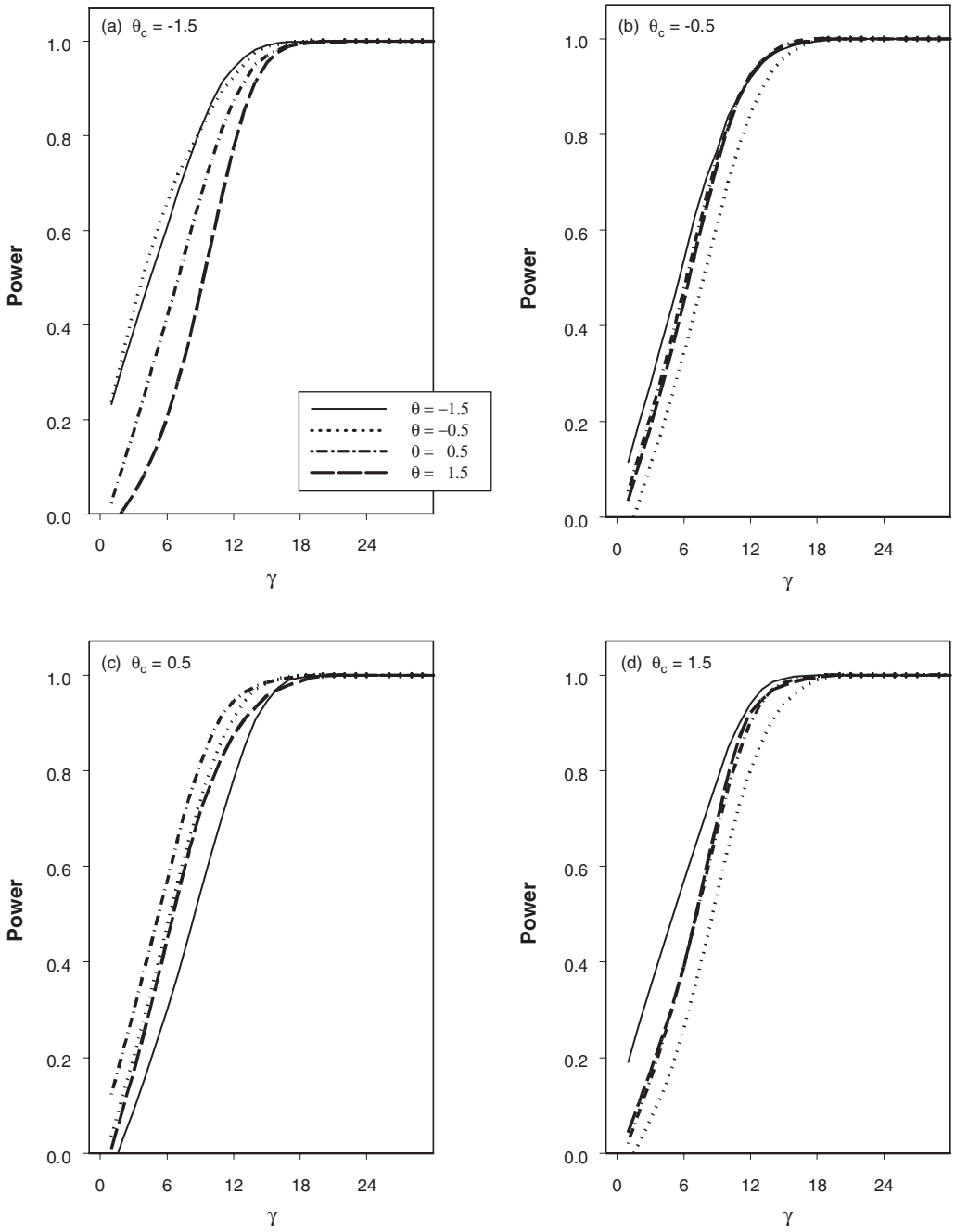
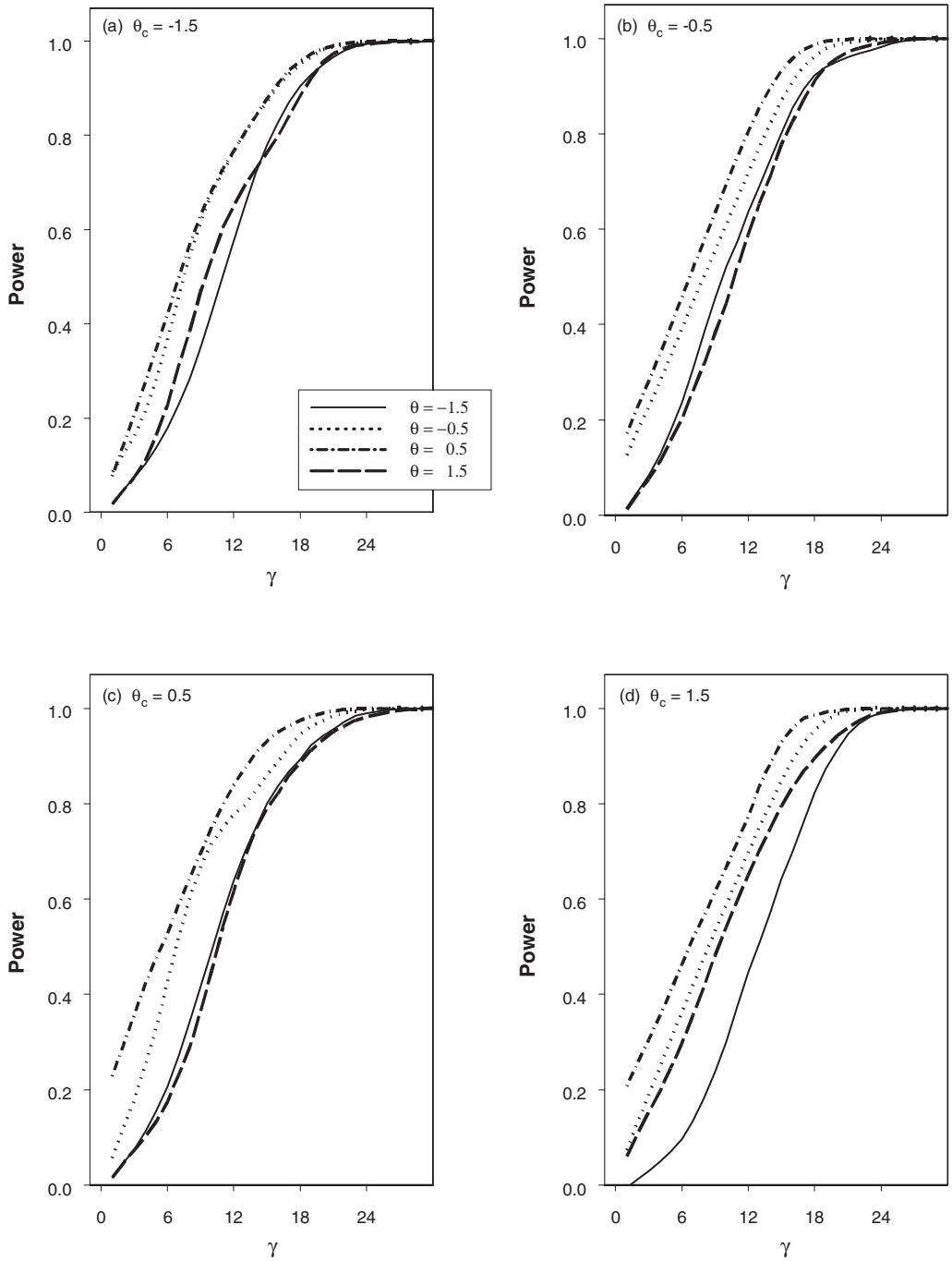


Figure 2
Power Functions for $\theta_s = -1.5, -0.5, 0.5, 1.5$
and $\theta_c = -1.5, -0.5, 0.5, 1.5$ at Significance Level $\alpha = .05 (N = 60)$



a more liberal test for extreme scores in the same direction can easily be remedied by choosing somewhat smaller than nominal significance levels for pairs of test takers with such scores.

More experience with a larger variety of test lengths, response probabilities, and score ranges conditional on which one should recode is needed to be able to generalize the results in the simulation study.

Although the authors of this study were attracted to using κ because it is a measure of agreement independent of the response vectors of any other test takers than c and s, they ended with a procedure that required the presence of a population of test takers to calculate the a-values of the items conditional on the number-correct scores. The result is not disappointing, though: The a-values are only used to recode the response alternatives on the items for c and s; their values are not used in the calculation of the κ statistic. Also, the number-correct score has been proven to be monotone in the ability parameter for a large variety of polytomous IRT models (Hemker, Sijtsma, Molenaar, & Junker, 1996, 1997). Using number-correct scores, therefore, implies no assumption on the response behavior of the test takers other than that their response probabilities follow an (unknown) model in this large collection.

Appendix
Computational Example of the Kappa Statistic

This appendix illustrates how to conduct a statistical test of answer copying based on the kappa statistic when the a-values are to be computed conditionally on the number-correct score of the copier. The same type of recoding was applied in Study 5 above.

The data in Table A1 are used to test the hypotheses that a copier, c, did not copy from five different sources, s_1, s_2, s_3, s_4, and s_5. The data are from a set of approximately 70,000 students on a large-scale Grade 8 mathematics achievement test consisting of 55 multiple-choice items with four answer categories. The answer key for the items is shown in the second column. The copier was randomly chosen from the examinees with a number-correct score of 30. The five sources were randomly chosen from examinees with number-correct scores 10, 20, 30, 40, and 50. The a-values were estimated using the responses of the 2,434 examinees in the data set with a number-correct score of 30. Because the data set was large, we were able to condition on the exact number-correct score of the copier. For smaller data sets, the a-values should be obtained by grouping the examinees within common score ranges, as described in Study 5.

The recoding scheme follows directly from the rank of the a-values, where the keyed answer category is given the first rank. In the case of a tie, the ranks are assigned randomly. For example, for Item 6, categories A and B had equal a-values; category A was randomly assigned rank 3 and category B rank 4. The recoding scheme is shown in columns 7 through 10. The actual responses of c and the five sources are in columns 11 through 16, respectively, and their recoded responses are in columns 17 through 22.

For all combinations of the copier and the sources, the 4×4 agreement tables are constructed from their recoded responses that are shown in Table A2, where the copier's responses are in the rows and the source's responses in the columns.

Using the table for the pair c-s_1 in Table A2 as an illustration, the test is conducted taking the following steps:

1. Compute the observed agreement: $p_o = (7 + 3 + 0 + 0)/55 = .1818$.

2. Compute the expected agreement: $p_e = (30 \times 10 + 9 \times 21 + 8 \times 14 + 8 \times 10)/55^2 = .2251$.

Table A1

Data for Computational Example With Conditioning on the Number-Correct Score of the Copier

i	Key	*a* value				Recoding Scheme				Actual Responses						Recoded Responses					
		A	B	C	D	A	B	C	D	c	s_1	s_2	s_3	s_4	s_5	c	s_1	s_2	s_3	s_4	s_5
1	B	0.08	0.84	0.03	0.05	2	1	4	3	B	B	B	B	B	B	1	1	1	1	1	1
2	B	0.03	0.91	0.04	0.02	3	1	2	4	A	C	D	B	B	B	3	2	4	1	1	1
3	B	0.09	0.89	0.01	0.02	2	1	4	3	B	D	B	A	B	B	1	3	1	2	1	1
4	D	0.00	0.03	0.01	0.96	4	2	3	1	D	D	D	D	D	D	1	1	1	1	1	1
5	B	0.03	0.80	0.04	0.13	4	1	3	2	A	C	C	B	B	B	4	3	3	1	1	1
6	D	0.02	0.02	0.16	0.80	3	4	2	1	D	C	C	D	D	D	1	2	2	1	1	1
7	C	0.15	0.11	0.72	0.02	2	3	1	4	C	B	C	C	C	C	1	3	1	1	1	1
8	C	0.16	0.13	0.56	0.14	2	4	1	3	C	A	D	C	C	C	1	2	3	1	1	1
9	B	0.26	0.41	0.15	0.18	2	1	4	3	A	A	B	C	B	B	2	2	1	4	1	1
10	C	0.14	0.04	0.69	0.12	2	4	1	3	C	B	A	D	C	C	1	4	2	3	1	1
11	A	0.47	0.11	0.26	0.17	1	4	2	3	A	D	D	A	A	A	1	3	3	1	1	1
12	A	0.85	0.04	0.10	0.01	1	3	2	4	B	D	C	A	A	A	3	4	2	1	1	1
13	A	0.95	0.01	0.00	0.03	1	3	4	2	B	A	A	A	A	A	3	1	1	1	1	1
14	D	0.02	0.01	0.04	0.93	3	4	2	1	D	D	D	D	D	D	1	1	1	1	1	1
15	A	0.11	0.28	0.22	0.4	1	3	4	2	A	B	D	D	C	A	1	3	2	2	4	1
16	D	0.03	0.01	0.02	0.94	2	4	3	1	D	A	A	D	D	D	1	2	2	1	1	1
17	D	0.19	0.03	0.05	0.73	2	4	3	1	D	C	D	D	D	D	1	3	1	1	1	1
18	C	0.04	0.15	0.78	0.03	3	2	1	4	B	C	C	C	C	C	2	1	1	1	1	1
19	C	0.24	0.11	0.49	0.15	2	4	1	3	D	A	D	A	C	C	3	2	3	2	1	1
20	D	0.19	0.38	0.24	0.19	4	2	3	1	C	B	D	D	A	D	3	2	1	1	4	1
21	A	0.66	0.03	0.30	0.01	1	3	2	4	B	C	C	A	A	A	3	2	2	1	1	1
22	C	0.05	0.19	0.61	0.15	4	2	1	3	B	D	D	C	B	C	2	3	3	1	2	1
23	B	0.12	0.62	0.17	0.09	3	1	2	4	B	A	D	B	D	C	1	3	4	1	4	2
24	C	0.29	0.24	0.19	0.28	2	4	1	3	C	A	C	B	C	C	1	2	1	4	1	1
25	B	0.18	0.79	0.03	0.00	2	1	3	4	B	D	B	B	B	B	1	4	1	1	1	1
26	C	0.03	0.01	0.94	0.02	2	4	1	3	C	A	C	C	C	C	1	2	1	1	1	1
27	B	0.11	0.54	0.32	0.03	3	1	2	4	B	A	B	C	B	B	1	3	1	2	1	1
28	A	0.49	0.32	0.13	0.07	1	2	3	4	A	B	A	B	A	A	1	2	1	2	1	1
29	A	0.72	0.11	0.03	0.13	1	3	4	2	B	C	D	D	A	A	3	4	2	2	1	1
30	A	0.59	0.12	0.26	0.03	1	3	2	4	C	D	B	B	B	A	2	4	3	3	3	1
31	B	0.06	0.4	0.14	0.41	4	1	3	2	A	C	D	B	B	B	4	3	2	1	1	1
32	B	0.09	0.57	0.32	0.03	3	1	2	4	D	A	D	C	B	B	4	3	4	2	1	1
33	D	0.18	0.20	0.30	0.33	4	3	2	1	D	D	B	B	D	D	1	1	3	3	1	1
34	B	0.39	0.5	0.03	0.08	2	1	4	3	C	A	D	B	B	B	4	2	3	1	1	1
35	A	0.21	0.34	0.27	0.18	1	2	3	4	B	D	C	C	C	C	2	4	3	3	3	3
36	B	0.14	0.22	0.18	0.46	4	1	3	2	D	C	C	D	D	B	2	3	3	2	2	1
37	B	0.35	0.38	0.18	0.08	2	1	3	4	B	B	A	B	B	B	1	1	2	1	1	1
38	A	0.32	0.25	0.26	0.16	1	3	2	4	A	B	A	D	D	A	1	3	1	4	4	1
39	C	0.09	0.12	0.40	0.39	4	3	1	2	C	C	A	A	C	C	1	1	4	4	1	1
40	A	0.41	0.38	0.06	0.15	1	2	4	3	A	B	C	A	B	A	1	2	4	1	2	1
41	B	0.47	0.41	0.09	0.03	2	1	3	4	B	D	C	A	B	B	1	4	3	2	1	1
42	B	0.07	0.66	0.14	0.13	4	1	2	3	A	C	C	B	B	B	4	2	2	1	1	1
43	B	0.29	0.59	0.07	0.05	2	1	3	4	B	D	B	B	A	B	1	4	1	1	2	1
44	D	0.09	0.15	0.35	0.41	4	3	2	1	C	C	D	C	D	D	2	2	1	2	1	1
45	A	0.45	0.12	0.37	0.06	1	3	2	4	C	C	C	A	A	A	2	2	2	1	1	1
46	D	0.19	0.32	0.2	0.29	4	2	3	1	A	C	C	A	B	D	4	3	3	4	2	1
47	B	0.2	0.34	0.14	0.32	3	1	4	2	A	D	D	D	C	A	3	2	2	2	4	3
48	A	0.35	0.1	0.47	0.08	1	3	2	4	A	C	A	D	B	A	1	2	1	4	3	1
49	C	0.14	0.22	0.57	0.06	3	2	1	4	C	D	A	C	C	C	1	4	3	1	1	1
50	A	0.39	0.26	0.23	0.12	1	2	3	4	D	B	B	B	A	A	4	2	2	2	1	1
51	D	0.16	0.24	0.21	0.38	4	2	3	1	A	B	B	D	B	D	4	2	2	1	2	1
52	B	0.02	0.81	0.04	0.13	4	1	3	2	B	D	C	B	B	B	1	2	3	1	1	1
53	B	0.38	0.29	0.08	0.25	2	1	4	3	B	C	A	A	B	A	1	4	2	2	1	2
54	B	0.34	0.29	0.23	0.14	2	1	3	4	A	B	B	A	B	C	2	1	1	2	1	3
55	C	0.5	0.15	0.23	0.13	2	3	1	4	C	C	A	A	D	C	1	1	2	2	4	1
Number correct score										30	10	20	30	40	50						

3. Compute an estimate of kappa: $\widehat{\kappa} = (.1818 - .2251)/(1 - .2251) = -.0559$.

4. Estimate the sampling variance of kappa under the null hypothesis from equation (9) with π_o and π_e replaced by p_e and p_0, respectively. Under the null hypothesis in equation (4), $\pi_o = \pi_e$, and

Table A2

Agreement Tables for the Five Copier-Source Pairs in the Computational Example

c-s_1

	1	2	3	4	Total
1	**7**	2	1	0	**10**
2	9	**3**	5	4	**21**
3	8	2	**0**	4	**14**
4	6	2	2	**0**	**10**
Total	**30**	**9**	**8**	**8**	**55**

c-s_4

	1	2	3	4	Total
1	**23**	5	6	6	**40**
2	2	**2**	0	2	**6**
3	1	2	**0**	0	**3**
4	4	0	2	**0**	**6**
Total	**30**	**9**	**8**	**8**	**55**

c-s_2

	1	2	3	4	Total
1	**14**	4	2	0	**20**
2	7	**1**	4	4	**16**
3	6	4	**1**	3	**14**
4	3	0	1	**1**	**5**
Total	**30**	**9**	**8**	**8**	**55**

c-s_5

	1	2	3	4	Total
1	**28**	7	7	8	**50**
2	2	**0**	0	0	**2**
3	0	2	**1**	0	**3**
4	0	0	0	**0**	**0**
Total	**30**	**9**	**8**	**8**	**55**

c-s_3

	1	2	3	4	Total
1	**17**	3	5	5	**30**
2	7	**3**	3	2	**15**
3	2	2	**0**	0	**4**
4	4	1	0	**1**	**6**
Total	**30**	**9**	**8**	**8**	**55**

Table A3

Summary of the Results for the Five Copier-Source Pairs in the Computational Example

Pair	p_o	p_e	$\hat{\kappa}$	$\sigma_{\hat{\kappa}}$	$z_{\hat{\kappa}}$	Decision
c-s_1	.1818	.2251	–.0559	.1066	–.5245	Reject H_1
c-s_2	.3091	.2962	.0183	.0946	.1938	Reject H_1
c-s_3	.3818	.3686	.0209	.0892	.2344	Reject H_1
c-s_4	.4545	.4383	.0288	.1027	.2808	Reject H_1
c-s_5	.5273	.5098	.0357	.1593	.2241	Reject H_1

the first term in the estimate of equation (9) simplifies to $p_e/(1-p_e) = .2905$. The second and third terms are equal to –.0943 and .42828, respectively. Thus, the sampling variance is estimated as $\sigma^2_{\hat{\kappa}} = (.2905 - .0943 + .4283)/55 = .01136$.

5. Compute the standardized kappa (equation (12)) under the null hypothesis: $z_{\hat{\kappa}} = -.0559/\sqrt{.01136} = -.5245$.

6. The critical value, z^*, follows from $\Pr(Z_{\hat{\kappa}} \geq z^*) = \alpha$ in equation (13). For a level of significance of $\alpha = .05$, it is equal to 1.645. Because $z_{\hat{\kappa}} = -.5245$ is much smaller, H_1 is rejected.

A summary of the results for all five copier-source pairs is shown in Table A3.

References

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley.

Bay, L. G. (1995, April). *Detection of cheating on multiple-choice examinations*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *46*, 443-459.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York: Springer.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

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

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

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large-sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323-327.

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

Hemker, B. T., Sijtsma, K., Molenaar, W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of total score. *Psychometrika*, *61*, 679-693.

Hemker, B. T., Sijtsma, K., Molenaar, W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331-347.

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-96-4). Princeton, NJ: Educational Testing Service.

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

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.

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.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer.

van der Linden, W. J., & Sotaridona, L. S. (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. S. (in press). Detecting answer copying when the regular response process follows a polytomous response model. *Journal of Educational and Behavioral Statistics*, *31*.

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

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

Acknowledgments

The article was written while W. J. van der Linden was a Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, CA. He is indebted to the Spencer Foundation for a grant awarded to the center to support his fellowship.

Author's Address

Address correspondence to Leonardo Sotaridona, CTB/McGraw-Hill, 7400 South Alton Court, Centennial, CO 80112; e-mail: leonardo_sotaridona@ctb.com.