DOCUMENT RESUME

ED 467 373 TM 034 299

AUTHOR	Sotaridona, Leonardo S.; Meijer, Rob R.
TITLE	Two New Statistics To Detect Answer Copying. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-01-07
PUB DATE	2001-00-00
NOTE	33p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Cheating; Multiple Choice Tests; *Responses; Sample Size; Simulation; *Statistical Analysis; Statistical Distributions; *Test Items
IDENTIFIERS	*K Index

ABSTRACT

Two new indices to detect answer copying on a multiple-choice test, S(1) and S(2) (subscripts), are proposed. The S(1) index is similar to the K-index (P. Holland, 1996) and the K-overscore(2), (K2) index (L. Sotaridona and R. Meijer, in press), but the distribution of the number of matching incorrect answers of the source (examinee s) and the copier (examinee c) is modeled by the Poisson distribution instead of the binomial distribution to improve the detection rate of K and K2. The S(2) index was proposed to overcome a limitation of the K and K2 indexes, namely their insensitiveness to correct answers copying. The S(2) index incorporates the matching correct answers in addition to the matching incorrect answers. A simulation study was conducted to investigate the usefulness of S(1) and S(2) for 40- and 80- item tests, 100 and 500 sample sizes, and 10%, 20%, 30% , and 40% answer copying. The Type I errors and detection rates of S(1) and S(2) were compared with those of the K2 and the w (omega) copying index (J. Wollack, 1997). Results show that all four indexes are able to maintain their Type I errors, with S(1) and K2 being slightly conservative compared to S(2) and w. Furthermore, S(1) had higher detection rates than K2. The S(2) index showed a significant improvement in detection rate compared to K and K2. (Contains 5 figures and 16 references.) (Author/SLD)

TM

ED 467 373

Two New Statistics to Detect Answer Copying

Research Report
01-07

Leonardo S. Sotaridona
Rob R. Meijer

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Two New Statistics to Detect Answer Copying

Leonardo S. Sotaridona

Rob R. Meijer

Abstract

Two new indices to detect answer copying on a multiple-choice test—S_1 and S_2—were proposed. The S_1 index is similar to the K-index (Holland, 1996) and the $\overline{K}_2$-index (Sotaridona & Meijer, in press) but the distribution of the number of matching incorrect answers of the source (examinee s) and the copier (examinee c) is modeled by the Poisson distribution instead of the binomial distribution to improve the detection rate of K and $\overline{K}_2$. The S_2 index was proposed to overcome a limitation of the K and $\overline{K}_2$ index, namely, their insensitiveness to correct answers copying. The S_2 index incorporates the matching correct answers in addition to the matching incorrect answers. A simulation study was conducted to investigate the usefulness of S_1 and S_2 for 40- and 80-item tests, 100 and 500 sample sizes, and 10%, 20%, 30% and 40% answer copying. The Type I errors and detection rates of S_1 and S_2 were compared with those of the $\overline{K}_2$ and the ω copying index (Wollack, 1997). Results showed that all four indices were able to maintain their Type I errors, with S_1 and $\overline{K}_2$ being slightly conservative compared to S_2 and ω. Furthermore, S_1 had higher detection rates than $\overline{K}_2$. The S_2 index showed a significant improvement in detection rate compared to K and $\overline{K}_2$.

Key Words: nominal response model, copying indices, cheating, Poisson distribution, loglinear model

Cheating on tests has a long and rich tradition (Cizek, 1999, Chap. 5). Among the cheating methods used are using forbidden materials, circumventing the testing process, or even using microrecorders. In the present study, we will be concerned with answer copying. In this type of cheating, one examinee copies the answers from another examinee, which may take place using all kinds of codes for transmitting answers and a code for doing so, for example, clicking of pens, tapping of the foot, and the like. Thus the examinees do not have to be in the physical neighborhood of each other. Because answer copying may invalidate an examinee's test score, it is necessary to prevent those practices by using well-instructed proctors and construct the seating arrangements so that there is ample room between the examinees. However, if a proctor observes some irregularities, statistical methods may be used to obtain additional evidence of answer copying.

Several methods have been proposed that all are based on determining the chance or likelihood that the observed score patterns of two examinees under suspicion are similar. A high likelihood may indicate answer copying. These chance methods can be classified into two types (Cizek, 1999, pp. 138-139). One type of method compares an observed pattern of responses to a known theoretical distribution (e.g., Frary, Tideman, & Watts, 1977; Wollack, 1997). In the second type of method, the probability of an observed pattern is compared with a distribution of values derived from independent pairs of examinees who took the same test. An example of such a statistic is the K-index (Holland, 1996).

Sotaridona and Meijer (in press) investigated the statistical properties of different forms of the K-index and compared the detection rate of these indices with the detection rate of the ω index (Wollack, 1997). The major difference between the indices is that the K-index does not assume any test model, whereas ω is based on item response theory modeling (e.g., van der Linden & Hambleton, 1997). Sotaridona and Meijer (in press) discussed that the K-index is less sensitive to answer copying when both the source and the copier have many matching correct answers. Lewis and Thayer (1998) and Sotaridona and Meijer (in press) found that the K-index that used the binomial distribution to model the matching incorrect answers had low power to detect substanstial amount of copying.

In this paper, we will propose an index S_2 that both takes the matching correct and the matching incorrect answers into account. Furthermore, we discuss an index S_1 which mathematical form is similar to the K-index (Holland, 1996) and the $\overline{K}_2$-index (Sotaridona & Meijer, in press) but the distribution of the number of matching incorrect answers of the source and the copier is modeled by the Poisson distribution.

This study is organized as follows. First, we introduce the K-index and the ω index. Second, we discuss two new indices S_1 and S_2 that may be used to obtain additional evidence of answer copying. Third, we conducted a simulation study to investigate the Type I error rate and detection rates of S_1 and S_2.

Existing Copying Indices

In this study, the copying indices ω (Wollack, 1997) and $\overline{K}_2$ (Sotaridona & Meijer, in press) are compared to the newly proposed copying indices, S_1 and S_2, with respect to the Type I errors and detection rates. A brief description of ω and $\overline{K}_2$ is given below followed by a more elaborate discussion of S_1 and S_2. The reader is referred to Sotaridona and Meijer (in press) and Wollack (1997) for a more detailed treatment of $\overline{K}_2$ and ω respectively.

The ω Index

Let examinee c, the copier, be suspected of copying answers from examinee s, the source. In a multiple-choice test with options $v = 1, 2, \ldots, k, \ldots, V$, let h_{cs} the number of items $i = 1, 2, \ldots, I$ where the response of c matches the response of s. Given that the response of s on i is k, let $P_{ik}(\theta_c)$ denotes the probability of c selecting the same option k on item i. Wollack (1997) used the nominal response model (Bock, 1972) to obtain this probability which is given by

$$P_{ik}(\theta_c)=\frac{\exp(\zeta_{ik} + \lambda_{ik}\theta_c)}{\sum\limits_{v=1}^{V} \exp(\zeta_{iv} + \lambda_{iv}\theta_c)}, \tag{1}$$

where ζ_{ik} and λ_{ik} are the intercept and slope parameters. The expected value of h_{cs} is computed, conditional on the ability level of the copier (θ_c), the item response vector of the source $\mathbf{U}_s = (U_{1s}, \ldots, U_{Is})$ where U_{is} is the response to item i, and the item parameters $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_I)$ with $\xi_i = (\zeta_{i1}, \ldots, \zeta_{iV}, \lambda_{i1}, \ldots, \lambda_{iV})$, as

$$E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi}) = \sum_{i=1}^{I} P_{ik}(\theta_c), \tag{2}$$

and the standard deviation of h_{cs} is

$$\sigma_{h_{cs}} = \sqrt{\sum_{i=1}^{I} [P_{ik}(\theta_c)]\,[1 - P_{ik}(\theta_c)]}. \tag{3}$$

The ω index is based on the residual between the observed and the expected value of h_{cs}. A standardized residual defines ω, which is asymptotically standard normally distributed (Wollack, 1997). The larger the value of ω, the stronger the evidence that c copied from s. The ω-statistic is given by

$$\omega = \frac{h_{cs} - E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi})}{\sigma_{h_{cs}}}. \tag{4}$$

The $\overline{K}_2$ Index

Define the number incorrect group $r = 1, 2, \ldots, c', \ldots, R$ such that examinees $j = 1, 2, \ldots, J_r$ have the same number of wrong answers, and c' indicate the group membership of c. The number of examinees in number incorrect group r is denoted by J_r so that $J_{c'}$ is the number of examinees with the same number of wrong answers as examinee c. Consequently, the two-letter index rj will be used to indicate an examinee j in number incorrect group r. Let U_{irj} be the response of examinee rj to item i and let W_s be the set of items, of size w_s, answered incorrectly by s.

For each examinee rj, an indicator variable A_{irj} equal to 1 if $U_{irj} = U_{is}$, and 0 otherwise. Note the item response of s is index by is indicating that s does not belong to any number incorrect group. The number of matching incorrect answers of rj and s, denoted by M_{rj} is then defined as

$$M_{rj} = \sum_{i \in W_s} A_{irj}. \tag{5}$$

For a particular $s - c$ pair, M_{rj} is observed for each examinee rj. For simplicity, M_{rj} will be denoted by M when it is not necessary to identify the examinee.

The $\overline{K}_2$ index is similar to the K index discussed by Holland (1996). For example, both indices are based on the random variable M, and are computed similarly as the sum of the binomial probabilities

$$\Pr(M_{c'c} \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} p^w (1-p)^{w_s - w}, \tag{6}$$

where $M_{c'c}$, with realization $m_{c'c}$, is the number of matching wrong answers between c and s, and p is the success probability parameter in the binomial distribution. The rationale behind the choice of the binomial distribution for M is discussed in Holland (1996). The main difference between K and $\overline{K}_2$ is the way p is estimated. For the K index, p is estimated by $\widehat{p} = \frac{\overline{M}_{c'c}}{w_s}$, where $\overline{M}_{c'c}$ is the mean of $M_{c'c}$ and w_s is the number of wrong answers of the source (Holland, 1996). For the $\overline{K}_2$ index, p is estimated by $\widehat{p}_2 = E(\beta_0 + \beta_1 Q_r + \beta_2 Q_r^2 + \varepsilon_r)$, where Q_r is the proportion of wrong answers of examinees in number incorrect group r. The parameters β_0, β_1, and β_2 are regression coefficients, and ε_r is an error term which is assumed to have a normal distribution with mean 0 and variance σ^2.

Note that the value of $\widehat{p}$ is obtained using data in the number incorrect group c' only, whereas the value of $\widehat{p}_2$ is obtained using all relevant information from R number incorrect groups. In this sense, $\widehat{p}_2$ contains more information than $\widehat{p}$, and therefore gives better estimate of p than $\widehat{p}$.

The $\overline{K}_2$ index is defined as

$$\overline{K}_2 = \Pr(M_{c'c} \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} \widehat{p}_2^w (1-\widehat{p}_2)^{w_s-w}. \tag{7}$$

The $\overline{K}_2$ index is an upper-tail probability. This probability can be compared to a chosen nominal level of significance α, such as 0.01. When it is less than or equal to this value, c may be identified as having a pattern of responses unusually similar to that of s.

Sotaridona and Meijer (in press) showed that the detection rates of the $\overline{K}_2$ index were in general higher than those of the K index, while ω yielded the highest detection rates. Furthermore, both $\overline{K}_2$ and ω were able to keep their empirical Type I errors below the nominal levels. The negative consequence of falsely identifying a noncopier as copier is severe, so we prefer an index that has Type I error at the nominal level or slightly below the nominal level.

Two New Indices

The S_1 Index

The S_1 index is similar to the $\overline{K}_2$ index in that it is also based on the random variable M. The S_1 index differs from the $\overline{K}_2$ index in the following ways. First for the $\overline{K}_2$ index, M is assumed to follow a binomial distribution whereas for S_1, M is assumed to follow a Poisson distribution. Secondly, the Poisson parameter μ is estimated using a loglinear model, whereas in the $\overline{K}_2$ index the binomial parameter p was estimated using a linear regression model. The motivation for proposing the Poisson distribution for M and the loglinear model for estimating μ as well as a statistic for checking the adequacy of the loglinear model are discussed, respectively, in the next three subsections. Once the estimate of μ for number incorrect group c', $\widehat{\mu}_{c'}$, is obtained, the S_1 index is computed as

$$S_1 = \sum_{w=m_{c'c}}^{w_s} \frac{e^{-\widehat{\mu}_{c'}} \widehat{\mu}_{c'}^{w}}{w!}. \tag{8}$$

$$S_1 = \Pr(M_{c'c} \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \frac{e^{-\widehat{\mu}_{c'}} \widehat{\mu}_{c'}^{w}}{w!},$$

Equation (8) is the probability of w_s being greater than $m_{c'c}$. The smaller the value of S_1, the stronger the evidence of answer copying.

The Choice for the Distribution of M

Several distributions have been assumed for the random variable M by previous researchers dealing with copying indices. Bay (1995) used the compound binomial distribution in developing the B_m copying index when all items in the item score pattern are considered. For the case where only the incorrect answers are considered, the ESA copying index (Belleza & Belleza, 1989), the K index (Holland, 1996), and the $\overline{K}_2$ index (Sotaridona & Meijer, in press) used the binomial distribution for M. Wollack (1997, p. 309) criticized the B_m and ESA indices for their inability to adjust the probabilities associated with an examinee's responses as a function of test score. Wollack (1997) found that B_m and ESA had lower detection rates compared to other indices based on classical test theory like the g_2 index (Frary et al., 1977). We did not include the g_2 index in this study since Wollack (1997) found that the Type I errors of g_2 are grossly inflated. Unlike g_2, the $\overline{K}_2$ index is able to control its Type I error below its nominal level.

Recall that the responses of the source to a set of test items are considered fixed and given these responses we count the number of wrong responses of the copier that matches that of the source and call it M. Since the binomial distribution did not yield high detection rates for K and $\overline{K}_2$ (Lewis & Thayer, 1998; Sotaridona & Meijer, in press), we propose the S_1 index that assumes a Poisson distribution as a reasonable approximation to the distribution of M. Hence, one may conceptualize S_1 as monitoring the rate or number of answer matches per incorrect answer by the source. If this rate is sufficiently high, then this provides evidence of answer copying. The extent to which the Poisson distribution approximates the distribution of M was investigated empirically.

Model for Estimating μ

To compute S_1 in Equation (8), we should know w_s, $m_{c'c}$, and μ. The value of w_s and $m_{c'c}$ are known whereas that of μ must be estimated. The mean of M differs across different ability levels. The values of M are small if most of the examinees have high ability level because the number of incorrectly answered items by the copier to match the wrongly answered items by the source is small. On the other hand, if most of the examinees have low ability level, the number of incorrectly answered items is large and the number of matched items is likely to be large. This information is taken into account in estimating μ by stratifying the examinees according to the number of wrong answers they obtained.

Since the Poisson distribution was assumed for M, it is standard practice to use the loglinear model to model the log of the mean of M (Agresti, 1996, p. 73). Using this model, it allows μ to be nonlinearly related to the predictor variable which in this case is the number of wrong answers. A study by Hanson (1994) revealed that the loglinear model is satisfactory for modeling M with M assumed a compound binomial distribution.

The relevant data for estimating μ are the number of wrong answers and the mean number of matching incorrect scores for each number incorrect group r. Let μ_r denote the expected value of the Poisson variate M_{rj}. The loglinear model has the form

$$\log(\mu_r) = \beta_0 + \beta_1 w_r, \forall\, r \tag{9}$$

where β_0 is the intercept term signifying the logarithm of the population mean across R number incorrect groups, and β_1 is the slope parameter. Estimation of β_0 and β_1 in Equation (9) is discussed in Agresti (1996, p. 93).

To obtain S_1, we need to determine the fitted mean for the number incorrect group to which the copier belongs. This fitted mean is

$$\widehat{\mu}_{c'} = e^{(\widehat{\beta}_0 + \widehat{\beta}_1 w_{c'})}.$$

Model Checking

The fit of the loglinear model in Equation (9) was investigated using the likelihood-ratio goodness-of-fit statistic, G^2, (Agresti, 1996, p. 89). The G^2 statistic can be used to test the null hypothesis that the model fits the data against the alternative that the model does not fit the data. Let $\widehat{\mu}_r$ be the fitted mean number of matching incorrect answers of number incorrect group r. The G^2 statistic is given by

$$G^2 = 2\sum_{r=1}^{R} \mu_r \log\left(\frac{\mu_r}{\widehat{\mu}_r}\right). \tag{10}$$

If the model perfectly fits the data, $\widehat{\mu}_r = \mu_r$. In such a case, $\log\left(\frac{\mu_r}{\widehat{\mu}_r}\right) = 0$ and consequently $G^2 = 0$. The distribution of G^2 is approximately chi-squared with degrees of freedom equal to R minus the number of model parameters. For the loglinear model in Equation (9), the number of model parameters is 2. The p-value to test the null hypothesis is the right-tail probability. Large values of G^2 or small p-values, for example, less than .01, would suggest a poor model fit (Agresti 1996, p. 89). If the fit of the model to the data is poor then it would not be appropriate to use Equation (8) as a statistical test of answer copying.

The S_2 Index

Copying indices that are based solely on the matching incorrect answers, such as the K and $\overline{K}_2$ indices, discard the additional information about copying that are available in the matching correct answers. By excluding the number of matching correct answers in the analysis of answer copying, we explicitly assume that c completely knows the answer to item i whenever c and s give a correct response to item i. However, this is not always the case. An examinee may obtain the correct answer to an item by copying or by guessing.

Note that the K and $\overline{K}_2$ indices are not sensitive to a copier who is copying only the correct answers of the source. This may be the case when s and c are friends and s shares his or her answers to c on items where he or she is almost sure of the correct answers. Another example is a high-stakes examination where c may bribe s for sharing his correctly answered items to c.

The new copying index S_2 is proposed to overcome this limitation. We propose S_2 to incorporate information about copying that are contained in the matching correct answers in addition to the information in the matching incorrect answers. Note that as used in K and $\overline{K}_2$, the evidence of answer copying is 1 if s and c choose the same wrong option to an item, and 0 if they are both correct or their response to an item did not match. For S_2, however, the evidence is 1 if s and c choose the same wrong option to an item, δ (to be described below) if the source and the copier are both correct, and 0 otherwise. The variable δ quantifies the amount of correct-answer copying information to an item for a particular source-copier pair.

Let i^* denotes an item that was answered correctly by s, and U_{i^*rj} the response of examinee rj to item i^*. Then, δ_{i^*rj} gives the estimate of copying information on item i^* by examinee rj. The value of δ_{i^*rj} satisfies the inequality

$$1 \geq \delta_{i^*rj} \geq 0,$$

that is, $\delta_{i^*rj} = 0$ if rj knows the correct answer to item i^* and $\delta_{i^*rj} = 1$ if rj is completely ignorant about the correct answer to item i^* (see conditions 1-2 below). The problem is to quantify the amount of knowledge that rj has on i^*. To do this we have to obtain the probability of rj answering item i^* correctly. This probability can be estimated as the proportion of examinees in number incorrect group r getting the correct answer to item i^*. A drawback of this approach is that the estimate is highly dependent on the population of examinees taking the test. For example, the estimate would tend to be low if the population of examinees are of high ability level while the estimate would tend to be high if the population of examinees are of low ability level. A sensible solution is to condition on the ability level of the suspected copier. For the rest of the presentation, unless specified otherwise, j will refer to an examinee that belongs to certain number incorrect group.

Let P_{i^*rj} denotes the probability of rj getting the correct answer to item i^*, and A_{i^*rj} an indicator variable equal to 1 if $U_{i^*rj} = u_{i^*s}$, and 0 otherwise. Note that P_{i^*rj} is a conditional probability, not to be mistaken as a joint probability that s and rj will give a common response to item i^*. Given U_{i^*s}, this probability is

$$P_{i^*rj} = \Pr\left(U_{i^*rj} = u_{i^*s} \mid U_{i^*s}\right), \tag{11}$$

and the maximum likelihood estimate of P_{i^*rj} is

$$\widehat{P}_{i^*rj} = \frac{\sum_{j=1}^{J_r} A_{i^*rj}}{J_r}. \tag{12}$$

Given the estimate of P_{i^*rj}, what remains is to transform this estimate into δ_{i^*rj}. A suitable transformation function, $f(P_{i^*rj})$, satisfy the following conditions:

1. $f(P_{i^*rj})$ approaches 0 as P_{i^*rj} approaches 1; that is, the evidence of answer copying deminishes as P_{i^*rj} approaches 1.
2. $f(P_{i^*rj})$ approaches 1 as P_{i^*rj} approaches 0; that is, the evidence of answer copying approaches 1 if the suspected copier is correct to an item despite low probability of getting the correct answer to such an item.
3. Test with different number of options must have different weight function. Let f and f' be two different weight functions and i^* and $i^{*\prime}$ are items taken from two tests with number of options V and V' such that $V < V'$. Then it holds that $f(P_{i^*rj}) > f'(P_{i^{*\prime}rj})$ whenever $P_{i^*rj} = P_{i^{*\prime}rj}$.

The basis for conditions 1-2 should be clear from the above discussions. Condition 3 arises from the idea that multiple-choice tests with different number of options should have different transformation functions that differ by a factor that is a function of the number of options. This calls for a function that account for the probability of guessing to an item as a scaling factor.

For notational convenience, let g denotes the probability of getting the correct answer to item i by guessing. Note, an often used value of g is 0.20 for a 5-option test and 0.25 for a 4-option test. A sensible function satisfying conditions 1-3 is shown in Equation 13.

$$\delta_{i^*rj} = f(P_{i^*rj}) = (d_1 e)^{d_2 P_{i^*rj}}, \tag{13}$$

where

$$d_2 = -\left(\frac{1+g}{g}\right) \text{ and } d_1 = \left(\frac{1+g}{1-g}\right)^{d_2 P_{i^*rj}}.$$

Equation (13) is a monotone decreasing function of P_{i^*rj} with g a scaling constant. Figure 1 shows the graph of Equation (13) with $g = .2$ (denoted as $F1-$ 5 options) and $g = .25$ (denoted as $F2-$ 4 options). As shown in the graph, the value of δ_{i^*rj} for both F_1 and F_2 approaches zero as P_{i^*rj} approaches 1 and δ_{i^*rj} approaches 1 as P_{i^*rj} approaches zero (conditions 1-2). Furthermore, $F1(P_{i^*rj}) < F2(P_{i^*rj})$ for $P_{i^*rj} \in (0, 1]$ (condition 3).

[Insert Figure 1 about here]

Let M^*_{rj} denotes the sum of the number of matching incorrect answers and weighted matching correct answers by examinee rj and examinee s. The expression for M^*_{rj} is given by

$$M^*_{rj} = M_{rj} + \sum_{i^*} \delta_{i^*rj}. \tag{14}$$

In Equation (14), the contribution of each item to the value of M^*_{rj} is 0 if the response of rj did not match that of s, 1 if the wrong response of rj matches that of s, and δ_{i^*rj} if the correct response of rj matches that of s. The value of M^*_{rj} would be large if most of the incorrect responses of rj matches the wrong responses of s or if P_{i^*rj} is small and most of the correct responses of rj matches the correct responses of s. The larger the value of M^*_{rj} relative to the number of items, the stronger the evidence of answer copying.

Note that if there are no matching correct answers between s and rj, the second term in Equation (14) sum up to zero and $M^*_{rj} = M_{rj}$. Hence, M_{rj} becomes a special case of M^*_{rj}. On the other hand, if there are no matching incorrect items but only matching correct answers, then $M_{rj} = 0$ and $M^*_{rj} = \sum_{i^*} \delta_{i^*rj}$. Thus, while M_{rj} is only sensitive to incorrect answer copying, M^*_{rj} is sensitive to both correct and incorrect answer copying.

In reality, the random variable M^*_{rj} is a nonnegative real-valued random variable. We treat M^*_{rj} as an integer by rounding it off to the nearest integer. Although some error is introduced by doing this, we expect that this will only have a minor influence on the effectiveness of the statistic. Like M_{rj}, we used the Poisson distribution for M^*_{rj} and the loglinear model to estimate its mean. We explored empirically the usefulness of the Poisson distribution to model M^*_{rj} using the G^2 statistics.

The S_2 index is then defined as

$$S_2 = \sum_{w=m^*_{c'c}}^{I} \frac{e^{-\widehat{\mu}_{c'}} \widehat{\mu}_{c'}^{w}}{w!}, \tag{15}$$

$$S_2 = \Pr(M^*_{c'c} \geq m^*_{c'c}) = \sum_{w=m^*_{c'c}}^{I} \frac{e^{-\widehat{\mu}_{c'}} \widehat{\mu}_{c'}^{w}}{w!},$$

where $M^*_{c'c}$, with realization $m^*_{c'c}$, is the sum of the number of matching incorrect and weighted matching correct answers between c and s. The smaller the value of S_2, the more likely that answer copying occurred.

Method

Data Generation and Simulation of Copying

The data were simulated in the same way as in Sotaridona and Meijer (in press). Multiple choice test items with five options were considered with test lengths 40 and 80 items and samples of 100 and 500 simulees were generated. Item parameters were chosen in accordance with the study by Wollack (1997). As described in Wollack (1997), the item parameters were estimated under the nominal response model using MULTILOG (Thissen, 1991) for an 80-item, 5-alternative English college placement test and a 40-item, 5-alternative mathematics college placement test used at a large Midwestern research university. We draw the ability parameter, θ, from $N(0, 1)$. Given the item and ability parameters, $P_{iv}(\theta)$ was computed based on the nominal response model. The item response was drawn randomly from $v = [1, 2, \ldots, V]$, each having probability of being drawn equal to $P_{i1}(\theta), P_{i2}(\theta), \ldots, P_{iV}(\theta)$ respectively. Then the source was drawn at random from a sample of simulees having ability percentile rank ranging from 40 to 90. In both 40- and 80-item tests, five percent copiers were selected randomly from the simulees with θ level below the θ level of the source. The percentage of items copied were 10, 20, 30, and 40.

We crossed the three factors – sample size (2 levels), number of items (2 levels), and percentage of items copied (4 levels)–resulting in $2 \times 2 \times 4 = 16$ testing conditions. The dataset in each condition was replicated 100 times.

Similar to Wollack (1997), copying was simulated by first randomly selecting a specified percentage of items from the copier and then altering the responses of c to match the responses of s on those items.

Type I Error and Detection Rates

A simulee was identified as a copier by $\overline{K}_2$, S_1, or S_2 index if the values were less than or equal to the level of significance α. The α levels were set to .0001, .0005, .001, .005, and .01; similar α levels used in Wollack (1997) with the exclusion of .05, .10 and .0025.

For the ω statistic, a simulee was identified as a copier when the value of ω was above the one-tailed critical value corresponding to the right tail of the standard normal curve. The ω was computed using the item and ability parameters that were used in the simulation. This was done partly for convenience and partly because Wollack & Cohen (1998) showed that the Type I error rate of ω is unaffected by estimating the item and ability parameters. As in Sotaridona and Meijer (in press), the copying indices were computed based on prior suspicion of a particular simulee copying from a specific source. Hence, the statistics were tested for significance without adjustment for the α level.

To determine the empirical Type I error rate, we computed the proportion of noncopier simulees who were identified by the copying index as copiers. This computation was based on 9400 non copiers (94 noncopiers per replication$\times$100 replications), for datasets with 100 examinees, and 47400 non copiers (474 noncopiers per replication$\times$100 replications), for datasets with 500 examinees.

Likewise, the detection rate was obtained by taking the proportion of true copier simulees who where classified as copier by $\overline{K}_2$, S_1, S_2, and ω. This computation was based on 500 true copiers (5 true copiers per replication$\times$100 replications), for datasets with 100 examinees, and 2500 true copiers (25 true copiers per replication$\times$100 replications), for datasets with 500 examinees. Ideally, we want an index which minimizes the Type I error rate and maximizes the detection rate.

Results

Adequacy of the Loglinear Model

The fit of the loglinear model given in Equation (9) was assessed using the G^2 statistic. The results are similar for M and M^* and also between 40$-$ and 80$-$item test so only the results for M^* with 40$-$item test are presented and discussed here.

Figure 2 shows the scatter plots of 100 p-values (x-axis) by rank (y-axis) for 40-item test with 100 and 500 simulees and for different percentages of items copied. Remember that the null hypothesis being tested is that the loglinear model fits the data; large p-values therefore supports the null hypothesis.

[Insert Figure 2 about here]

The loglinear model fits the data very well in every situation simulated as reflected by the high p-values both for $J = 100$ and $J = 500$. For example, at $J = 100$, the minimum p-value for 10% copying is 0.332, 0.418 for 20% copying, 0.182 for 30% copying, and 0.481 for 40% copying (Figure 2 a-d), whereas at $J = 500$, all the p-values are nearly 1 across four percentages of copying (Figure 2 e-h). The fit of the model are quite similar for different percentages of copying.

Type I Error Rate

Figure 3 shows the Empirical Type I error of ω, $\overline{K}_2$ (denoted as $K2$), S_1, and S_2 (denoted as $S1$ and $S2$) for different α-levels and across combinations of sample sizes and test lengths. The solid line in the graph is a boundary line indicating perfect agreement between the nominal and empirical Type I errors. A copying index having Type I errors that is above the boundary line is liberal in classifying the simulee as copier and below the boundary is conservative. An ideal copying index maintains its Type I error on or slightly below the nominal α level, but not too far below; otherwise, its detection rate will be reduced.

[Insert Figure 3 about here]

The S_2 index holds its Type I error for $J = 100$ (Figures 3a-b) and tend to be slightly liberal for $J = 500$ (Figure 3 c-d). The ω index on the other hand is slightly liberal for $J = 100$ and slightly conservative for $J = 500$. Both the S_1 and $\overline{K}_2$ were able to hold their Type I errors below the nominal levels for which, in most cases, lower than the Type I errors of S_2 and ω. The most conservative index for $J = 100$ was S_1 and for $J = 500$ was $\overline{K}_2$.

Detection Rate

The detection rates for $\overline{K}_2$, S_1, S_2, (denoted $K2$, $S1$ and $S2$, respectively), and ω for different α-levels, percentages of copying, and test lengths are shown in Figure 4 for 100 simulees and Figure 5 for 500 simulees. The detection rates for all the indices increased with percentage of copying. For example for 40 items and 100 simulees, the detection rates in Figure 4a (40% copying) are higher than the detection rates in Figure 4b (30% copying) which are both higher than that in Figure 4c (20% copying) and Figure 4d (10% copying). Similar trends were observed for other combinations of sample sizes and test lengths.

[Insert Figures 4 and 5 about here]

Consistent with the findings of Wollack (1997) and Sotaridona and Meijer (in press), the detection rate of ω increased with test length but not with sample size. For example, Figure 4 shows that for 100 examinees, the detection rate of ω was higher for the 80-item test than for the 40-item test. The same observation was noted for 500 examinees (Figure 5). For a fixed test length, changing the sample size from 100 to 500 did not change the detection rate of ω (compare Figure 4 with Figure 5).

On the other hand, the test length, the sample size, or a combination of both test length and sample size affect the detection rates of $\overline{K}_2$, S_1 and S_2. In particular, increasing the test length (compare Figure 4 with Figure 5) or sample size (compare Figure 4 a-d with Figure 4 e-h, and Figure 5 a-d with Figure 5 e-h) resulted in increased detection rates. What follows, we compare the detection rates for the four indices.We should keep in mind, however, that the empirical Type I errors are not exactly similar, though the differences are small.

None of the index is best in all testing conditions considered. The S_2 index outperformed the other indices if the amount of copying are 20% and 10% regardless of test length and sample size (Figure 4 c-d, g-h and Figure 5 c-d, g-h), or if the amount of copying is 30% or 40% with 40 items and 500 simulees (see Figure 5 a-b).

Furthermore, with 30% or 40% copying, the S_2 index approximately equalled the ω index on 40 items and 100 simulees (Figure 4 a-b) and on 80 items and 500 simulees (Figure 5 e-f), whereas the ω index has the highest detection rates on 80 items and 100 simulees (Figure 4 e-f). The four indices are equally effective with nearly 100% detection rates (for $\alpha = .01$, 40% copying, and $I = 80$; see Figures 4e and 5e). In general, the $\overline{K}_2$ index had the lowest detection rate compared to the other indices.

Discussion

We proposed the S_1 index to detect answer copying as an alternative to the $\overline{K}_2$ index. In the S_1 index the Poisson distribution was used instead of the binomial distribution for M. The S_2 index was also proposed to overcome the limitation of $\overline{K}_2$ and subsequently that of S_1 that are not sensitive to answer copying the correct item scores. Crucial in the application of S_1 and S_2 is obtaining reliable estimates of the means of M and M^*. We approached this concern by using the loglinear model. We evaluated the fit of the loglinear model using the G^2 statistic. The Type I errors and detection rates of S_1 and S_2 were compared with the Type I errors and detection rates of $\overline{K}_2$ and ω.

The results did not provide convincing evidence against using the Poisson distribution for M and M^*. In particular, using the Poisson distribution, instead of the binomial distribution, resulted in S_1 having detection rates considerably higher than that of $\overline{K}_2$. The S_2 index, which incorporates information from the matching correct scores in addition to the matching incorrect–scores, lead to a significant improvement in detection rate of S_1. In general, a copying index is not sensitive when only few item scores are copied. This initial study reveals that S_2 showed noticeable improvement over the best copying index ω if the amount of copying is 20% of the total number of items or less.

As shown in this study and in Sotaridona and Meijer (in press), if the item parameters in the nominal response model can be estimated reliably, ω seems to be the best choice for detecting answer copying because it is sensitive across all ability levels of the copier and can also be used to detect answer copying on examinations with only the source and the copier as examinees; S_1 and S_2 cannot be used in this latter case. However, considering the computational simplicity and less restrictive assumptions imposed on S_2, the S_2 index may be a good alternative to use in practice.

Results concerning the Type I errors of ω and $\overline{K}_2$ were also consistent with the result of previous study (Sotaridona & Meijer, in press) which showed that the ω is sligtly liberal at $J = 100$ whereas $\overline{K}_2$ is conservative. In general, although the empirical Type I errors of the four indices are not perfectly in agreement with the nominal Type I errors, the deviations are small.

The present study only considered five percent copiers and the items copied by the copiers were selected at random. There is some indication that the magnitude of the difference in the ability level of the source and the copier affects the performance of a copying index. For future research, it might be interesting to study the Type I errors and detection rates of S_1 and S_2 for varying mode of answer copying and for different concentrations of copiers, percentages of correct answers copying, and various ability level of the source.

References

Agresti, A. (1996). *An introduction to categorical data analysis.* NY: Wiley.

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

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

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

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

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

Hanson, B. A. (1994). *Statistical Indexes of Response Similarity Derived from the Compound Binomial Distribution.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Holland, P. W. (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.

Kotz, S. & Johnson, N. L. (1986). *Encyclopedia of statistical science*, 7, NY: Wiley.

Lewis, C. & Thayer, D. T. (1998). The power of the k-index (or PMIR) to detect copying (*ETS Research Report No. 98-49*). Princeton, NJ: Educational Testing Service.

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

Thissen, D. (1991). MULTILOG *user's guide* (Ver. 6). Chicago: Scientific Software, Inc.

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). NY: Springer-Verlag.

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

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.

Figure Captions

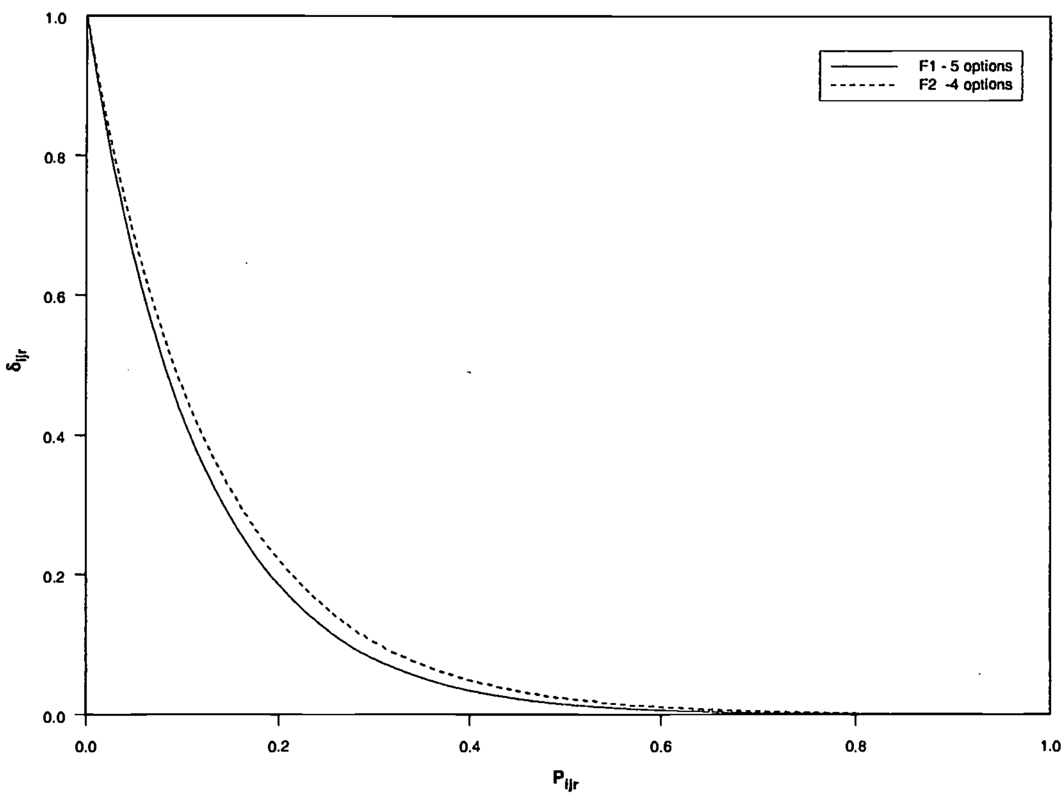
Figure 1. Graph of δ_{i^*jr} as a Function of P_{i^*jr} with $p_g = 0.25$ and $p_g = 0.20$.

Figure 2. Scatter Plots of 100 p-values of G^2 Statistics, Ranked in Increasing Order, for 40-Item Test

Figure 3. Nominal and Empirical Type I Error Rate as a Function of Simulee Size and Test Length

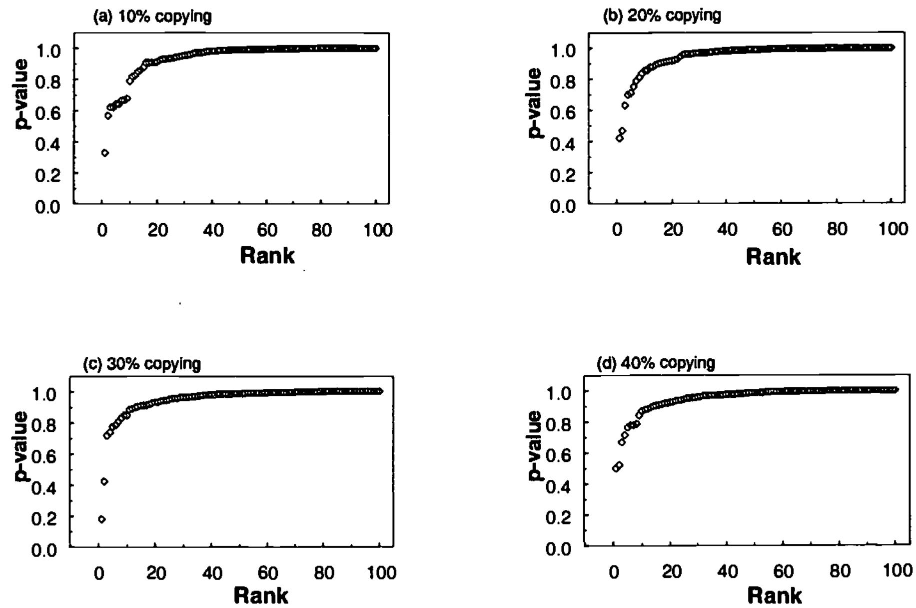
Figure 4. Detection Rate of ω, $\overline{K}_2$, S_1, and S_2 on Test with 100 Simulees

Figure 5. Detection Rate of ω, $\overline{K}_2$, S_1, and S_2 on Test with 500 Simulees

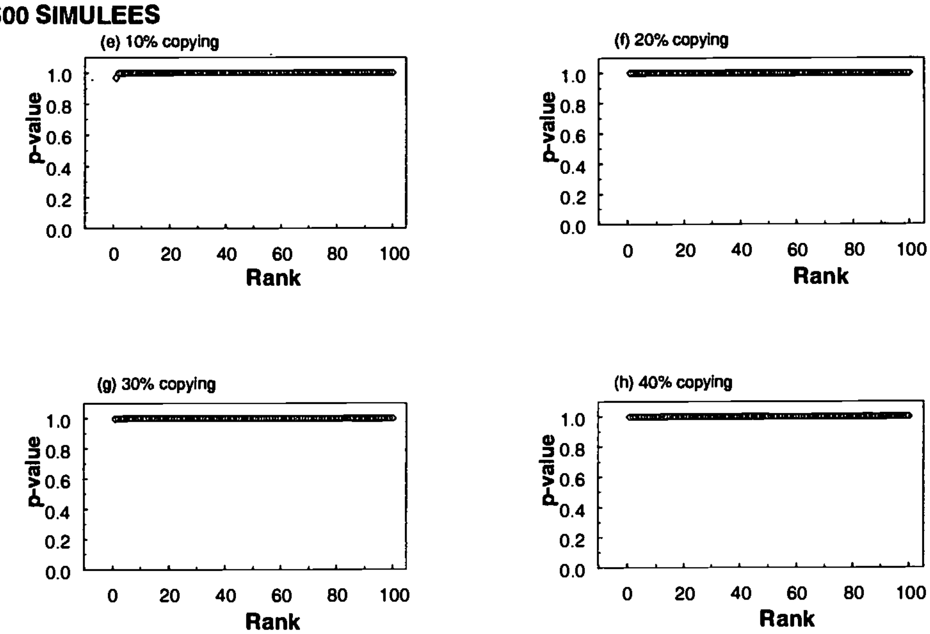


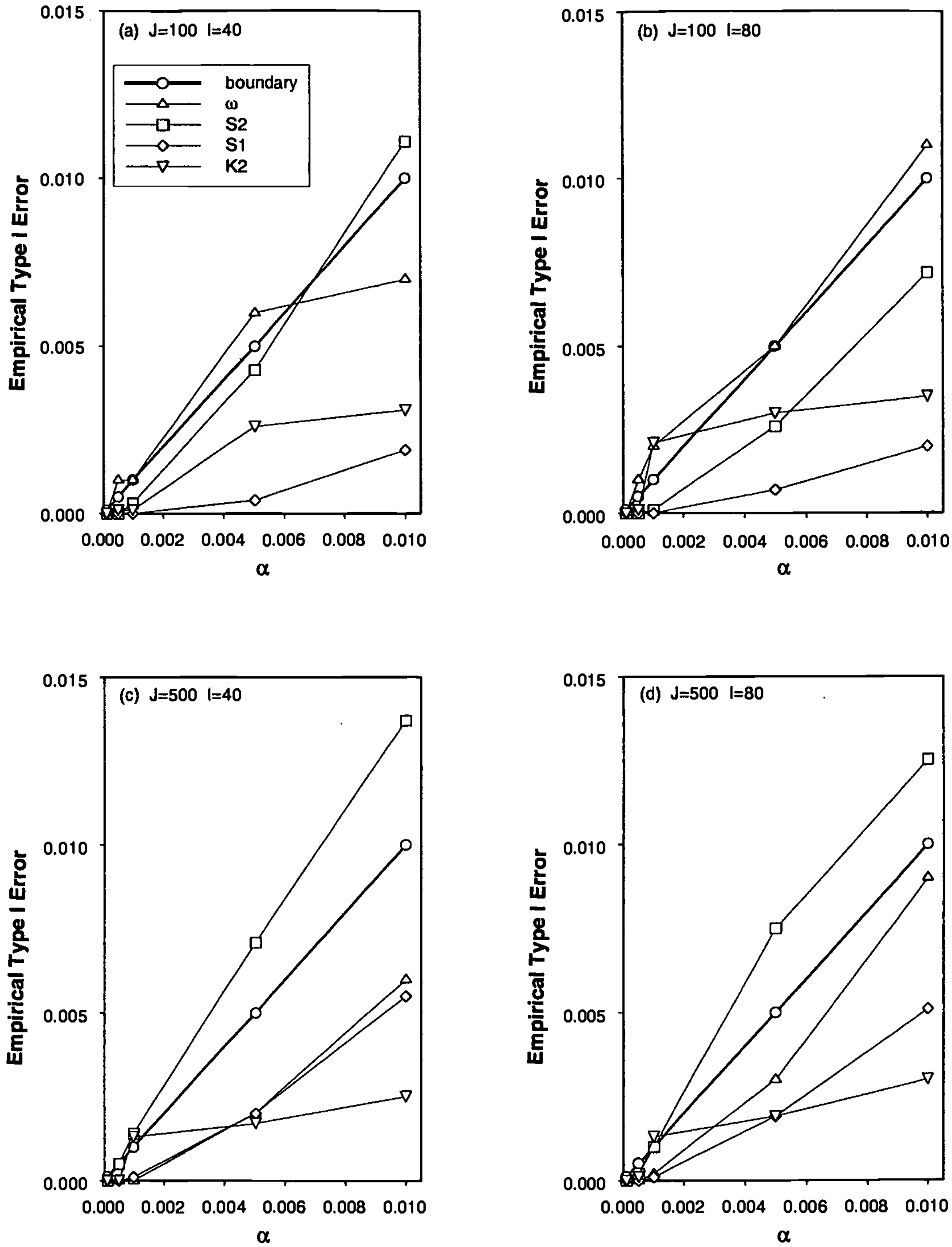
BEST COPY AVAILABLE

100 SIMULEES

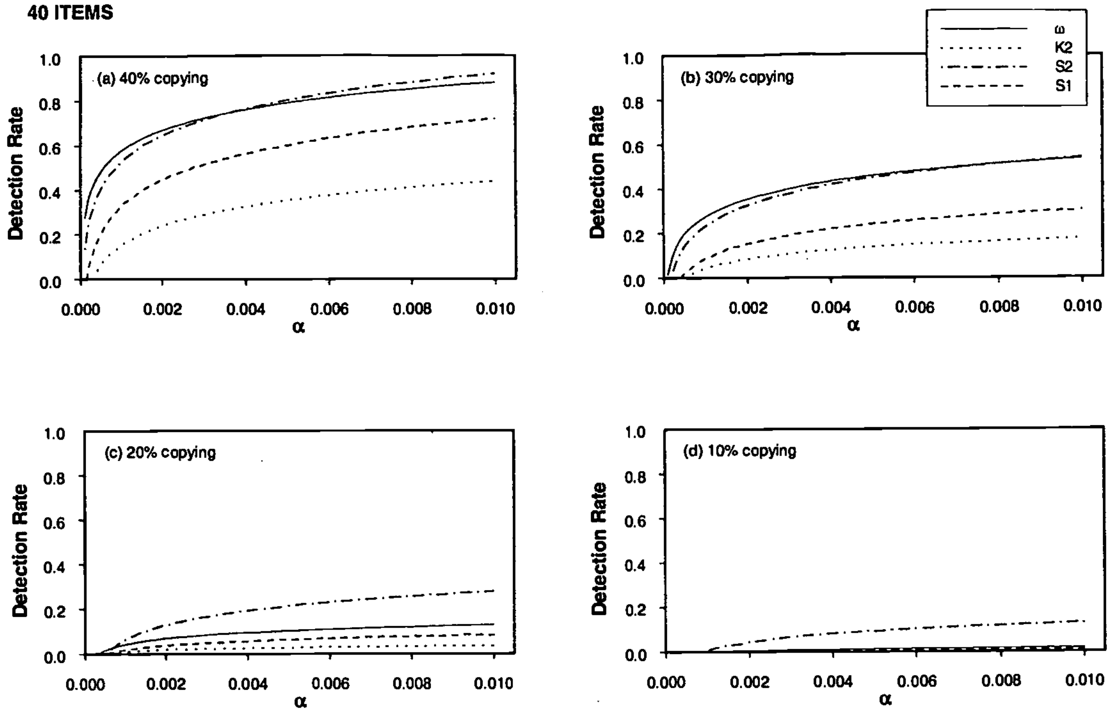


500 SIMULEES

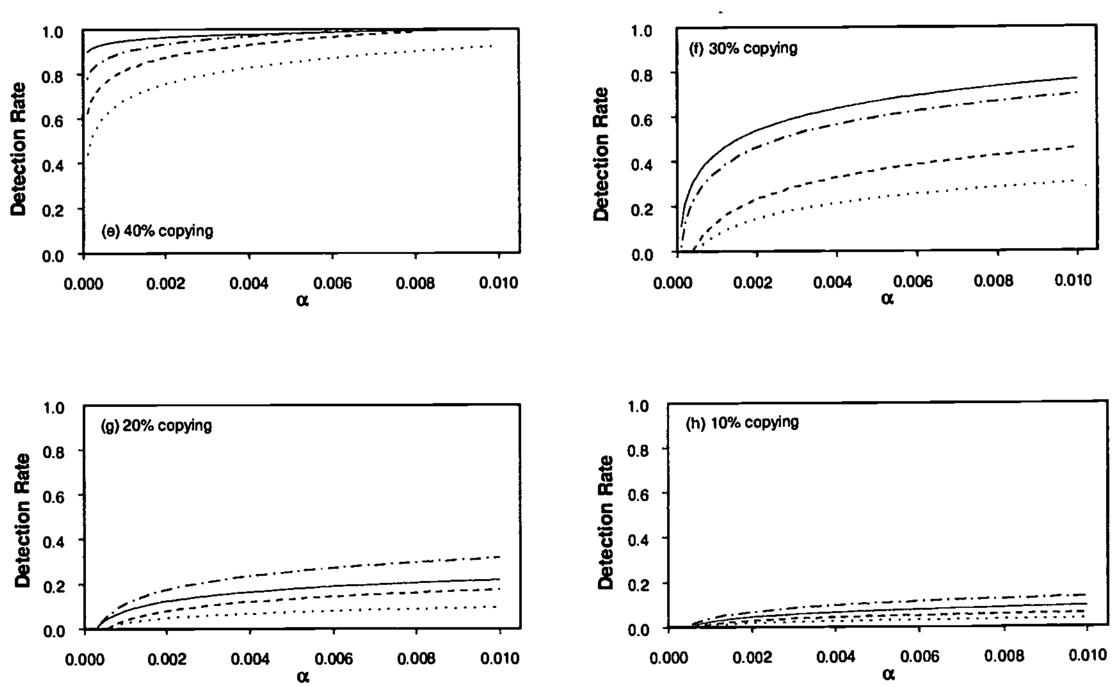



(a) J=100 I=40
(b) J=100 I=80
(c) J=500 I=40
(d) J=500 I=80
boundary
ω
S2
S1
K2
Empirical Type I Error
α
0.000
0.005
0.010
0.015
0.000 0.002 0.004 0.006 0.008 0.010

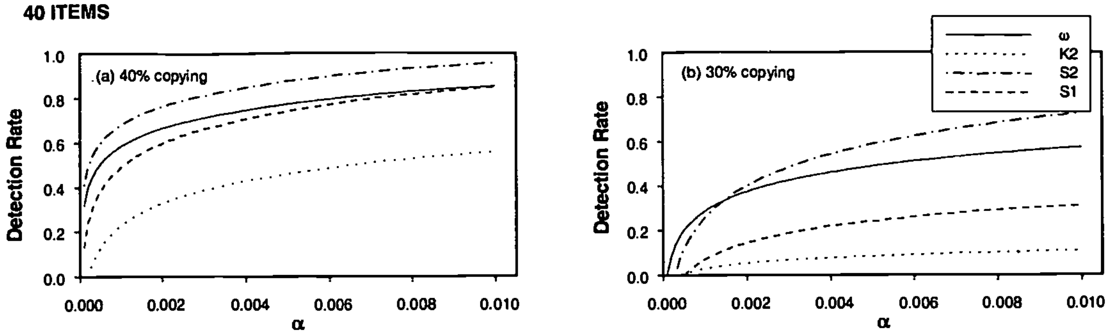
40 ITEMS



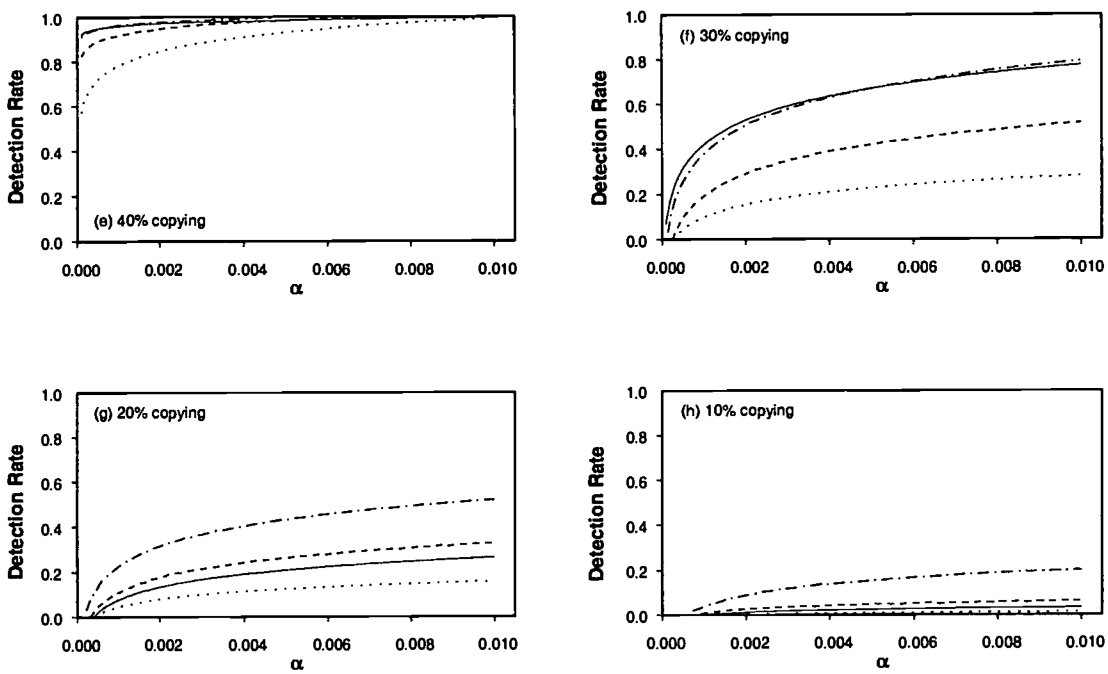
80 ITEMS



40 ITEMS



80 ITEMS



Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-01-07	L.S. Sotaridona & R.R. Meijer, *Two New Statistics to Detect Answer Copying*
RR-01-06	L.S. Sotaridona & R.R. Meijer, *Statistical Properties of the K-index for Detecting Answer Copying*
RR-01-05	I. Hendrawan, C.A.W. Glas, & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*
RR-01-04	R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*
RR-01-03	R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*
RR-01-02	R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*
RR-01-01	W.J. van der Linden & H. Chang, *Implementing Content Constraints in Alpha-Stratified Adaptive testing Using a Shadow test Approach*
RR-00-11	B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*
RR-00-10	W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score equating*
RR-00-09	W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*
RR-00-08	L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*
RR-00-07	W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*
RR-00-06	C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*
RR-00-05	B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*
RR-00-04	B.P. Veldkamp, *Constrained Multidimensional Test Assembly*
RR-00-03	J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*
RR-00-02	J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*
RR-00-01	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

NOTICE

Reproduction Basis

X This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").