

Detecting Answer Copying When the Regular Response Process Follows a Known Response Model

Wim J. van der Linden
Leonardo Sotaridona
University of Twente

A statistical test for detecting answer copying on multiple-choice items is presented. The test is based on the exact null distribution of the number of random matches between two test takers under the assumption that the response process follows a known response model. The null distribution can easily be generalized to the family of distributions of the number of random matches under the alternative hypothesis of answer copying. It is shown how this information can be used to calculate such features as the maximum, minimum, and expected values of the power function of the test. For the case of the nominal response model, the test is an alternative to the one based on statistic ω . The differences between the two tests are discussed and illustrated using empirical results.

Keywords: answer copying, cheating, generalized binomial distribution, item response models, nominal response model

All existing statistical tests for detecting answer copying on multiple-choice questions (e.g., Angoff, 1974; Frary, Tideman, & Watts, 1977; Holland, 1996; Sotaridona & Meijer, 2002; van der Linden & Sotaridona, 2004; Wollack, 1997) are based on a statistic defined as the number of matching responses between the test taker suspected of copying (the “copier”) and the test taker whose answers may have been copied (the “source”). However, they do differ as to the set of items on which the statistic is defined, whether or not it is standardized, as well as the null distribution it is postulated to have. More specifically, these differences are as follows.

First, as for the set of items on which the statistic is defined, the K index (Holland, 1996; Lewis & Thayer, 1998) is defined on the number of matching alternatives between the source and the copier on the set of items the source had incorrect. Its sampling distribution is conditional on the actual incorrect responses of the source and the number of incorrect responses of the copier. The same type of conditioning is used in Sotaridona and Meijer (2002) and van der Linden and Sotaridona (2004). For the test based on the assumption of the knowledge-copying-or-random-guessing

The article was written while the first author was a Fellow at 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. The authors are indebted to Charles Lewis for putting them on the right track as to the issue of conditioning in statistical tests of cheating.

model in van der Linden and Sotaridona, the conditioning is even necessary. As these authors show, under this assumption, including the items the source has correct in the definition of the test statistic would lead to the testing of confounded hypotheses. All other statistical tests referred to in the introductory paragraph have statistics defined on the full set of items in the test. The issue of conditioning in statistical tests is rather delicate. We return to it later in this article.

Second, the g_2 (Frery et al., 1977) and ω statistic (Wollack, 1997; see also Wollack & Cohen, 1998) are standardized versions of the number of matching alternatives. Both statistics derive the mean and standard deviation needed for standardization from the probabilities with which the copier selects his/her alternatives. For the g_2 statistic these probabilities are estimated from the proportions of test takers selecting the alternatives in the population of test takers corrected by the copier's number-correct score and the average number-correct score in the population. Statistic ω assumes the fit of the nominal response model (Bock, 1972, 1997) to the regular response process and uses estimates of the parameters in the model to derive the probabilities. All other known test statistics are defined as the unstandardized number of matching incorrect alternatives between the copier and source.

Third, the most critical differences between the statistical tests, however, are their postulated null distributions. The K index and the tests in Sotaridona and Meijer (2002) and van der Linden and Sotaridona (2004) have null distributions related to the binomial family. The null distribution of the K index is a parametric binomial with a success parameter obtained by piecewise linear regression of the proportion of matching incorrect alternatives on the proportion of incorrect responses in the population of test takers. The test by Sotaridona and Meijer is based on the same type of null distribution but with quadratic instead of piecewise linear regression. The test in van der Linden and Sotaridona has a null distribution that is a binomial with a shifted support, where the shift is determined by the number of items the copier knows.

On the other hand, the tests based on the g_2 and the ω statistic have a null distribution that is postulated to be normal. For the ω statistic the postulate is based on the central limit theorem (Wollack, 1997, p. 331). Generally, we have to be careful to invoke such theorems when we have a case of independent but nonidentically distributed variables, as in the current application. However, for the binary variables to be defined in Equation 4 the conditions in the central limit theorem for non-identical variables, known as the Liapunov theorem (see, e.g. Lehmann, 1999, sect. 2.7), hold, and large-sample normality for the number of matching alternatives is guaranteed.

In this article, we derive the exact null distribution for the number of matching response alternatives under the assumption of a known polytomous item response theory (IRT) model. Unlike the ω statistic, which only holds for increasing test length, the statistical test based on this distribution can be applied to multiple-choice tests of any length. An additional advantage is the test applies to a response model with any response format. Moreover, it seems to be easy to generalize the null distribution to the family of distributions under the alternative hypothesis of

copying. We show how these alternative distributions can be used to analyze the power function of a statistical test based on the number of matching alternatives.

Hypotheses and Test Statistic

Let $i = 1, \dots, n$ denote the items in the test, each with response options $a = 1, \dots, k$. We use index j for a test taker checked on copying answers from source s and random variables U_{ji} with possible values $1, \dots, k$ to represent the response of j to i .

Assumptions

Throughout this article, we assume that the probabilities with which a test taker who has not copied any answers chooses a response alternative follow a known response model. In IRT, a variety of dichotomous and polytomous response models is available (for introductions to these models, see van der Linden & Hambleton, 1997), and the statistical test proposed in this article holds for any choice of model. Whatever model is chosen, however, it is most critical that it has satisfactory fit to the response data. We assume that the items have been calibrated with enough precision to consider their values for the item parameters as known. As for the person, for ease of exposition, we develop our theory assuming that the values of test takers j and s for the parameters θ_j and θ_s are known (see end of this section). For a test of answer copying when the response model is unknown, see Sotaridona, van der Linden, and Meijer (2006).

In the section with numerical results below, we temporarily make the additional assumption that the responses follow the nominal response model. This enables us to make an empirical comparison between the results for the test in this article and the one based on the ω statistic. The nominal response model gives the probability of a test taker with ability level θ choosing alternative a of item i as

$$\pi_{i_a}(\theta) = \frac{\exp(\zeta_{i_a} + \lambda_{i_a} \theta)}{\sum_{a=1}^k \exp(\zeta_{i_a} + \lambda_{i_a} \theta)}, \quad (1)$$

where ζ_{i_a} and λ_{i_a} are the intercept and slope parameters for alternative a on item i (Bock, 1972, 1997). The model can be estimated by using the computer program Multilog (Thissen, 1991). For a study on the (minor) effect of estimation error in the parameter values in this model on the properties of ω test, see Wollack and Cohen (1998). We do not expect the effects for the statistic in this article to differ much from Wollack and Cohen's findings.

Hypotheses

The parameter we are interested in is the unknown number of answers j copied from s . Let γ_{js} represent this number. The statistical hypotheses to be tested against each other, therefore, are

$$H_0 : \gamma_{js} = 0, \tag{2}$$

and

$$H_1 : \gamma_{js} > 0. \tag{3}$$

Let I_{jsi} be a (random) indicator function used to establish if the response of j and s on item i match. That is,

$$I_{jsi} = \begin{cases} 1 & \text{if } U_{ji} = U_{si} \\ 0 & \text{if } U_{ji} \neq U_{si} \end{cases}. \tag{4}$$

The number of items with matching choices of alternatives by j and s is defined as

$$M_{js} = \sum_{i=1}^n I_{jsi}. \tag{5}$$

Note that the alternative hypothesis in Equation 3 is *not* that j copied all answers in the test. If it were, the test would boil down to a simple check if $M_{js} = n$. We need a statistical test just because under the alternative hypothesis γ_{js} is an unknown parameter with a large range of possible values.

If j and s have no access to each other's answers, they choose their alternatives for item i independently, with probabilities following the response model that holds, such as the one in Equation 1. The probability of j and s choosing alternative a is then equal to

$$\pi_{jsi_a} = \pi_{i_a}(\theta_j) \pi_{i_a}(\theta_s). \tag{6}$$

It follows that if j does not copy the answer by s to item i , the probability of a random match between the choices by j and s is equal to

$$\Pr\{I_{jsi} = 1\} = \pi_{jsi} = \sum_{a=1}^k \pi_{jsi_a}. \tag{7}$$

On the other hand, if j did copy the answer by s , the probability of a match between their responses would be equal to one. In Equation 7, k is the number of response alternatives on item i . For the case of a dichotomous response model, such as the well-known three-parameter logistic (3PL) model, the theory in this paper holds for $k = 2$.

Shifting our focus to the entire test, the following two hypotheses can be formulated:

$$H_0 : \pi_{jsi} = \sum_{a=1}^k \pi_{jsi_a} \quad \text{for } i = 1, \dots, n, \quad (8)$$

$$H_1 : \pi_{jsi} = \begin{cases} \sum_{a=1}^k \pi_{jsi_a} & \text{for } n - \gamma_{js} \text{ items} \\ 1 & \text{for } \gamma_{js} > 0 \text{ items.} \end{cases} \quad (9)$$

The null hypothesis specifies a vector of known probabilities of a match between the alternatives chosen for the items in the test. The alternative hypothesis specifies two possible values for each of these probabilities. We know that one of these values holds for a subset of $n - \gamma_{js}$ items and the other for a subset of γ_{js} items. But we do not know which items belong to which subset. Because of the role of γ_{js} in the definition of these subsets, the hypotheses in Equations 8 and 9 are equivalent to the original hypotheses in Equations 2 and 3.

Observe that the definition of the probabilities in Equation 7 and, therefore, the two hypotheses in Equations 8 and 9, are symmetric in j and s . We assume independent answers if there is no copying. If there is copying, there is not only dependence between the responses of the two test takers but even agreement. Both independence and agreement are symmetric notions. The test we propose is for the case in which we have no prior information identifying s as the source, for example, a test as part of a screening of the response vectors of a group of test takers to identify potentially suspicious agreement, with a subsequent search for possible evidence of cheating by pairs of students with significant results. In this case, the distinction between j and s is arbitrary; a test of j copying from s is also a test of the reverse possibility, as well as the possibility that j and s have cooperated and solved some of the items jointly.

The alternative case in which we do have prior evidence of s having been served as source for j is addressed later in this article.

Null Distribution

It follows that, under the null hypothesis in Equation 8, the number of matching alternatives is the result from a series of independent Bernoulli trials, each with a different probability of a random match, π_{jsi} , given in Equation 7. Therefore, the distribution of M_{js} belongs to the family of the generalized binomial (e.g., Lord, 1980, sect. 4.1), sometimes also called the compound binomial (although this name is more appropriate for the case of random sampling of items).

The family of generalized binomial distributions does not have a probability function in closed form, but its probabilities can easily be calculated using the generating function

$$\prod_{i=1}^n [Q_i + zP_i], \tag{10}$$

where P_i is the probability of success on the i th trial (here the probability of a match on item i) in a series of n trials and $Q_i = 1 - P_i$.

Let $f_n(m)$ be the probability of m successes in n trials. This probability is given by the coefficient of z^m . If, for example, the total number of trials $n = 2$, the coefficients of z^m for $m = 0, 1$ and 2 are equal to $Q_1Q_2, Q_1P_2 + P_1Q_2$, and P_1P_2 , respectively.

For larger tests, these probabilities can easily be calculated using the recursive procedure in Lord and Wingersky (1984). The procedure begins the calculation of the probability distribution of the number of successes in a series consisting only of the first trial, and at each recursion step t adds a next trial, until $t = n$. More formally, if $t = 1, f_t(m) = Q_1$ for $m = 0$ and $f_t(m) = P_1$ for $m = 1$. For each new step, the probabilities $f_t(m)$ follow from the probabilities in the previous step as

$$f_t(m) = Q_t f_{t-1}(m) + P_t f_{t-1}(m - 1), \tag{11}$$

where $f_t(m) = 0$ if $m < 0$ or $m > t$.

Alternative Distributions

We use Γ_{js} to denote the (unknown) set of γ_{js} indices of the items in the test for which j did copy the answer by s . The distribution of M_{js} under the alternative hypothesis has probabilities

$$f_n(m; \gamma_{js}) = \begin{cases} 0 & \text{for } m < \gamma_{js} \\ f_{n-\gamma_{js}}(m - \gamma_{js}; \bar{\Gamma}_{js}) & \text{for } \gamma_{js} \leq m \leq n, \end{cases} \tag{12}$$

where $f_{n-\gamma_{js}}(m - \gamma_{js}; \bar{\Gamma}_{js})$ is the probability function of the generalized binomial defined over the $n - \gamma_{js}$ items in $\bar{\Gamma}_{js}$. The distribution in Equation 12 can be motivated as follows: Because under the alternative hypothesis γ_{js} items were copied, the probability of fewer than γ_{js} matches is equal to zero. Furthermore, the event $M_{js} = \gamma_{js}$ is possible only if zero random matches occur on the remaining $n - \gamma_{js}$ items. For $m = \gamma_{js}, f_{n-\gamma_{js}}(m - \gamma_{js}; \bar{\Gamma}_{js})$ gives the probability of zero random matches on the items in $\bar{\Gamma}_{js}$. Likewise, for $m = \gamma_{js} + 1, f_{n-\gamma_{js}}(m - \gamma_{js}; \bar{\Gamma}_{js})$ gives the probability of one random match, and so forth.

Stochastic Order

The two hypotheses in Equations 8 and 9 imply a right-sided statistical test. Thus, it is important that the upper tail of any alternative distribution in Equation 12 is always to the right of the upper tail of the null distribution. This property can be established by showing that the family of distributions in Equation 12 is stochastically increasing in the number of answers copied, $\gamma_{js} = 0, 1, \dots, n$. Because an

increase in γ_{js} is identical to the jump of the probabilities π_{jsi} to the value of one for some of the test items, the same result can be obtained by showing that the generalized binomial family is increasing in its success parameters.

Using the property in Equation 11, van der Linden (submitted) shows that the following relation holds for the cumulative distribution function of the generalized binomial:

$$F_n(m) = (1 - \pi_i) f_{n-1}(m) + F_{n-1}(m - 1), \quad (13)$$

where π_i is the success probability for an arbitrary trial i , $F_n(m)$ is the distribution function for all n trials, and $f_{n-1}(m)$ and $F_{n-1}(m - 1)$ are the probability and distribution function for the $n - 1$ trials that remain if trial i is removed from the set. Because the factor $1 - \pi_i$ is decreasing in π_i , the generalized binomial is stochastically increasing in its parameters, and the upper tail of an alternative distribution in Equation 9 condition is always to the right of the tail of the null distribution.

Statistical Test

A (nonrandomized) right-sided test with significance level α has as its critical value the smallest value of M_{js} , to be denoted as m^* , for which the generalized binomial distribution with the probabilities of a match, π_{jsi} , defined in Equation 7 satisfies

$$\Pr\{M_{js} \geq m^*\} \leq \alpha. \quad (14)$$

Critical value m^* can easily be calculated from the recursive procedure in Equation 13.

Comparison with Statistic ω

Statistic ω is a standardized version of M_{js} given by

$$\omega = \frac{M_{js} - E[M_{js} | \mathbf{u}_s]}{\text{Var}[M_{js} | \mathbf{u}_s]^{1/2}}, \quad (15)$$

where the expected value and variance are defined as

$$E[M_{js} | \mathbf{u}_s] = \sum_{i=1}^n \pi_{i_a^{(s)}}(\theta_j), \quad (16)$$

and

$$\text{Var}[M_{js} | \mathbf{u}_s] = \sum_{i=1}^n \pi_{i_a^{(s)}}(\theta_j) [1 - \pi_{i_a^{(s)}}(\theta_j)], \quad (17)$$

\mathbf{u}_s is the vector of alternatives chosen by s , and $\pi_{i_a}^{(s)}(\theta_j)$ denotes the probability of j choosing alternative a on item i , which was also chosen by s (Wollack, 1997, eqs. 6 and 8). The value of ω is compared with an appropriate critical value, say ω^* , from the standard normal distribution, and the null hypothesis is rejected in favor of a one-tailed alternative (that copying has taken place) if $\omega \geq \omega^*$.

Observe that the mean $E[M_{js} | \mathbf{u}_s]$ and variance $\text{Var}[M_{js} | \mathbf{u}_s]$ in Equation 15 are conditional on the response vector of s . The test based on ω is, therefore, a conditional test. The number of matching alternatives, M_{js} , is the same as in the test based on Equation 14; the conditioning is introduced only by the standardization and the null distribution.

Exact Conditional Test

If the interest is in a conditional statistical test of answer copying given the response vector produced by s , the following *exact* test is offered as an alternative to ω .

Let $i_a^{(s)}$ still denote the alternative on item i chosen by s . The conditional probability of a matching choice by j is equal to

$$\Pr\{I_{jsi} = 1 | i_a^{(s)}\} = \pi_{i_a}^{(s)}(\theta_j), \tag{18}$$

that is, the marginal probability of j choosing the given alternative. An exact conditional test of copying also has a generalized binomial as null distribution, but with the probabilities of a match given in Equation 7 replaced by the conditional probabilities in Equation 18.

If a conditional test is preferred, our choice would be the one based on Equation 18 instead of ω . The normal approximation for ω holds only asymptotically. For shorter tests, we expect the approximation to be problematic, particularly if the generalized binomial distribution is skewed. This happens if the probabilities of a match in Equation 18 are smaller than .50—a condition that is easily satisfied, for example, if j and s respond at entirely different ability levels.

Discussion

This issue of conditioning in a statistical test, such as in Equations 15 and 18, is delicate. The choice whether or not to condition on a certain event can only be motivated by the nature of the application, for example, the intended interpretation of the outcome of the test (Lehmann, 1986, chap. 10).

We believe that, for the case of symmetry between j and s introduced above, it is enough to condition on the person and item parameters in the response model, as was done in the definition of the probabilities of a random match in Equation 7. These parameters capture the structural aspects of the situation. If we ignored them, the test would become sensitive to these aspects and possibly become confounded, for example, with the differences in ability between the two test takers.

If we introduce additional conditioning on the response vector by one of the test takers, as in Equation 18, the symmetry inherent in the hypotheses in Equations 8 and 9 is given up. One of the consequences is that if two test takers decide to cooperate and produce common answers on a portion of the test, one of them might be identified as a copier by the test and the other as a noncopier—clearly an undesirable outcome.

One case in which the conditional test based on Equation 18 offers advantages is when we have prior evidence that s has served as source but wonder if j has been a copier. If such evidence is present, the conditional test in Equation 18 offers us control of the Type I error at level α for the actual response vector produced by s rather than over replications of this vector. Also, we do not have to make any assumptions on the response probabilities of s ; particularly, we do not have to know θ_s .

Power Analysis

The power function of the test is given by the probabilities $\Pr\{M_{js} \geq m^* | \gamma_{js}\}$, where M_{js} has the distribution associated with the true alternative among the set of alternatives in Equation 9. This distribution is not only dependent on the unknown number of answers copied by j , γ_{js} , but also on the probabilities of a random match, π_{jsis} , on the items in the subset for which j did not copy any answer, $\bar{\Gamma}_{js}$.

However, for a given value of γ_{js} , although the probabilities of a random match for the items in $\bar{\Gamma}_{js}$ are unknown, we can still calculate such descriptive information as the average, minimum, maximum, and p th percentile of the power of the test over all possible subsets of items of size $n - \gamma_{js}$. For larger values of n , if γ_{js} approaches $n/2$, the number of possible subsets quickly becomes prohibitively large. In such cases, we can estimate these quantities using a sufficiently large set of random samples of $n - \gamma_{js}$ items from the test.

Although not discussed in Wollack (1997), the same logic underlying the approximation of the null distribution of statistic ω can be used to approximate its alternative distributions. This strategy will be used in our power analyses in the next section. For γ_{js} items copied, the power of a test based on ω is approximated as

$$\Pr\left\{Z \geq \frac{M_{js}^* - \gamma_{js} - E[M_{js} | \mathbf{u}_s, \bar{\Gamma}_{js}]}{\text{Var}[M_{js} | \mathbf{u}_s, \bar{\Gamma}_{js}]^{1/2}}\right\}, \quad (19)$$

with Z being a standard normal variable, M_{js}^* the number of matches corresponding with the critical value in a test based on Equation 15, and the mean and variance computed given the responses by s on the items in $\bar{\Gamma}_{js}$. The fact that $\bar{\Gamma}_{js}$ is unknown can be dealt with in the same way as suggested above. Because our main interest is in an evaluation of the generalized binomial test based on Equation 7, we compare the power of this test with results for ω based on Equation 19 only. In particular, a comparison between the power of ω and the test using the conditional probabilities in Equation 18 is omitted.

Numerical Examples

The same set of item parameters for the nominal response model in Equation 1 as in Wollack (1997) was used. (The set is available from the authors upon request.) For these values, the null distributions and power functions for tests of 10, 20, and 30 items under the generalized binomial given by Equation 10 and the normal approximation assumed for Equation 15 were analyzed. The differences between the two types of distributions depend on the degree to which the probabilities π_{jsi} are bounded away from .50; that is, the true null distribution is skewed. The analyses were therefore done for four pairs of test takers with probabilities π_{jsi} systematically covering the interval (0, .50). Observe that this range is for probabilities for which the power of the test is largest.

The pairs of values (θ_j, θ_s) for these test takers were selected as follows. First, a grid of values for (θ_j, θ_s) , each running from -2.0 to 2.0 with step size .10, was chosen. Second, for each pair (θ_j, θ_s) , the average of the π_{jsi} values, $\bar{\pi}_{js}$, was calculated from the item parameters in the test. Third, all averages larger than .50 were ignored, and the remaining averages were ordered in size. Finally, the four pairs of (θ_j, θ_s) values that produced averages $\bar{\pi}_{jsi}$ at the 0th, 33th, 67th, and 100th percentiles in this ordered set were used in this example. The procedure was repeated for each set of $n = 10, 20,$ and 30 items. The results are summarized in Table 1.

Null Distributions

Plots of the probability functions of the generalized binomial null distribution of M_{js} and the implied normal density function for M_{js} by the test based on ω are given in Figures 1–3. For each test length, we have four plots, one for each pair of test takers with an average $\bar{\pi}_{jsi}$ selected as described above. The null distributions of ω depend on the response vector by s . Therefore, for each of the four values for

TABLE 1
Selected Pairs of Ability Values (θ_j, θ_s) and Average Probabilities of a Match $\bar{\pi}_{jsi}$ for Test Lengths $n = 10, 20,$ and 30

n	(θ_j, θ_s)	$\bar{\pi}_{jsi}$
10	(-2.0, 2.0)	.17
	(-1.0, 0.8)	.35
	(0.7, 1.5)	.42
	(1.3, 2.0)	.50
20	(-2.0, 2.0)	.16
	(-0.6, 1.0)	.35
	(-0.9, 2.0)	.42
	(-2.0, -0.9)	.50
30	(-2.0, 2.0)	.18
	(-0.8, 1.0)	.36
	(0.9, 2.0)	.43
	(-0.9, -0.6)	.50

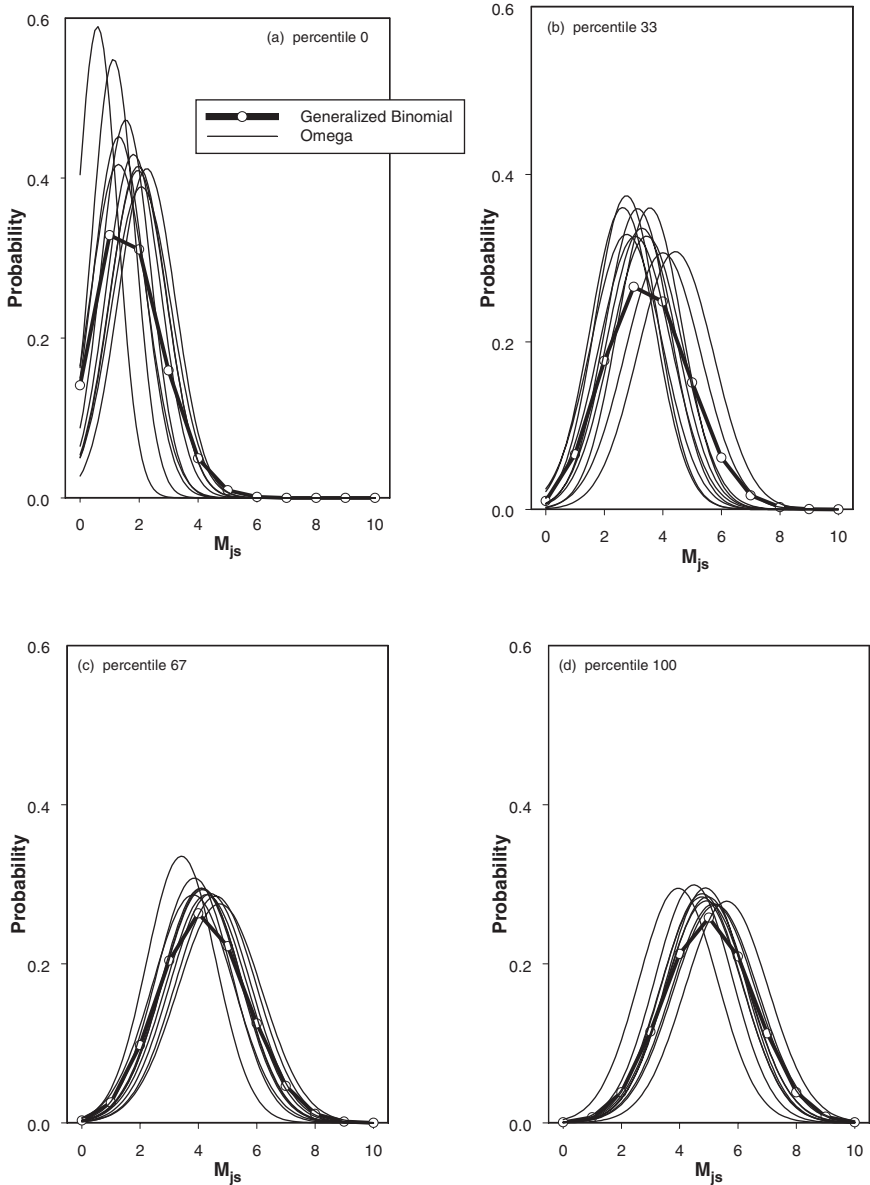


FIGURE 1. Null distributions of the generalized binomial test (bold line) and the test based on statistic ω for samples of 10 response vectors for the same value of θ_s (thin lines) for pairs of examinees with average probability of a random match $\bar{\pi}_{jsi} = .17, .35, .42,$ and $.50$ ($n = 10$).

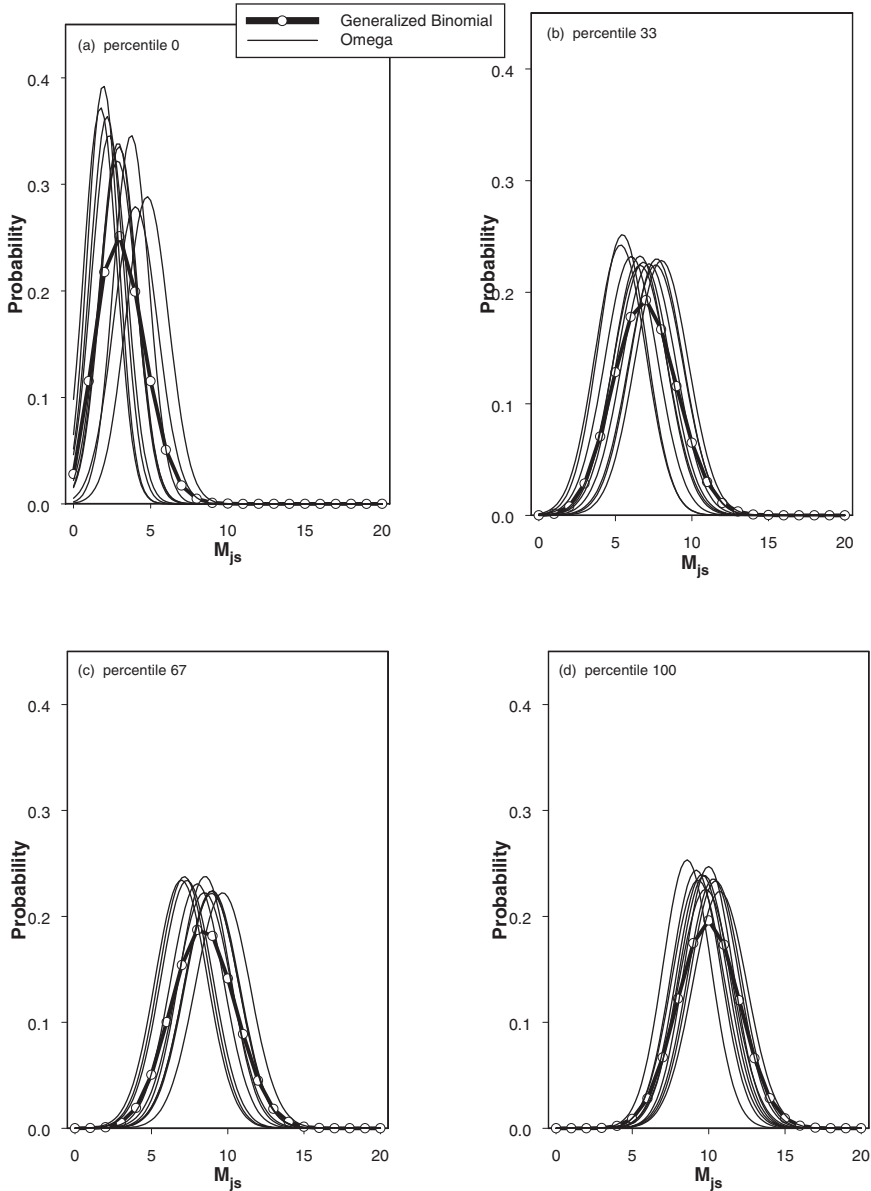


FIGURE 2. Null distributions of the generalized binomial test (bold line) and the test based on statistic ω for samples of 10 response vectors for the same value of θ_s (thin lines) for pairs of examinees with average probability of a random match $\bar{\pi}_{jsi} = .16, .35, .42, \text{ and } .50$ ($n = 20$).

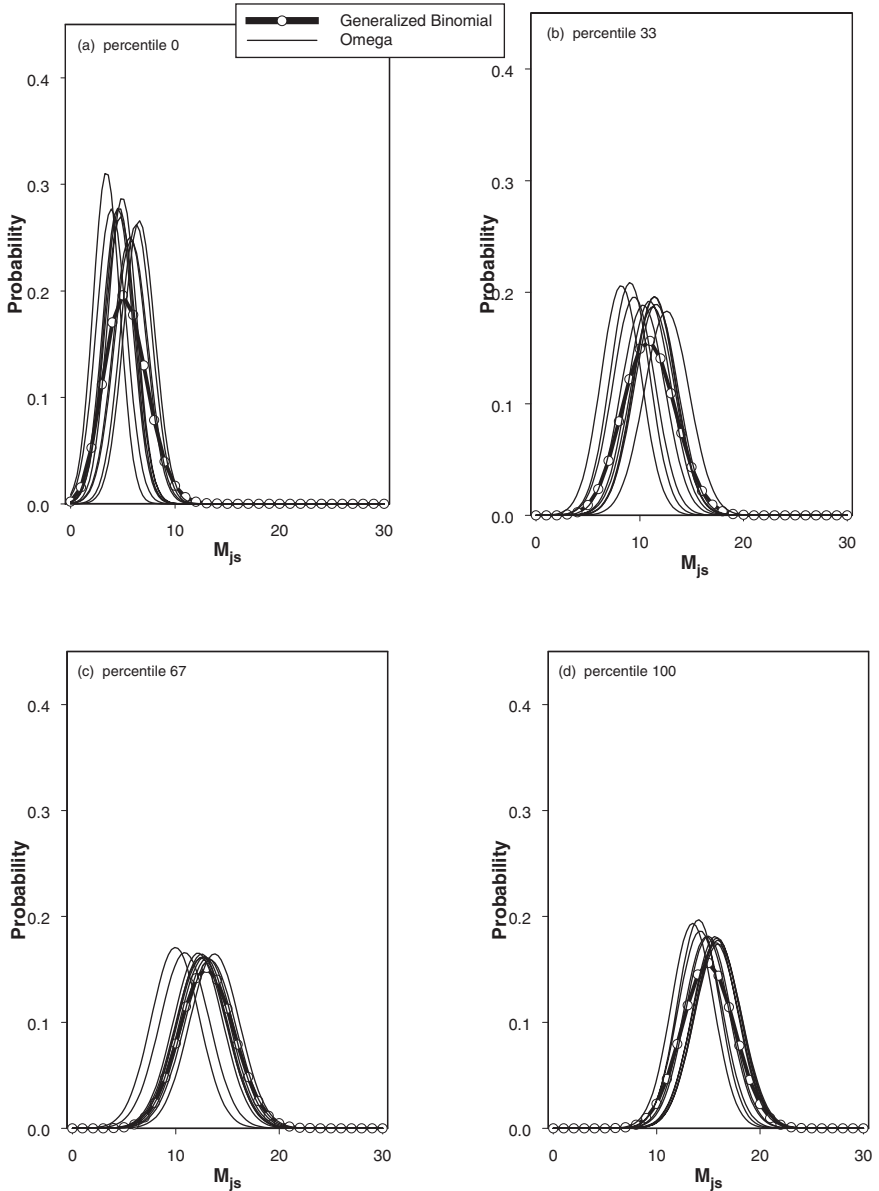


FIGURE 3. Null distributions of the generalized binomial test (bold line) and the test based on statistic ω for samples of 10 response vectors for the same value of θ_s (thin lines) for pairs of examinees with average probability of a random match $\bar{\pi}_{jsi} = .18, .36, .43, \text{ and } .50$ ($n = 30$).

θ_s , 10 response vectors were generated randomly. This number was chosen because it was small enough to allow visual inspection of these distributions and large enough to get an impression of their variation. Because we compare an unconditional with a conditional test, our focus is not on the individual null distributions of ω for a given response vector, but on the difference between the trend they show and the generalized binomial.

As expected, the null distributions for the generalized-binomial test in this article are quite skewed for the pair of values (θ_j, θ_s) with probabilities of a match far away from 0.50. The tests based on ω for these cases seem to entail critical values that tend to be much lower than the one for the generalized binomial. The reason is the approximation of a strongly skewed distribution by a symmetric one. This observation holds particularly for $n = 10$. For this test length, the values (θ_j, θ_s) with the lowest probabilities of a match show upper tails for the normal approximations to the null distribution that are all to the left of the tail of the exact unconditional distribution. In fact, the majority of these approximations would result in a critical value lower than half the size of the true value. For pairs of values (θ_j, θ_s) with probabilities of a match closer to .50, the generalized binomial distribution is approximately symmetric and tends to be covered much better by the normal approximations associated with ω , although the latter still show a slight bias toward a smaller than the true variance.

Power Calculations

Figures 4–6 show the plots with the results of the power analysis of the generalized binomial test for $\alpha = .05$ (which is not necessarily the best choice for α in actual applications of the test). The test lengths and the pairs of ability values (θ_j, θ_s) were the same as for Figures 1–3. The plots are based on all possible subsets of γ_{js} items from the test, provided the number was smaller than 250; for larger numbers, a random sample of this size was taken. Each plot shows the curves for the 0th, 25th, 50th, 75th, and 100th percentile in the distribution of power over these subsets. Obviously, the curves cannot cross, and curves for the higher percentiles are more to the left.

The dominant impression from these plots is that the lower the average $\bar{\pi}_{jsi}$ (that is, the more skewed the null distribution), the more the power curves approach the ideal of a steep curve close to $M_{js} = 0$. Also, for a lower average $\bar{\pi}_{jsi}$ the variation in power becomes smaller. Both trends seem to hold for any test length.

The procedure was replicated to calculate the conditional power function for the tests based on ω in Equation 19. The results were calculated for ten different response vectors for each of the values for θ_s . In Figures 7–9, the same percentiles in the distribution of power as in Figures 4–6 are shown. The curves for ω with the lower probabilities of a match are generally steeper than those for the exact test in Figures 4–6. The reason for this increase in steepness is the well-known trade-off between Type I and Type II errors in statistical hypothesis testing. The test based on ω has a tendency to set the critical value too low for the lower probabilities of a random match, and hence to a lower probability of a Type II error. If the critical

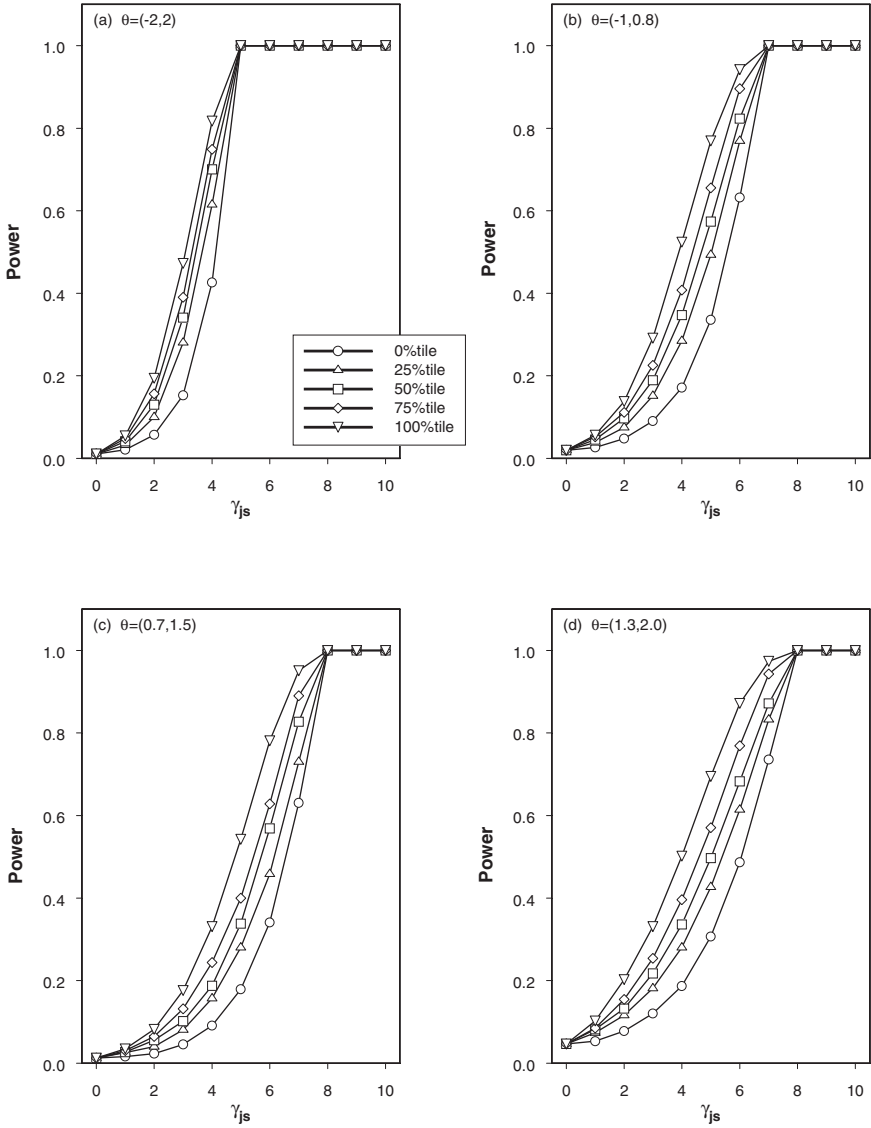


FIGURE 4. Curves for the 0th, 25th, 50th, 75th, and 100th percentiles in the power distribution of the generalized binomial test for pairs of examinees with average probability of a random match $\bar{\pi}_{jsi} = .17, .35, .42, \text{ and } .50$ ($n = 10$).

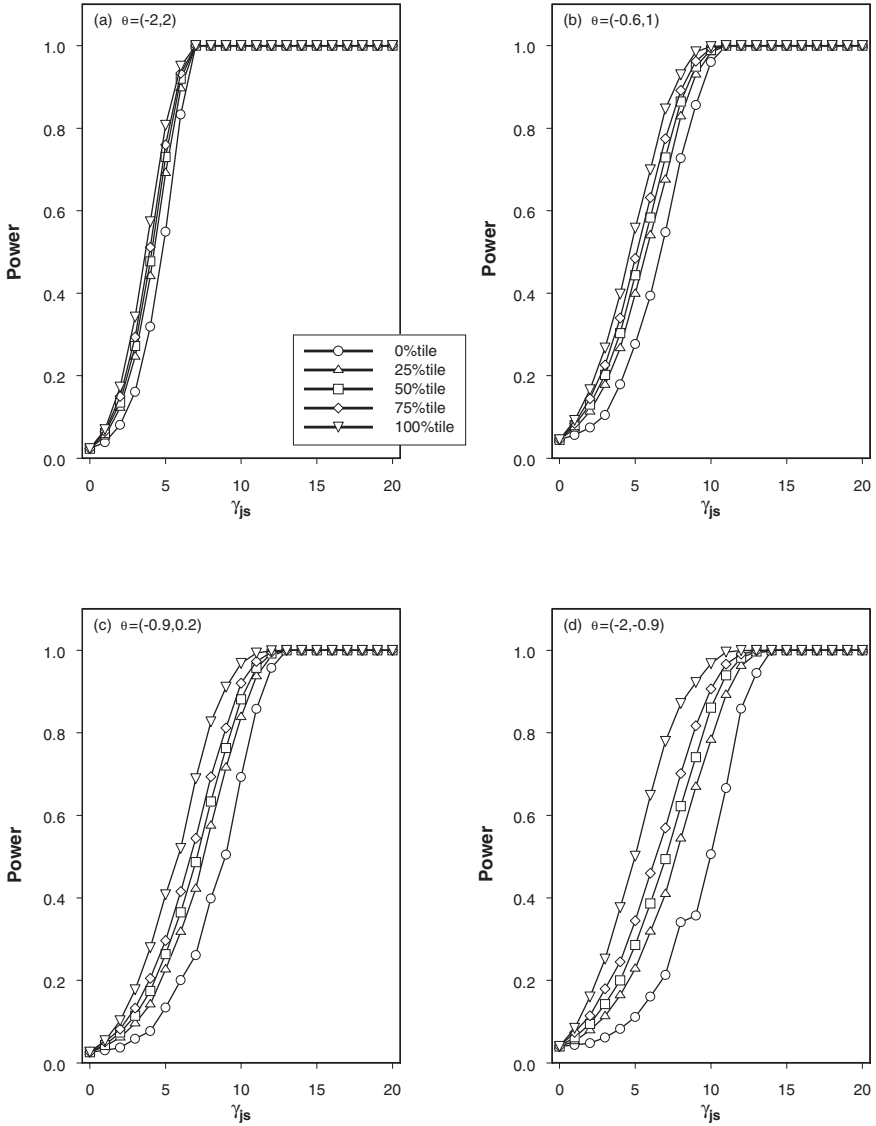


FIGURE 5. Curves for the 0th, 25th, 50th, 75th, and 100th percentiles in the power distribution of the generalized binomial test for pairs of examinees with average probability of a random match $\bar{\pi}_{jsi} = .16, .35, .42, \text{ and } .50$ ($n = 20$).

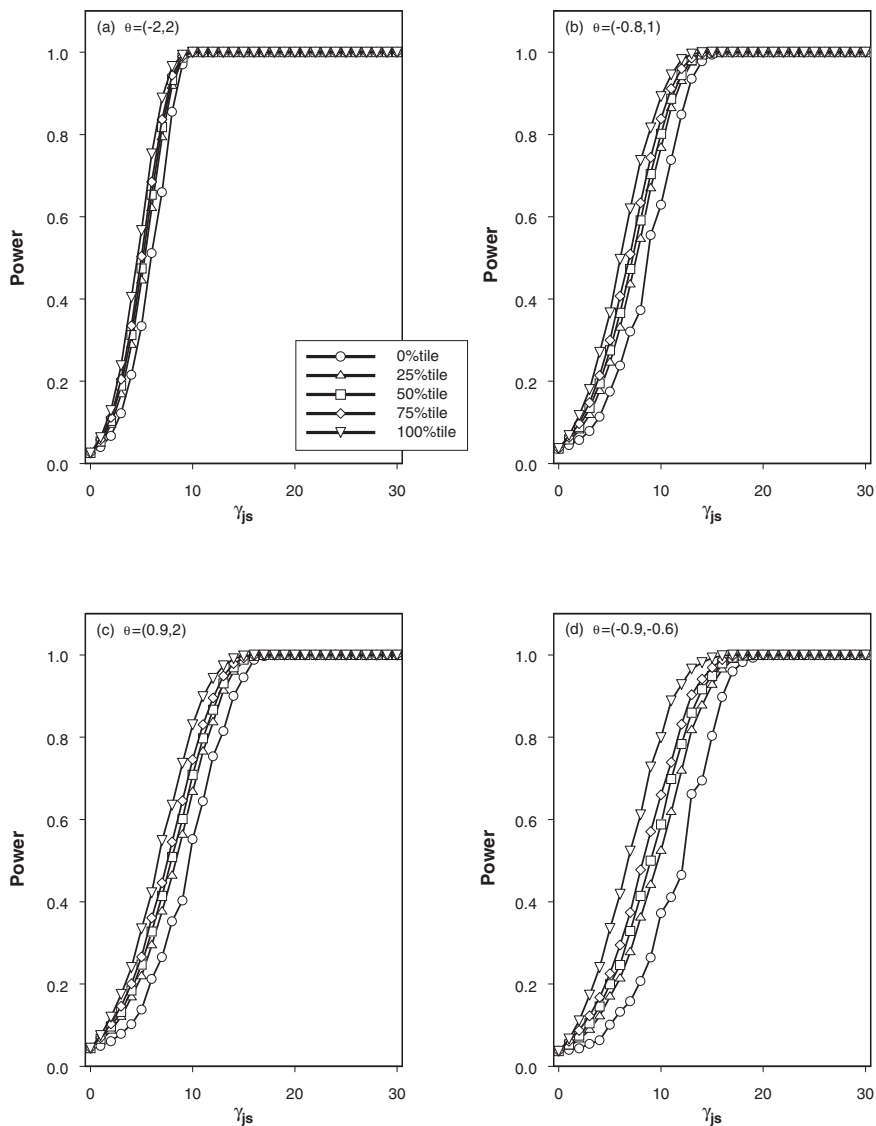


FIGURE 6. Curves for the 0th, 25th, 50th, 75th, and 100th percentiles in the power distribution of the generalized binomial test for pairs of examinees with average probability of a random match $\bar{\pi}_{jsi} = .18, .36, .43, \text{ and } .50$ ($n = 30$).

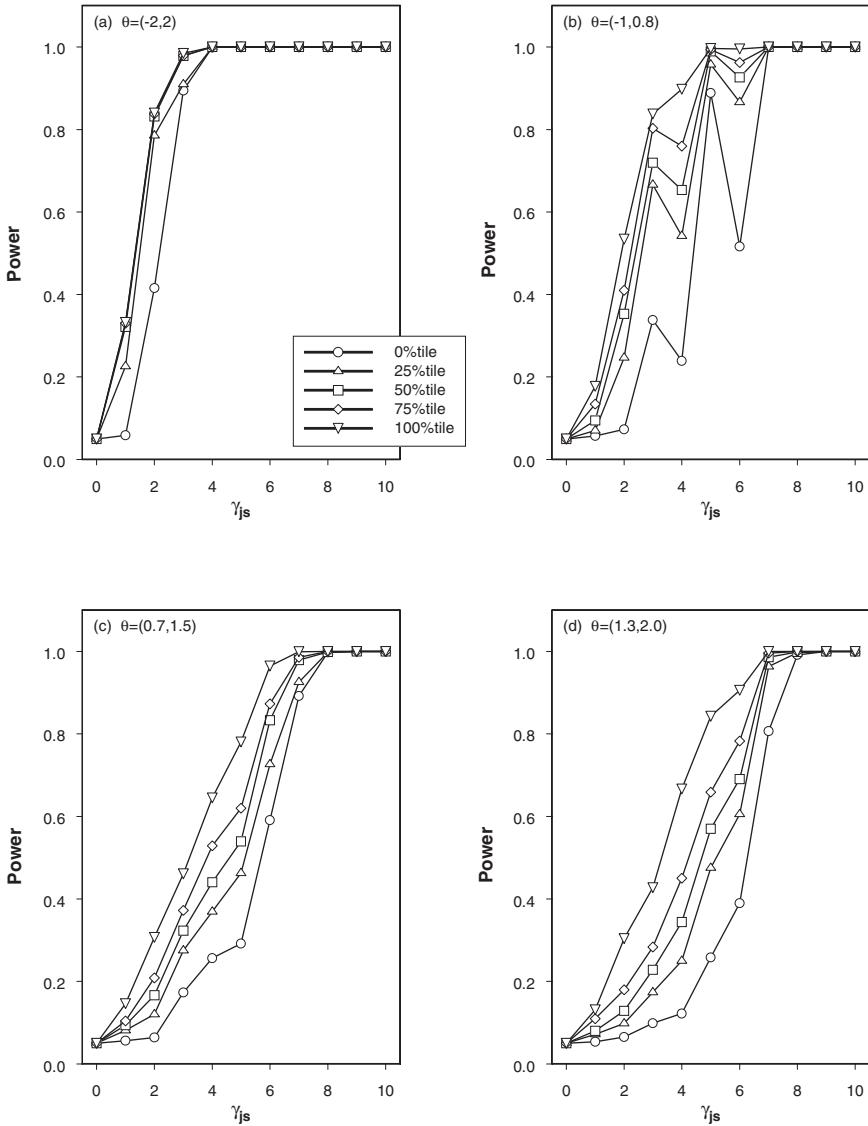


FIGURE 7. Curves for the 0th, 25th, 50th, 75th, and 100th percentiles in the power distribution of the test based on statistic ω for pairs of examinees with average probability of a random match $\bar{\pi}_{jst} = .17, .35, .42, \text{ and } .50$ ($n = 10$).

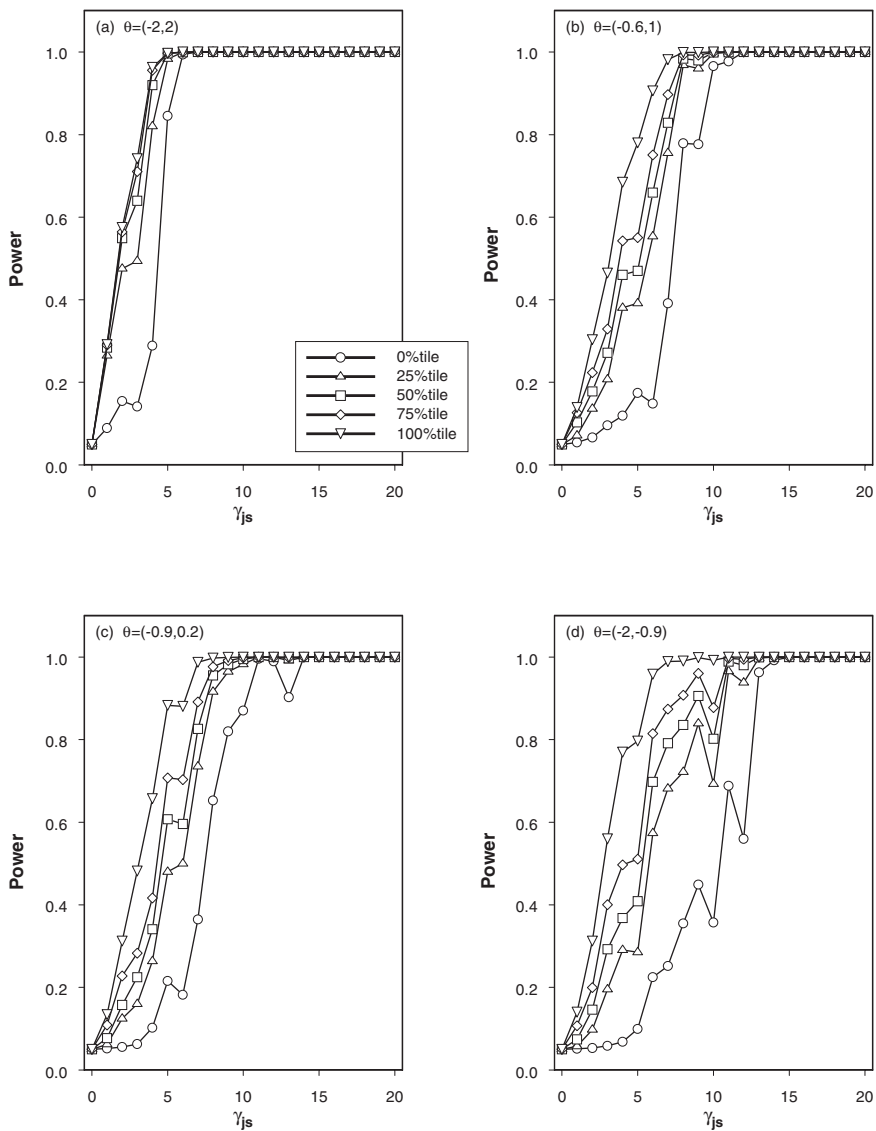


FIGURE 8. Curves for the 0th, 25th, 50th, 75th, and 100th percentiles in the power distribution of the test based on statistic ω for pairs of examinees with average probability of a random match $\bar{\pi}_{jsi} = .16, .35, .42, \text{ and } .50$ ($n = 20$).

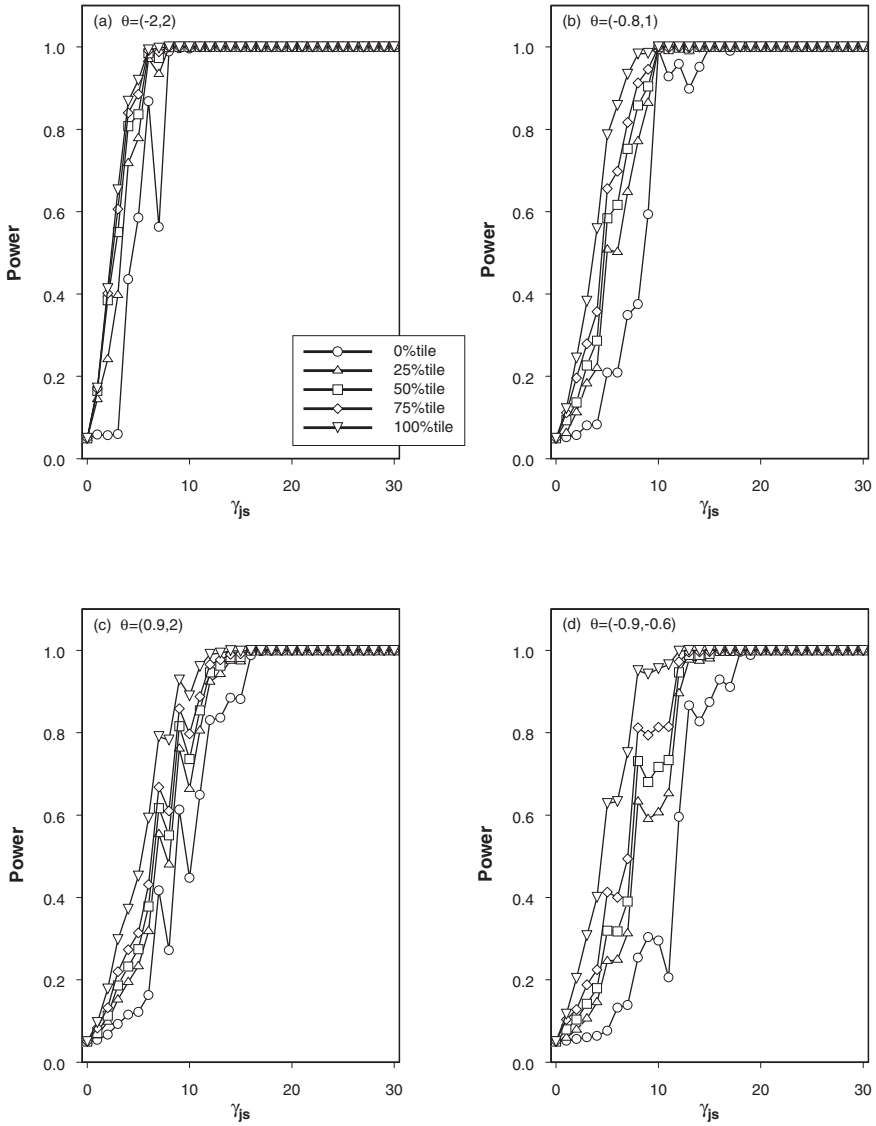


FIGURE 9. Curves for the 0th, 25th, 50th, 75th, and 100th percentiles in the power distribution of the test based on statistic ω for pairs of examinees with average probability of a random match $\bar{\pi}_{jst} = .18, .36, .43, \text{ and } .50$ ($n = 30$).

value were set equal to zero, the test would even show a perfect power curve (but always make a Type I error when there is no copying).

Conclusions

The null distribution for a test of answer copying on multiple-choice items was derived, and it was indicated how a power analysis for this test can be conducted. The test assumes a known response model fitting the regular response process. In applications, the parameters in this model as well as the pair of abilities (θ_j , θ_s) have to be estimated from actual response data. The impact of this parameter estimation, which we do not expect to deviate much from the results for the ω statistic in Wollack and Cohen (1998), will be the subject of a future study. The following conclusions are, therefore, tentative.

For a wide range of pairs of abilities (θ_j , θ_s), a test of size $\alpha = .05$ based on the generalized binomial distribution seems to have enough power to detect copying on some 30–40% of the items in the test with certainty. The power becomes better the larger the difference between the abilities θ_j and θ_s . Best power was obtained for the cases with a low ability for the copier, j , and a higher ability for the source, s . The reason for this result is a lower probability of a random match and hence a null distribution more skewed to the right.

The properties of the test were compared with those of ω , which is the only other test available in the literature based on the assumption of a known response model. This test is conditional on the response vector produced by s . The number of matches corresponding with its critical value, therefore, varies as a function of this vector. Because it assumes a normal approximation to the null distribution, the test shows a tendency to set much lower values for shorter tests and lower probabilities of matching alternatives, exactly the condition under which the test based on the generalized binomial showed maximum power.

References

- Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69*, 44–49.
- 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 and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33–49). New York: Springer.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *6*, 152–165.
- 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 96-7). Princeton, NJ: Educational Testing Service.
- Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Springer.
- Lehmann, E. L. (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.

- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8, 452–461.
- 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., van der Linden, W. J., & Meijer, R. R. (2006). *Detecting answer copying using statistic kappa*. *Applied Psychological Measurement*, 30, 412–431.
- Thissen, D. (1991). *MULTILOG user's guide* (Version 6). Chicago: Scientific Software, Inc.
- van der Linden, W. J. (submitted). *Order properties in observed-score distributions*.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. NY: Springer-Verlag.
- 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.
- 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.

Authors

- WIM J. VAN DER LINDEN is Professor of Measurement and Data Analysis, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; w.j.vanderlinden@utwente.nl. His areas of specialization include test theory, applied statistics, and research methods.
- LEONARDO SOTARIDONA was Research Assistant, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. His current address is: CTB/McGraw-Hill, 7400 S. Alton Court, Centennial, CO, 80112; leonardo_sotaridona@ctb.com. His areas of specialization are test theory and detection of cheating on tests.

Manuscript received August 1, 2003
Revision received May 13, 2004
Accepted December 29, 2004