PSYCHOMETRIKA—VOL. 80, NO. 3, 689–706
SEPTEMBER 2015
DOI: 10.1007/s11336-014-9409-x

BAYESIAN CHECKS ON CHEATING ON TESTS

WIM J. VAN DER LINDEN

CTB/MCGRAW-HILL

CHARLES LEWIS

FORDHAM UNIVERSITY

Posterior odds of cheating on achievement tests are presented as an alternative to p values reported for statistical hypothesis testing for several of the probabilistic models in the literature on the detection of cheating. It is shown how to calculate their combinatorial expressions with the help of a reformulation of the simple recursive algorithm for the calculation of number-correct score distributions used throughout the testing industry. Using the odds avoids the arbitrary choice between statistical tests of answer copying that do and do not condition on the responses the test taker is suspected to have copied and allows the testing agency to account for existing circumstantial evidence of cheating through the specification of prior odds.

Key words: answer copying, Bayesian checks, cheating on tests, erasure analysis, generalized binomial distribution, statistical hypothesis testing..

1. Introduction

Although cheating on achievement tests is not common, it is definitely on the rise. For instance, in a recent survey based on a randomized response technique among a population of university students, Fox & Meijer (2008) found up to 20–31 % of the students reporting such types of cheating as conferring during the test, looking at the test papers of someone else, or allowing others to copy their own work. One of the main weapons of the testing industry against cheating is statistical detection of its occurrence. The current statistical methods address four different kinds of cheating: (i) test takers copying answers from others; (ii) fraudulent erasures on answer cheats after completion of the test; (iii) attempts to memorize items during testing to share them with later students; and (iv) preknowledge of unreleased test items. The main methods used in these categories are briefly reviewed here; more specific information on some of these methods is provided later in this paper.

Early methods for the detection of answer copying on multiple-choice tests were proposed by Angoff (1974), Frary et al. (1977), and Saupe (1960). These methods have the statistical form of a test of the null hypothesis of no answer copying based on the number of matching responses between two test takers. Their null distributions, however, were derived from rather ad hoc assumptions about the response process. Versions of the same type of tests derived from more sophisticated assumptions are the K-index (Holland 1996; Lewis & Thayer 1998; Sotaridona & Meijer 2002), as well as the tests proposed by van der Linden & Sotaridona (2004) and Wesolowsky (2000). Tests directly based on standard item response models widely used in the testing industry were provided by Wollack (1997) and van der Linden & Sotaridona (2006). Although these models are based on response-model assumptions, it is a standard operating procedure in the testing industry to check their validity while field testing the items, and to discard items with a

Correspondence should be sent to Wim J. van der Linden, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940. Email: wim_vanderlinden@ctb.com.

less satisfactory fit. An entirely different approach was recently offered by Belov & Armstrong (2010). Their method uses the Kullback–Leibler divergence to detect inconsistencies in test takers' performances between common sections administered to all of them and rotating sections with pretest items, and then utilizes the K-index for secondary analysis to find evidence of answer copying on the common sections.

Fraudulent erasures on answer sheets arise when teachers or school administrators "help students" by improving wrong answers on their answer sheets after the test, or do so to hide underachievement by their class or school. One of the first to look into the possibility of detecting this type of fraud using optical scanning of the answer sheets for erasures was Qualls (2001). Her main goal was to obtain estimates of the typical distributions of the numbers of regular answer changes by test takers on low-stakes tests, where fraud is generally absent, for use with high-stakes tests to detect outliers. Jacob and Levitt (2003a,b, 2004) used multinomial logits of numbers of erasures regressed on performances by the same students in other years as well as several kinds of background information to detect unusual erasure patterns. Their case study for a Chicago school district, which created considerable interest thanks to its inclusion in Levitt & Rubner (2005), led to the identification of several teachers who had been erasing answers. A statistical test and method of residual analyses for the detection of fraudulent erasures based on item response modeling of the probability of wrong-to-right (WR) answer changes were given by Linden & Jeon (2012).

Most detection of memorization and preknowledge of test items involves methods of person-fit analyses from item response theory. For computerized testing, they become more powerful when combined with the analysis of the response times recorded for the items. Attempts to memorize test items typically lead to careless responses that are given only to move from one item to the next, as well as response times not representative of the actual amount of labor required to solve the items. The latter also holds for items that test takers have already seen before the test is administered to them, in combination with unlikely correct responses given their ability demonstrated on the other items. An earlier review of person-fit analyses used to detect aberrant responses was given by Meijer & Sijtsma (2001). Later methods include Bayesian options by Glas & Meijer (2003) and McLeod et al. (2003), as well as methods based on a Bayesian analysis of residual response times by van der Linden & Guo (2008).

Several of the methods just reviewed are based on—or were at least motivated by—notions derived from statistical hypothesis testing. They typically focus on an obvious test statistic (for instance, the number of matching incorrect responses on multiple-choice items or the number of wrong-to-right (WR) erasures on an answer sheet), and then choose a statistical rationale for distinguishing between test takers with high but still likely values for the statistic and those with unlikely high values. Clear advantages of this approach relative to the heuristic approaches in the early literature on the detection of cheating are the use of explicit assumptions about the probability distribution of the test statistic for regular examinees as well as the distinction between regular and irregular behavior. However, its focus is exclusively on the null hypothesis of regular behavior. On the other hand, newer developments as in Linden & Jeon (2012) and van der Linden & Sotaridona (2006) also try to set up an explicit alternative hypothesis with a probabilistic structure expected to hold for those who cheat, which is then tested against the null hypothesis of no cheating (e.g., Linden & Jeon 2012; van der Linden & Sotaridona 2006).

A still unresolved issue already emerging in the very first reports on the detection of answer copying is the question of how to deal with the responses by the source. Three different answers have been given, sometimes with different motivations. One answer has been to focus on the incorrect responses by the source only—a choice usually motivated by the observation that matching correct responses constitute weak evidence of copying since both the source and the copier may just know the answer (e.g., Holland 1996) or by demonstrating confounding of the hypotheses to be tested if correct responses by the source are admitted (van der Linden & Sotaridona 2004). This approach typically views knowing an answer to an item as a deterministic event. Two alternative

approaches have been proposed: (i) treating all responses by the source as fixed and (ii) treating all of them as random. The first in the literature to struggle with the choice between the two options was Frary et al. (1977). More recently, conditional and unconditional versions of a test based on the same multinomial model for the response alternatives were proposed by van der Linden & Sotaridona (2006). Suppose an item has alternatives $a = 1, \ldots, k$, and let π_{ci_a}, and π_{si_a} denote the probabilities of c (the copier) and s (the source) choosing alternative a on item i under the response model. One version is based on the null hypothesis of c and s working independently (i.e., without any coping) on the items with probabilities of a matching choice from the alternatives equal to

$$\Pr\{\text{match on } i\} = \sum_{a=1}^{k} \pi_{ci_a}\pi_{si_a}. \tag{1}$$

The alternative is to take the responses of the source as given. Suppose s has chosen alternative a' on an item. The probability of a regular matching response by c given the choice by s is then equal to

$$\Pr\{\text{match on } i\} = \pi_{ci_{a'}}. \tag{2}$$

Obviously, for the same level of significance, the two tests may flag different test takers. Classical statistics is unable to offer any criterion for the choice between the two types of test. The underlying statistical issue is the choice between unconditional and conditional inference (Lewis 2006). Both types of inference generally have the same expected power, and the choice between them can be motivated only by such extra-statistical considerations as the nature of the application or the intended use of the results (Lehmann & Romano 2005, Sect. 10.1). For an application of the proof of equal expected power to the problem of answer copying, see Appendix 1.

Fortunately, there exists another option in the form of the calculation of the posterior odds of cheating. This option entirely avoids the arbitrary choice between unconditional and conditional inference in that, following the Bayesian logic of treating all data as given, it automatically conditions on all responses, both by the source and the copier. An additional advantage of the use of posterior odds is the ability to account for existing empirical evidence of cheating through the specification of prior odds. For instance, when the hypothesis of cheating is tested for test takers from a well-defined population, it is impossible for the types of tests just reviewed to account for our knowledge of the typical incidence of cheating in the population. The only control they offer is through the choice of significance level of the test. But that choice only impacts the proportion of cases flagged among those not involved in cheating; it does not account in any way for the proportion that actually did cheat.

An important question is how to use statistical evidence of cheating in the actual practice of testing. The general rule followed by the testing industry is not to accuse any test taker of cheating but to inform those with highly suspicious answer sheets that, due to statistical irregularities, their response patterns do not warrant any scoring, and offer them a retake. The rule has been supported by jurisdiction, acknowledging the rights of testing organizations not to release test scores that are not in agreement with their professional standards. The posterior odds derived below are intended to help testing organizations reach such conclusions, combining whatever circumstantial evidence they may have collected with the statistical information in the test taker's responses.

2. Posterior Odds of Cheating

The basic procedure we propose is first illustrated as an alternative for the current statistical tests of answer coping on multiple-choice tests reviewed above. We then give an example for the calculation of the posterior odds of fraudulent erasures on answer sheets.

2.1. Answer Copying

Let N be the set of items in the test for which we want to check on the copying of any answers by a hypothetical copier c from a source s. We use M to denote its subset of items with matching alternatives between c and s. The subset of items on which c actually copied is denoted as Γ. Although the sets M and Γ vary across pairs of test takers, for convenience, we do not index them by c and s.

Thus, it holds that

$$\Gamma \subseteq M \subseteq N. \tag{3}$$

Also, note that M is observed but Γ is unknown. In fact, it may very well hold that $\Gamma = \varnothing$ (c has not copied at all). The prior probability of Γ being a specific subset of items in the test is given by probability function

$$p(\Gamma), \quad \Gamma \in \mathcal{P}(N), \tag{4}$$

where $\mathcal{P}(N)$ is the power set of N (i.e., the set of all its possible subsets).

It may be tempting to check blindly (i.e., without any prior evidence) on answer copying between test takers on the entire set of items in a test for a given administration, but we are not in favor of this option. First, as the use of miniature electronic communication devices in cheating is on the rise, it would no longer suffice to check just on test takers sitting adjacent to each other, whereas checking on all possible pairs would be infeasible given the sheer number of them. Second, for the full set of items in a real-world test, $P(N)$ would become prohibitively large (e.g., 2^{40} for a test of 40 items), making it impossible to specify a meaningful probability distribution over it. We are therefore not in favor of such blind procedures. Rather, we recommend making checks only when there exists prior empirical evidence for individual test takers suggesting scrutiny of a specific portion of the test, for instance, a page of the answer sheet or a section in the test for which a proctor has reported suspicious behavior. Such specific analyses are more convincing, provided the evidence was collected and the prior distribution in (4) was specified before observing any of the actual responses by c and s. Finally, note that the option of specifying different prior distributions for different test takers allows us to tailor the analysis to different cases of suspicion.

As for the response model, two choices are considered. The first is the nominal response model, which (with appropriate identifiability restrictions) specifies the probability of a regular test taker with ability level θ choosing response category $u_i = 1, \ldots, k_i$ for item i as

$$p(u_i \mid \theta) \equiv \frac{\exp(\zeta_{u_i} + \lambda_{u_i}\theta)}{\sum_{u=1}^{k_i} \exp(\zeta_{u_i} + \lambda_{u_i}\theta)}, \tag{5}$$

where ζ_{u_i} and λ_{u_i} are the intercept and slope parameters for category u_i (Bock 1972, 1997). Although the methods that follow are more informative for this polytomous type of model than one of the currently popular dichotomous models, the latter tend to fit achievement test data much better and are more frequently applied. In general, dichotomous responses ($u_i = 1, 2$) have probability function

$$p(u_i \mid \theta) = p(1 \mid \theta)^{u_i}[1 - p(1 \mid \theta)]^{1-u_i}. \tag{6}$$

For this case, we use the well-known three-parameter logistic (3PL) response function as an example, which gives the probability of a correct response as

$$p_i(1 \mid \theta) \equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{7}$$

where $b_i \in (-\infty, \infty)$, $a_i \in (0, \infty)$, and $c_i \in [0, 1]$ are parameters that may be interpreted as the difficulty, discriminating power, and the height of the lower asymptote required to deal with the effects of guessing on item i, respectively. The generalization of the results in this paper to other types of dichotomous, polytomous, multidimensional, etc. response models is straightforward. Also, throughout our treatment, we assume that the item parameters have been estimated with enough precision to consider their values as known. The estimation of the ability parameter will be discussed separately below.

The conditional probability function of response $U_{ci} = u_{ci}$ for copier c on item i given the response $U_{si} = u_{si}$ by s is equal to

$$p(u_{ci} \mid \theta_c, \Gamma, u_{si}) = \begin{cases} p(u_{ci} \mid \theta_c), & \text{if } \ i \notin \Gamma, \\ 1, & \text{if } \ i \in \Gamma \text{ and } u_{ci} = u_{si}, \\ 0, & \text{if } \ i \in \Gamma \text{ and } u_{ci} \neq u_{si}. \end{cases} \tag{8}$$

The probability function can be explained as follows: If c did not copy ($i \notin \Gamma$), the probability of his/her response given the one by s is just the regular marginal response probability in (5) or (6). On the other hand, if c did copy ($i \in \Gamma$), the responses match and the probability of u_{ci} given u_{si} is equal to one. The event of c coping with responses u_{ci} and u_{si} that do not match is treated as impossible (that is, we exclude the unlikely case of c making a writing error when copying); its probability is therefore equal to zero.

Applying the standard IRT assumption of conditional independence, the joint probability of the entire response vectors $\mathbf{u}_c = (u_{ci})$ and $\mathbf{u}_s = (u_{si})$ may be written as

$$\begin{aligned} p(\mathbf{u}_c, \mathbf{u}_s \mid \theta_c, \theta_s, \Gamma) &= p(\mathbf{u}_c \mid \theta_c, \Gamma, \mathbf{u}_s) p(\mathbf{u}_s \mid \theta_s) \\ &= \prod_{i=1}^{n} p(u_{ci} \mid \theta_c, \Gamma, u_{si}) \prod_{i=1}^{n} p(u_{si} \mid \theta_s). \end{aligned} \tag{9}$$

Consider the posterior probability of c not copying any of the items; that is, $\Gamma = \varnothing$. Combining (4) and (9), the probability is equal to

$$\begin{aligned} p(\varnothing \mid \theta_c, \theta_s, \mathbf{u}_c, \mathbf{u}_s) &= \frac{p(\varnothing) \prod_{i=1}^{n} p(u_{ci} \mid \theta_c, \varnothing, u_{si}) \prod_{i=1}^{n} p(u_{si} \mid \theta_s)}{\sum_{\Gamma} p(\Gamma) \prod_{i=1}^{n} p(u_{ci} \mid \theta_c, \Gamma, u_{si}) \prod_{i=1}^{n} p(u_{si} \mid \theta_s)} \\ &= \frac{p(\varnothing) \prod_{i=1}^{n} p(u_{ci} \mid \theta_c)}{\sum_{\Gamma} p(\Gamma) \prod_{i=1}^{n} p(u_{ci} \mid \theta_c, \Gamma, u_{si})} \\ &= \frac{p(\varnothing) \prod_{i=1}^{n} p(u_{ci} \mid \theta_c)}{\sum_{\Gamma \subseteq M} p(\Gamma) \prod_{i=1}^{n} p(u_{ci} \mid \theta_c, \Gamma, u_{si})}. \end{aligned} \tag{10}$$

Remember that Γ is the hypothetical set of items that were actually copied. Consequently, for any Γ not a subset of M, it holds that $u_{ci} \neq u_{si}$ for some item, and therefore that $p(u_{ci} \mid \theta_c, \Gamma, u_{si}) = 0$; see (8).

Observe that the response probabilities for the source in the numerator and denominator have canceled; the only probabilities on which the expression depends are the response probabilities for the copier. The expression can be further simplified by distributing the product operator over the items in set M and its complement, $\overline{M}$, the result being

$$p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s) = \frac{p(\varnothing)\prod_{i\in M} p(u_{ci} \mid \theta_c,)\prod_{i\in \overline{M}} p(u_{ci} \mid \theta_c)}{\sum_{\Gamma\subseteq M} p(\Gamma)\prod_{i\in M} p(u_{ci} \mid \theta_c, \Gamma, u_{si})\prod_{i\in \overline{M}} p(u_{ci} \mid \theta_c)}$$
$$= \frac{p(\varnothing)\prod_{i\in M} p(u_{ci} \mid \theta_c)}{\sum_{\Gamma\subseteq M} p(\Gamma)\prod_{i\in M} p(u_{ci} \mid \theta_c, \Gamma, u_{si})}. \tag{11}$$

Thus, the posterior probability of c not copying does not depend on the responses to any of the items outside the set with matching responses, M. Further, although prior distribution $p(\Gamma)$ still needs to be specified across all possible sets Γ to be a proper distribution, none of the probabilities beyond $\Gamma \subseteq M$ is actually used.

The posterior odds of c cheating on at least one of the items in N are equal to

$$\frac{1 - p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)} = \frac{\sum_{\substack{\Gamma\subseteq M\\ \Gamma\neq\varnothing}} p(\Gamma)\prod_{i\in M} p(u_{ci} \mid \theta_c, \Gamma, u_{si})}{p(\varnothing)\prod_{i\in M} p(u_{ci} \mid \theta_c)}. \tag{12}$$

Distributing the product operator in the numerator of (12) over the items in the set with copied answers Γ and its complement relative to M, we obtain

$$\frac{1 - p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)} = \frac{\sum_{\substack{\Gamma\subseteq M\\ \Gamma\neq\varnothing}} p(\Gamma)\prod_{i\in\Gamma} p(u_{ci} \mid \theta_c, \Gamma, u_{si})\prod_{i\in M\setminus\Gamma} p(u_{ci} \mid \theta_c, \Gamma, u_{si})}{p(\varnothing)\prod_{i\in M} p(u_{ci} \mid \theta_c)}$$
$$= \frac{\sum_{\substack{\Gamma\subseteq M\\ \Gamma\neq\varnothing}} p(\Gamma)\prod_{i\in\Gamma} p(u_{ci} \mid \theta_c, \Gamma, u_{si})\prod_{i\in M\setminus\Gamma} p(u_{ci} \mid \theta_c)}{p(\varnothing)\prod_{i\in M} p(u_{ci} \mid \theta_c)}$$
$$= \frac{\sum_{\substack{\Gamma\subseteq M\\ \Gamma\neq\varnothing}} p(\Gamma)\prod_{i\in M\setminus\Gamma} p(u_{ci} \mid \theta_c)}{p(\varnothing)\prod_{i\in M} p(u_{ci} \mid \theta_c)}, \tag{13}$$

where the last step follows from the fact that the conditional response probabilities for the items with matching responses in Γ are equal to one [see (8)]. Observe that the odds of c cheating are not directly dependent on the ability of s. The only relationship between c and s exists through their observed set of matching responses, M.

Now assume that the prior probabilities of copying are independent probabilities γ_i, $i = 1, \ldots, n$, across the items in N. Of course, their values should be chosen to be consistent with the desired prior probability of no cheating on the entire set of items; that is, it should hold that $p(\varnothing) = \prod_{i\in N}(1 - \gamma_i)$. Although this seems a reasonable choice, the alternative of some of the prior probabilities being dependent deserves study. For instance, it may be argued that items on the same page or varying substantially in difficulty require a different specification of prior probabilities. Abbreviating the notation of the regular response probabilities $p(u_{ci} \mid \theta_c)$ as p_{ci}, the posterior odds of cheating can be written as

$$\frac{1 - p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)} = \frac{\sum_{\substack{\Gamma\subseteq M\\ \Gamma\neq\varnothing}} \prod_{i\in\Gamma}\gamma_i \prod_{i\in N\setminus\Gamma}(1-\gamma_i)\prod_{i\in M\setminus\Gamma} p_{ci}}{\prod_{i\in N}(1-\gamma_i)\prod_{i\in M} p_{ci}}$$
$$= \frac{\prod_{i\in N\setminus M}(1-\gamma_i)\sum_{\substack{\Gamma\subseteq M\\ \Gamma\neq\varnothing}} \prod_{i\in\Gamma}\gamma_i \prod_{i\in M\setminus\Gamma}(1-\gamma_i)p_{ci}}{\prod_{i\in N}(1-\gamma_i)\prod_{i\in M} p_{ci}}$$
$$= \frac{\sum_{\substack{\Gamma\subseteq M\\ \Gamma\neq\varnothing}} \prod_{i\in\Gamma}\gamma_i \prod_{i\in M\setminus\Gamma}(1-\gamma_i)p_{ci}}{\prod_{i\in M}(1-\gamma_i)p_{ci}}. \tag{14}$$

It is illuminating to analyze the combinatorial nature of this expression. First, observe that $(1-\gamma_i)p_{ci}$ is the product of the prior probability of no copying on item i and the regular response probability of c for this item. The product is proportional to the posterior probability of c not having copied on item i. Likewise, $\gamma_i * 1$ is the product of the prior probability of c copying on item i and the probability of c having the same response as s when copying [see (8)]. Thus, γ_i is proportional to the posterior probability of c having copied on i. As (14) is the ratio of two posterior probabilities, the absence of norming does not matter. It follows that (14) equals the ratio of (i) the sum of the posterior probabilities associated with all possible combinations of copying on at least one item in M and not copying on any of its other items (numerator) and (ii) no cheating on any of the items in M at all (denominator).

In order to prepare efficient computation of (14), let

$$\xi_{ci} \equiv (1-\gamma_i)p_{ci}. \tag{15}$$

Renumbering the items so that items $1, 2, \ldots, m$ are in M and $m+1, \ldots, n$ are in $N\backslash M$, the expression can be written as

$$\frac{1-p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \theta_c, \mathbf{u}_c, \mathbf{u}_s)} = \frac{\boldsymbol{\gamma}'\boldsymbol{\xi}_c - \prod_{i\in M}\xi_{ci}}{\prod_{i\in M}\xi_{ci}}, \tag{16}$$

where $\boldsymbol{\gamma} = (1, \boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(m)})$ and $\boldsymbol{\xi}_c = (\boldsymbol{\xi}_c^{(m)}, \boldsymbol{\xi}_c^{(m-1)}, \ldots, \boldsymbol{\xi}_c^{(1)}, 1)$ are column vectors of length 2^m that have components

$$\begin{aligned}
\boldsymbol{\gamma}^{(1)} &= (\gamma_1, \gamma_2, \ldots, \gamma_m);\\
\boldsymbol{\gamma}^{(2)} &= (\gamma_1\gamma_2, \gamma_1\gamma_3, \ldots, \gamma_{m-1}\gamma_m);\\
&\vdots\\
\boldsymbol{\gamma}^{(m)} &= (\gamma_1\gamma_2\ldots\gamma_m)
\end{aligned} \tag{17}$$

and

$$\begin{aligned}
\boldsymbol{\xi}_c^{(m)} &= (\xi_{c1}\xi_{c2}\ldots\xi_{cm})\\
\boldsymbol{\xi}_c^{(m-1)} &= (\xi_{c2}\xi_{c3}\ldots\xi_{cm}, \xi_{c1}\xi_{c3}\ldots\xi_{cm}, \ldots, \xi_{c1}\xi_{c2}\ldots\xi_{c(m-1)});\\
&\vdots\\
\boldsymbol{\xi}_c^{(1)} &= (\xi_{c1}, \xi_{c2}, \ldots, \xi_{cm}),
\end{aligned} \tag{18}$$

respectively.

An example of all possible products of γ_i and ξ_{ci} involved in the calculation of (16) for a copier and source with $m = 3$ matching responses is given in Table 1. The first column lists all possible combinations of items, while the second and third column contain the components of the $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ vectors, respectively, in the order in which they are defined in (16)–(18). The last column in the table lists the component-wise products of $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$. This column enables us to calculate the posterior odds; its first entry is the denominator $\prod_{i\in M}\xi_{ci}$ of (16) while the sum of all other entries is its numerator.

The structure of Table 1 reminds us of the generalized (or compound) binomial probability distribution of the number-correct score on a test of m items by a test taker with ability θ_c. The only

TABLE 1.
Example of all products of γ_i and ξ_{ci} in the posterior odds in Eq. 16 ($m = 3$).

Items	Products of γ_i and ξ_{ci}		
	$\boldsymbol{\gamma}$	$\boldsymbol{\xi}_c$	$\boldsymbol{\gamma}'\boldsymbol{\xi}_c$
$\varnothing$	1	$\xi_{c1}\xi_{c2}\xi_{c3}$	$\xi_{c1}\xi_{c2}\xi_{c3}$
{1}	γ_1	$\xi_{c2}\xi_{c3}$	$\gamma_1\xi_{c2}\xi_{c3}$
{2}	γ_2	$\xi_{c1}\xi_{c3}$	$\xi_{c1}\gamma_2\xi_{c3}$
{3}	γ_3	$\xi_{c1}\xi_{c2}$	$\xi_{c1}\xi_{c2}\gamma_3$
{1,2}	$\gamma_1\gamma_2$	ξ_{c3}	$\gamma_1\gamma_2\xi_{c3}$
{1,3}	$\gamma_1\gamma_3$	ξ_{c2}	$\gamma_1\xi_{c2}\gamma_3$
{2,3}	$\gamma_2\gamma_3$	ξ_{c1}	$\xi_{c1}\gamma_2\gamma_3$
{1,2,3}	$\gamma_1\gamma_2\gamma_3$	1	$\gamma_1\gamma_2\gamma_3$

thing we have to do to obtain this distribution is replace ξ_{ci} and γ_i by the response probabilities p_{ci} and $q_{ci} \equiv 1 - p_{ci}$, respectively. (Of course, ξ_{ci} and γ_i do not sum to one, but this is not a problem here.) The first entry in the last column then becomes the probability of $x = 3$ items correct, the sum of the entries for {1}, {2}, and {3} becomes the probability of $x = 2$ items correct, etc., all the way down to the last entry, which now is the probability of $x = 0$ items correct. Continuing the analogy, it follows that the denominator of (16) corresponds to the probability of a number-correct score equal to $x = m$, whereas its numerator corresponds to the sum of the probabilities of the number-correct scores from $x = 0$ through $m - 1$. An efficient recursive algorithm for the calculation of the posterior odds of cheating suggested by this analogy is presented in Appendix 2.

2.1.1. Alternative Models The K-index procedure for detecting answer copying is designed for a test with multiple-choice items that are scored dichotomously. It assumes the sampling of c from a population of independently operating test takers with c's number of wrong answers. The sampling distribution considered is for the number of matching incorrect alternatives conditional on the items s had wrong. This distribution is taken to be binomial, with a success parameter p_{cs} obtained by a conservative piecewise linear model for the regression of the proportion of matches to the incorrect answers by s on the proportion of incorrect answers for the entire population of test takers. Let w_c denote the proportion of incorrect answers by c. Formally, the success parameter is defined as

$$p_{cs} = \begin{cases} 0.085 + \beta w_c, & \text{if} \quad 0.0 < w_c < 0.3, \\ 0.085 + 0.3\beta + 0.4\beta(w_c - 0.3), & \text{if} \quad 0.3 \leq w_c < 1.0, \end{cases} \tag{19}$$

where β is estimated empirically for each population and test. For a motivation and further details of the binomial model with this success parameter, see Holland (1996) and Lewis & Thayer (1998). The K-index is defined as the upper tail probability for this distribution, evaluated at the observed number of matching incorrect alternatives for c and s.

As a result of the conditioning on the items s has wrong, we need to redefine our notation. Again, N denotes the set of items that is considered, but its choice now is restricted to the items s had wrong. In addition, M is the subset of N with matching incorrect alternatives by c and s, and Γ the subset of M for which c copied the answers. Likewise, u_{si} now represents the incorrect alternative for item i chosen by s, and the prior probabilities γ_i are for the event of c copying the incorrect response to item i by s. The conditional probability of response $U_{ci} = u_{ci}$ given $U_{si} = 0$ is equal to

$$p(u_{ci} \mid \Gamma, u_{si}) = \begin{cases} p_{cs}, & \text{if} \quad i \notin \Gamma, \\ 1, & \text{if} \quad i \in \Gamma \text{ and } u_{ci} = u_{si}, \\ 0, & \text{if} \quad i \in \Gamma \text{ and } u_{ci} \neq u_{si}. \end{cases} \tag{20}$$

Analogously to (16), the posterior odds of c not cheating under the model for the K-index are equal to

$$\frac{1 - p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)} = \frac{\boldsymbol{\gamma}' \boldsymbol{\xi}_c - \prod_{i \in M} \xi_{csi}}{\prod_{i \in M} \xi_{csi}}, \tag{21}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}_c$ are vectors with the same format as in (17)–(18) but the latter now has entries

$$\xi_{csi} \equiv (1 - \gamma_i) p_{cs}. \tag{22}$$

The odds can be calculated using the algorithm in the Appendix 2 with ξ_{ci} in (15) replaced by ξ_{csi} in (22).

The test of answer copying based on a binomial test (van der Linden & Sotaridona 2004) is also for dichotomously scored multiple-choice items. Although, like the K-index, its null distribution is derived from the binomial, it does not assume any sampling of c from a population of independently operating test takers. Instead, the result follows from the assumption of three alternative response processes for an individual test taker: First, if test takers know the correct answer, it is assumed that they will give it. Second, if test takers do not know the answer, they may copy it from a neighbor. Third, if test takers do not want to copy or have no access to the answers by any of the neighbors, they will guess randomly. In fact, the response process is the one of the familiar knowledge-or-random guessing model extended with the option of copying.

The binomial test also conditions on the items s has wrong. Continuing our new notation, the relevant response probabilities are

$$p(u_{ci} \mid \Gamma, u_{si}) = \begin{cases} (k-1)^{-1}, & \text{if} \quad i \notin \Gamma, \\ 1, & \text{if} \quad i \in \Gamma \text{ and } u_{ci} = u_{si}, \\ 0, & \text{if} \quad i \in \Gamma \text{ and } u_{ci} = u_{si}, \end{cases} \tag{23}$$

where k is the number of alternatives. It follows immediately that the posterior odds of cheating are equal to

$$\frac{1 - p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)} = \frac{\boldsymbol{\gamma}' \boldsymbol{\xi} - \prod_{i \in M} \xi_i}{\prod_{i \in M} \xi_i}, \tag{24}$$

with $\boldsymbol{\xi}_c$ replaced by $\boldsymbol{\xi}$, which has entries

$$\xi_i \equiv (1 - \gamma_i)(k-1)^{-1}. \tag{25}$$

The odds can be calculated using the algorithm in Appendix 2 with p_{ci} in (15) replaced by $(k-1)^{-1}$.

In the empirical examples below, we will address the case of a proctor who is positive of communication between c and s while working on a section of the test but unable to be more specific. The case can be represented by equal prior probabilities of copying on the individual items; that is, the choice of $\gamma_i = \gamma$, $i = 1, \ldots, n$, with γ a positive number following from $p(\varnothing) = (1 - \gamma)^n$. For checks based on the K-index and the binomial test above, the posterior odds in (21) and (24) simplify to

$$\frac{1 - p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)} = \sum_{g=1}^{m} \binom{m}{g} \left(\frac{\gamma}{(1 - \gamma) p_{cs}} \right)^g, \tag{26}$$

and

$$\frac{1-p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)}{p(\varnothing \mid \mathbf{u}_c, \mathbf{u}_s)} = \sum_{g=1}^{m} \binom{m}{g} \left(\frac{\gamma(k-1)}{1-\gamma}\right)^g, \tag{27}$$

respectively, where g denotes the size of set Γ. The simplification does not hold for response models with item parameters, such as the ones in (5) and (6)–(7). Although it seems tempting to use (26) as a rough approximation to (14), with the average of p_{ci} across the items substituted for p_{cs}, the result will be a considerable loss of power. We will illustrate this point later.

2.2. *Fraudulent Erasures*

The detection of fraudulent erasures of wrong answers on answer sheets requires a model for the probability of a regular test taker changing an answer on a multiple-choice item found to be incorrect during review. The proposed model is based on a two-stage response process: In the first stage, the test taker produces initial answers to the items in the test. Once the answers have been given, the second stage begins, in which (s)he reviews all answers, confirming the ones still thought to be okay but changing those for which a better alternative seems available. Of course, a test taker could change the answers more than once, but we focus on the final response. Observe that, as a result of optical scanning of the answer sheet, we know both the erased initial and this final responses for each item.

The assumption of this two-stage process can only hold when the test is not speeded and the test taker thus has enough testing time to complete the reviews. Erasures may also occur through other non-fraudulent processes, for instance, when a test taker discovers systematic misreading of the item numbers on the answer sheets. Such cases have to be excluded by additional analysis and/or direct inspection of the answer sheets.

Suppose the test consists of items that have been pretested and calibrated under the 3PL model in (7). It then makes sense to assume probabilities of the first-stage responses that follow the same model, which we now denote as

$$p(1 \mid \theta) \equiv c_i + (1-c_i)\frac{\exp[a_i(\theta^{(1)} - b_i)]}{1+\exp[a_i(\theta^{(1)} - b_i)]}, \tag{28}$$

where a_i, b_i, and c_i are the item parameters estimated during item calibration. The test taker's ability $\theta^{(1)}$ is estimated from the initial responses. Because of the lack of identifiability of the model, in order to capture the change in response probability due to the effects of the initial responses, we need a constrained version of it for the second-stage responses. We therefore fix θ at its initial level, but let the a_i and b_i parameters free to capture the change. In addition, we assume that if the test taker still does not know the answer and already guessed during the initial stage, (s)he will not guess again. The result is

$$p(u_i^{(2)} = 1 \mid \theta^{(1)}, u_i^{(1)} = 0) = \frac{\exp[a_{0i}(\theta^{(1)} - b_{0i})]}{1+\exp[a_{0i}(\theta^{(1)} - b_{0i})]}, \tag{29}$$

and

$$p(u_i^{(2)} = 1 \mid \theta^{(1)}, u_i^{(1)} = 1) = \frac{\exp[a_{1i}(\theta^{(1)} - b_{1i})]}{1+\exp[a_{1i}(\theta^{(1)} - b_{1i})]}, \tag{30}$$

where $u_i^{(1)}$ and $u_i^{(2)}$ are the initial and final responses to item i and (a_{i0}, b_{i0}) and (a_{i1}, b_{i1}) are the parameters for item i given $u_i^{(1)} = 0$ and 1, respectively. The latter are free parameters that can be estimated from the final responses given the initial responses using logistic regression; for

more details of the model and its estimation as well as the generalization to polytomous response models as in (5), see Linden & Jeon (2012).

Observe that (29) is the probability of a WR change. These changes are the ones of interest when there is suspicion of fraudulent behavior by school teachers or administrators; we therefore ignore (30).

Let E_i be a binary variable indicating whether ($E_i = 1$) or not ($E_i = 0$) an observed change on item i was a WR change. As before, N denotes the set of items considered, $M \subseteq N$ the subset of items with an observed WR erasure, and $\Gamma \subseteq M$ its subset of fraudulent erasures. Prior probability γ_i is for the event of a fraudulent erasure on item i. The probability of $E_i = e_i$ is

$$p(e_i \mid \Gamma) = \begin{cases} p(u_i^{(2)} = 1 \mid \theta^{(1)}, u_i^{(1)} = 0), & \text{if} \quad i \notin \Gamma \text{ and } e_i = 1, \\ 1 - p(u_i^{(2)} = 1 \mid \theta^{(1)}, u_i^{(1)} = 0), & \text{if} \quad i \notin \Gamma \text{ and } e_i = 0, \\ 1, & \text{if} \quad i \in \Gamma \text{ and } e_i = 1, \\ 0, & \text{if} \quad i \in \Gamma \text{ and } e_i = 0. \end{cases} \tag{31}$$

The second probability is only given for completeness; our current focus is exclusively on WR changes, which have the first probability. The last two probabilities reflect the assumptions of the teacher/administrator changing incorrect answers only and always replacing them by the correct answer.

Following the same argument as before, the posterior odds of cheating are

$$\frac{1 - p(\varnothing \mid \mathbf{e})}{p(\varnothing \mid \mathbf{e})} = \frac{\boldsymbol{\gamma}'\boldsymbol{\xi} - \prod_{i \in M} \xi_i}{\prod_{i \in M} \xi_i}, \tag{32}$$

where vector $\boldsymbol{\xi}$ has entries

$$\xi_i \equiv (1 - \gamma_i)\, p(u_i^{(2)} = 1 \mid \theta^{(1)}, u_i^{(1)} = 0). \tag{33}$$

Again, the odds can easily be calculated using the algorithm in Appendix 2 with p_{ci} in (15) replaced by $p(u_i^{(2)} = 1 \mid \theta^{(1)}, u_i^{(1)} = 0)$ in (29).

3. A Few Numerical Examples

As each of the proposed checks implies an expression for its posterior odds differing only in the probabilities of the specific type of cheating behavior on the items it addresses, we restrict our examples to the fully developed case of answer copying under a known response model in (5)–(16). All examples are for the item parameters for the nominal response model in (5) for the same set of 40 items as used in Wollack (1997). Each of the items had five answer choices.

Figure 1 highlights the probabilistic structure of random matches between pairs of test takers on the entire test derived from the item parameters only. Each of the plots shows the distribution of the number of random responses for the typical case of a lower-ability test taker suspected to copy from a more able one we may have to check on in practice: $(\theta_c, \theta_s) = (-2.0, 1.0)$, $(-1.5, 1.0)$, $(-1.0, 1.0)$, and $(-0.5, 1.0)$. The distributions are known to be generalized or compound binomial and can be generated from the probabilities of a random match on the individual items using another version of the recursive algorithm discussed in Appendix 2 (for details, see van der Linden & Sotaridona 2006). For each item, the probability of a random match was calculated according to (1) (i.e., without any response data). The average probabilities for the four pairs of

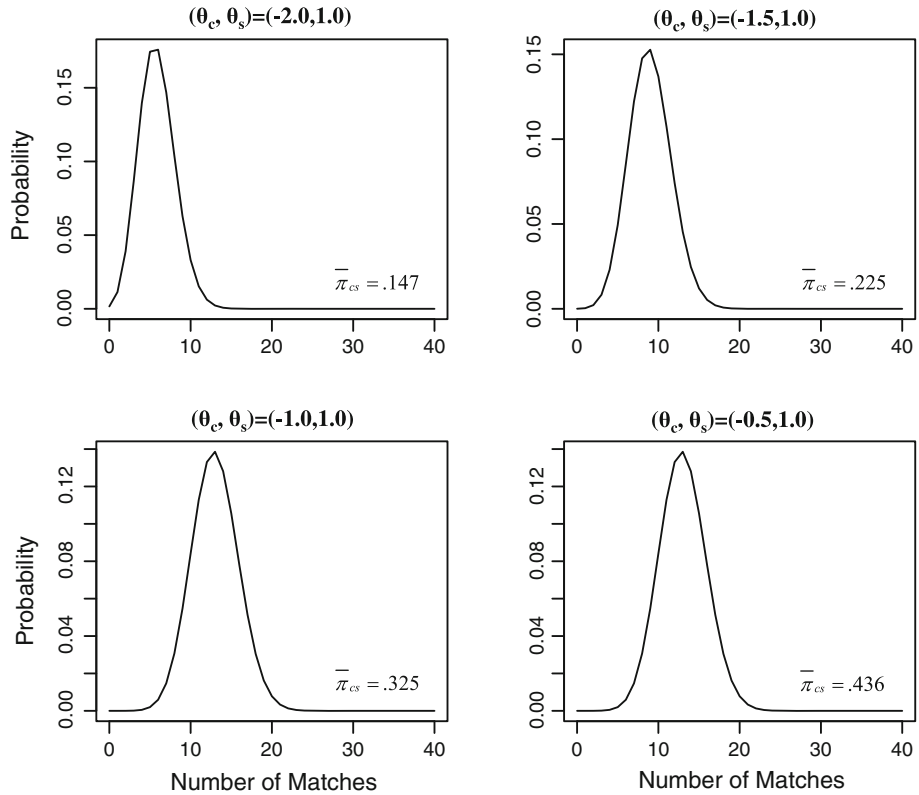

FIGURE 1.
Distributions of random matches between the responses of a hypothetical copier and source on a test of 40 five-choice items for four different combinations of ability levels. *Note* lower-right corner of each plot shows the average probability of a match on an item.

test takers across all items were equal to $\overline{\pi}_{cs} = .147$, .225, .325, and .436, respectively. As the distributions in Figure 1 confirm, it is not unlikely for two test takers to work independently and nonetheless produce substantial numbers of matching responses. The problem of separating these cases from fraudulent matches is our statistical challenge.

Suppose the test consists of different sections and we have prior evidence of c copying some of the answers from s in the first section of ten items, for instance, in the form of a proctor who has observed suspicious communications between the pair of test takers. The proctor is positive that the communications occurred while c and s were working on this section but is unable to be more specific. The strength of the evidence can be measured in the form of a prior probability of c having copied at least one of the answers from s in the section; that is, the specification of the prior probability $1 - p(\varnothing)$. Since the proctor is ignorant as to the specific items on which copying might have occurred, it seems natural to adopt a constant prior probability of c having copied on any of them; that is, $\gamma_i = \gamma, i = 1, \ldots, 10$. For the general case of n items, probability calculus allows us to derive γ as

$$\gamma = 1 - (1 - p(\varnothing))^{-n}. \tag{34}$$

A set of response data was generated from the item parameters in the first section for the same four ability pairs (θ_c, θ_s) as in Figure 1. The number of observed matches between c and s happened to be equal to two for all four pairs of θ values. In addition, for each of these pairs we simulated different levels of answer copying by systematically replacing responses by c with those by s for the items that did not have a random match, beginning with the first non-matching

TABLE 2.
Posterior odds given numbers of fraudulent matches between pairs of test takers on a section of ten items in the test as a function of their ability levels and the prior probability of at least one fraudulent match.

(θ_c, θ_s)	(−2.0,1.0)			(−1.5,1.0)		
Prior odds	1/3	1	3	1/3	1	3
No. of matches						
2	**0.62**	**1.71**	**4.28**	**0.31**	**0.82**	**1.98**
3	2.04	7.56	27.58	0.41	1.18	3.18
4	2.42	10.51	45.50	0.76	2.52	8.50
5	3.15	16.00	444.05	0.90	3.20	12.31
6	6.06	45.44	681.63	1.62	7.11	38.02
7	6.82	57.40	*	4.48	28.91	255.20
8	13.96	188.98	*	5.12	37.47	407.13
9	15.59	239.49	*	12.89	157.66	*
10	19.85	391.38	*	15.11	219.86	*
(θ_c, θ_s)	(−1.0,1.0)			(−0.5,1.0)		
No. of matches						
2	**0.19**	**0.50**	**1.15**	**0.12**	**0.31**	**0.69**
3	0.67	1.99	5.56	0.20	0.53	1.27
4	0.98	3.36	11.78	0.34	0.98	2.65
5	1.14	4.18	16.82	0.47	1.43	4.38
6	1.38	5.65	27.56	0.63	2.11	7.50
7	1.63	7.36	42.35	0.75	2.65	10.55
8	8.97	64.74	658.47	0.85	3.18	14.02
9	9.56	74.34	858.08	0.94	3.68	17.70
10	11.10	101.38	*	1.12	4.72	26.35

Bold numbers are for random matches. *Asterisks* indicate posterior odds larger than 1,000.

item in our files. Three levels of the prior probability were analyzed: $1 - p(\varnothing) = .25$, .50., and .75. Observe that these choices amount to prior odds of copying on at least one item equal to 1:3, 1:1, and 3:1, respectively; that is, a case of weak belief of cheating, indifference between cheating and no cheating, and strong belief of cheating. For each of the numbers of matches $m = 2, ..., 10$, the posterior odds of answer copying were calculated using the algorithm in Appendix 2. Table 2 shows the results for all simulated conditions.

Several things can be observed from these examples. First, as expected, the posterior odds increased with the numbers of fraudulent answer changes. They happen to do so at a different rate for each of the simulated conditions, though. One of the reasons for these different rates is the dependence of the odds on the actual responses simulated for the source, which, obviously, were different for each simulated condition. Second, not surprisingly, the posterior odds did increase with the prior odds of cheating. In fact, moving from the prior odds of 3:1 against cheating to 3:1 in favor of it—a change by a factor equal to nine—had a strong effect on the posterior odds for each given number of matches. Finally, the posterior odds showed a tendency to decrease as the difference between θ_c and θ_s decreases. Again, this was as expected: The closer the abilities of the copier and source, the more likely a random match on any of the items, and the Bayesian response to this is lower posterior odds.

4. Discussion

Although the tendencies in Table 2 are clear, they are for twelve specific conditions only. We therefore should not generalize without further study.

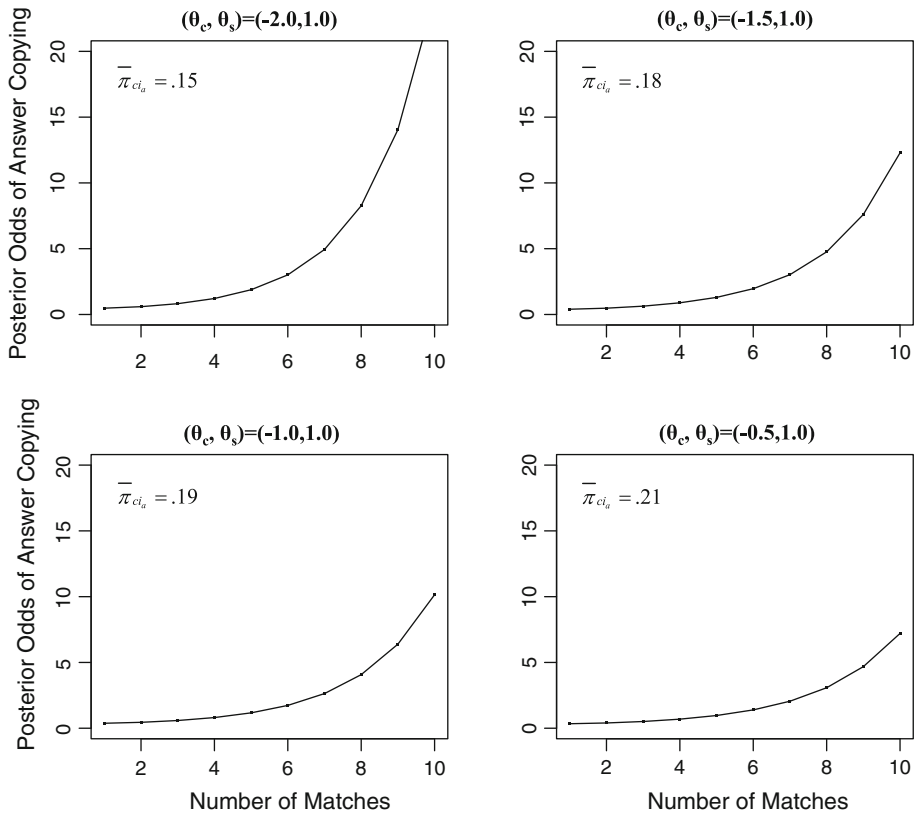

FIGURE 2.
Posterior odds of answer copying as a function of the number of matches according to Eq. 26, with the probability of a random match on an item calculated as the average of Eq. 2 across all items in the first section of the test. *Note* upper-left corner of each plot shows the average probability of Eq. 2 across the items.

In fact, one of the dominant impressions derived from our current set of examples is their data dependency. For instance, for the case of prior indifference, if we had accepted posterior odds greater than five as evidence of answer copying, the critical numbers of matches given the responses for each of the four ability pairs would have been two, five, five, and ten matches (Table 2, middle columns), respectively. Such differences should not come as a surprise, though; in a Bayesian approach, all inferences are conditional on the data.

Our earlier derivation of the posterior odds allows us to be more specific about the factors that have an impact on them. As demonstrated by our derivation of (10), in order to detect answer copying, the responses by c and s outside the subset of items with a match, M, appear to be redundant. On the other hand, the odds of having copied critically depend on the specific alternatives chosen by the copier. Consequently, in order to account for the likelihood of a random match with the source on them, we need to know the parameters that control the copier's response probabilities for these alternatives; that is, both the pertinent item parameters and ability parameter θ_c. Our derivations also reveal that the ability of the source, θ_s, does not matter (Eq. 13)—a fact that makes perfect sense. The only thing that counts when c copies from s are the actual responses by the latter; it is unnecessary to know the ability that generated these responses.

If any of the relevant quantities is left out of the equation, the posterior odds are miscalculated. Two cases in point are presented in Figures 2 and 3. Both figures are for the same indifference prior odds of 1:1 used for each of the four ability pairs in Table 2 (middle columns). Figure 2

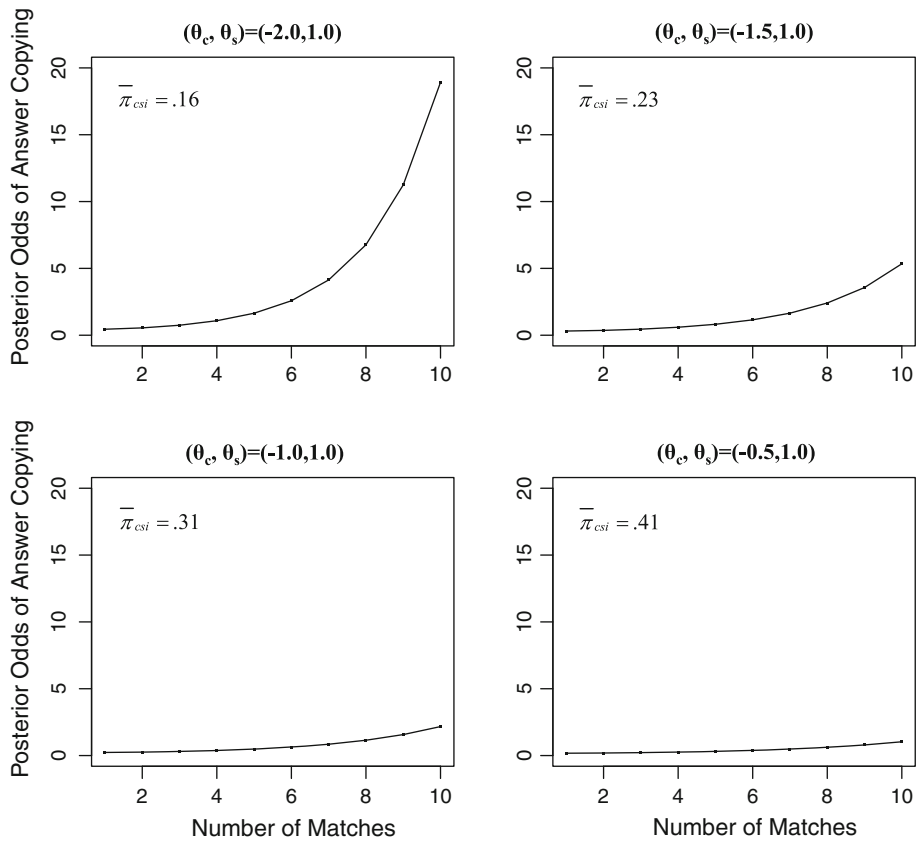

FIGURE 3.
Posterior odds of answer copying as a function of the number of matches according to Eq. 26, with the probability of a random match on an item calculated as the average of Eq. 1 across all items in the first section of the test. *Note* upper-left corner of each plot shows the average probability of Eq. 1 across the items.

presents the posterior odds of copying as a function of the number of matches, with the probability of a match calculated as the average of the probabilities $\pi_{ci_{a'}}$ in (2) across the same ten items and for the same responses by the source as in the table. The curves follow directly upon substitution of the average probability for p_{cs} in (26). This case thus amounts to intentional ignorance as to the differences between the items and their alternatives while still conditioning on the responses by the source. Observe that, with a slight exception for $(\theta_c, \theta_s) = (-0.5, 1.0)$, the odds of copying in Figure 2 are much smaller than in Table 2. As predicted earlier, the effect of ignoring relevant statistical information on the differences between the items and their alternatives thus tends to result in loss of power, in the current Bayesian context in the form of much less discrimination between the odds associated with the lower and higher numbers of matches. The curves in Figure 3 are based on the average probabilities of π_{csi} in (1); that is, when the conditioning on the source's responses is omitted as well. The effects of the additional ignoring of these responses imply even larger loss of power.

Again, we recommend not to use the Bayesian checks in this paper routinely for all test takers but only when prior empirical evidence suggests a specific type of cheating by a test taker on a specific set of items in the test. This restriction has two advantages. First, it automatically entails the formulation of prior probabilities for the items involved in the check which are much more informative than the weak prior probabilities following from the routine use of (34) for full-length

tests. Second, use of the checks for a smaller subset of items enables us to estimate θ_c , the only unknown quantity in (14), from the responses to the other items in the test. Although the use of a simple plug-in estimate of θ_c would be computationally convenient, its estimation error would propagate in the calculation of the posterior odds. The Bayesian way to proceed is to introduce a prior distribution for θ_c as well and account for any remaining uncertainty about this parameter by integrating it out of the posterior quantities of interest. The questions of how to do so and how serious the impact on the posterior odds would be will be subject of subsequent research.

Our final comment is on the combinatorial nature of the posterior odds of cheating in (5)–(16). Naively, one might have expected just an application of the rather straightforward "posterior odds = likelihood ratio × prior odds" factorization explained in most introductory texts to Bayesian statistics. However, this type of factorization only holds for the case of a statistical model for identically distributed independent variables with common parameters. In the current context, it would apply only if all items could be treated as exchangeable. However, as just demonstrated by the differences between the results in Table 2 and Figures 2–3, test items are not exchangeable. They differ substantially in their properties, and therefore considerably in their probabilities of a random match between independently working test takers. Consequently, in order to derive the posterior odds for an observed number of matches, we have to evaluate all possible combinations of cheating on at least one item in M and no cheating on the rest. Fortunately, the algorithm in Appendix 2 enables us to deal with this job efficiently.

Appendix 1: Conditional and Unconditional Tests of Answer Copying

The experiment of c and s responding to the same test can be conceived of as a two-stage process, in which we first observe response vector $\mathbf{U}_s = \mathbf{u}_s$ by s and then $\mathbf{U}_c = \mathbf{u}_c$ by c given the former. For known item parameters, the two vectors have distributions $\mathbf{u}_s \sim F_{\theta_s}$ and $\mathbf{u}_c \sim F_{\theta_c}$ depending on the ability parameters θ_s and θ_c, respectively. Given $\mathbf{U}_s = \mathbf{u}_s$, the observation of $\mathbf{U}_c = \mathbf{u}_c$ is equivalent to that of the number of matches $M_{cs} = m_{cs}$ between c and s, with $M_{cs} \sim F_{\gamma_{cs}}$, where γ_{cs} is the unknown number of items copied by c (see the main text for examples of the distribution of $F_{\gamma_{cs}}$).

Conditional inference of γ_{cs} from $\mathbf{U}_s = \mathbf{u}_s$ has expected power equal to inference from the observed joint distribution of $\mathbf{U}_c$ and $\mathbf{U}_s$, when $\mathbf{U}_s$ is partially ancillary with respect to γ_{cs}, that is, (i) its distribution, F_{θ_s}, does not depend on γ_{cs} and (ii) γ_{cs} and θ_s are distinct (Lehmann & Romano 2005, Sect. 10.1–10.2). In the current application, both conditions are fulfilled.

Appendix 2: Calculation of Posterior Odds

The distribution of the number-correct score on a test, which is known to be generalized or compound binomial, lacks a closed-form probability function, but its probabilities are easily calculated using a well-known recursive algorithm introduced in the test-theory literature by Lord & Wingersky (1984). The earlier analogy between these probabilities and the entries in the last column of Table 2 is the rationale for the following modification of the algorithm, which can be used to calculate the posterior odds in (16).

Let z denote the size of the relative complement $M \backslash \Gamma$, that is, the subset of items in M for which the answers were not copied. In addition, we use $\pi_m(z)$ to represent the sum of all possible joint products of the z quantities ξ_{ci} for the items in $M \backslash \Gamma$ (but remember that these are not probabilities) and the $m - z$ prior probabilities γ_i for the items in Γ [for example, for $x = 1$, the sum of products in the last column of Table 1 for the rows {1}, {2}, and {3}]. Finally, we define $\pi_m(z) = 0$ for $z < 0$ and $z > m$.

Beginning with the first item ($t = 1$), the algorithm runs through steps $t = 2, ..., m$, each time adding one extra item to the set that is considered:

1. For $t = 1$, set $\pi_1(0) = \gamma_1$ and $\pi_1(1) = \xi_{c1}$;
2. For $t = 2, \ldots, m$ and $z = 0, \ldots, t$ calculate the values of $\pi_t(z)$ from the recursive relation

$$\pi_t(z) = \begin{cases} \gamma_t \pi_{t-1}(0), & \text{for} \quad z = 0 \\ \gamma_t \pi_{t-1}(z) + \xi_{ct} \pi_{t-1}(z-1), & \text{for} \quad z = 1, \ldots, t-1, \\ \xi_{ct} \pi_{t-1}(t-1), & \text{for} \quad z = t. \end{cases}$$

3. Finally, calculate the posterior odds in (16) as

$$\frac{1 - p(\varnothing \mid \theta_c, \mathbf{u}_c)}{p(\varnothing \mid \theta_c, \mathbf{u}_c)} = \frac{\sum_{z=0}^{m-1} \pi_m(z)}{\pi_m(m)}. \tag{35}$$

Observe that, unlike the original version of the algorithm for use with number-correct score distributions, its current modification can be used for dichotomous and polytomous response models alike: The response probabilities p_{ci} that figure in $\xi_{ci} \equiv (1-\gamma_i)\, p_{ci}$ are just the probabilities for the copier's actual responses to the items in M, no matter whether the items are scored dichotomously under models as in (7), polytomously under models as in (5), or as a mixture of both.

References

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

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

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.

Fox, J.-P., & Meijer, R.R. (2008). Using item response theory to obtain individual information from randomized response data: An application using cheating data. *Applied Psychological Measurement*, *32*, 595–610.

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

Glas, C.A.W., & Meijer, R.R. (2003). A Bayesian approach to person-fit analysis in item response theory. *Applied Psychological Measurement*, *27*, 217–233.

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.

Jacob, B.A., & Levitt, S. (2003a). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, *118*, 843–877.

Jacob, B.A., & Levitt, S. (2003b). Catching cheating teachers: The results of an unusual experiment in implementing theory. Brookings-Wharton Papers on Urban Affairs, 185–209.

Jacob, B.A., & Levitt, S. (2004, Winter). To catch a cheat. *Education Next*, 68–75.

Lehmann, E.L., & Romano, J.P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.

Levitt, S., & Rubner, S. (2005). *Freakonomics: A rogue economist explores the hidden side of everything*. New York: Harper Collins.

Lewis, C. (2006). A note on conditional and unconditional hypothesis testing: A discussion of an issue raised by van der Linden and Sotaridona. *Journal of Educational and Behavioral Statistics*, *31*, 305–309.

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., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 452–461.

McLeod, L.D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*, 121–137.

Meijer, R.R., & Sijtsma, K. (2001). Methodology review: Evaluation of person fit. *Applied Psychological Measurement*, *25*, 107–135.

Qualls, A.L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, *20*(1), 9–16.
Saupe, J.L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, *20*, 475–489.
Sotaridona, L.S., & Meijer, R.R. (2002). Statistical properties of the *K*-index for detecting answer copying. *Journal of Educational Measurement*, *39*, 115–132.
van der Linden, W.J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.
van der Linden, W.J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, *37*, 180–199.
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. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, *31*, 283–304.
Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple-choice exams. *Journal of Applied Statistics*, *27*, 909–921.
Wollack, J.A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, *21*, 307–320.

Manuscript Received: 22 OCT 2013
Published Online Date: 11 JUN 2014