Comparison of the L_z* Person-Fit Index and ω Copying-Index in Copying Detection

First Year Paper

July 2014

Lieke Voncken

ANR: 163620

Tilburg University

School of Social and Behavioral Sciences

Department of Methodology and Statistics

Supervisor: dr. J. Tijmstra

Second Reader: dr. W.H.M. Emons

Abstract

The purpose of this study was to compare the l_z* person-fit index and the ω copying-index in the detection of answer copying. We expected the copying-index to perform better than the person-fit index because it is specifically created to detect answer copying. The current study simulated answer copying pairs for each of 20 conditions (2 test length × 5 amount of copying × 2 sample size) with 3 significance levels to compare the Type I error rate and power of the l_z* person-fit index and Wollack's ω copying-index to detect copying behavior. Results indicate that the l_z* index has higher detection rates than the ω index in most conditions. Nevertheless, the conditions in which ω performed better are the conditions with the highest absolute levels of detection for both indices. Empirical type I error rates were far below its nominal level for all conditions of ω and in most conditions of l_z*. However, in some conditions, the empirical type I error rates for l_z* were about equal to or slightly above the nominal level. Directions for future research are discussed.

Comparison of the L_z* and Omega Indices in Cheating Detection

People have to take high-stake tests in their lives many times. For instance, you repeatedly have to complete tests that determine your course grades at school and tests that determine whether you are granted a driver's license or are hired for a job. These high stakes may create an incentive to cheat. There are many different ways how examinees can do this, e.g., they may steal answer sheets, communicate with someone outside the examination room or copy someone else's answers. Studies found that approximately two-third of college students in the U.S. admitted to cheating at least once during college (McCabe, 1993) and 5% will copy answers during any given test administration (Bellezza & Bellezza, 1989; Frary & Olson, 1985). A later study (McCabe, Trevino, & Butterfield, 2001) found that about 60% of the students reported that they had copied at least once from another on a test. In the survey study of Rakovski and Levy (2007), 1.8% of the business college students reported that they had engaged in answer copying during exams more than five times. If we take into account social desirability these numbers may be even higher.

The problem with cheating is that it biases the inferences about the examinee's performance, i.e., the validity of the test is threatened (Meijer, 1997; Schmitt, Chan, Sacco, AIcFarland, & Jennings, 1999). The ability level of cheaters is often estimated higher than the true ability level. Because decisions are often based on test scores, inaccurate ability estimates of the examinees may invalidate high stake decisions. Hence, testing agencies try to prevent cheating by using different versions of a test, minimizing item exposure, prohibiting the use of communication devices and the presence of examination supervisors (Shu, 2010). However, it is not possible to prevent cheating from happening completely. Fortunately, different cheating detection methods are available that allow the detection of cheating after test completion. In this study, we will compare two of these cheating detection methods. More specifically, we will focus on copying behavior because this is a popular way of cheating (Fox

& Meijer, 2008).

Person-Fit Indices

One way to detect cheating is to assess the person-fit. Examinees who cheat may generate responses that are unlikely under an item response theory model. A cheating examinee is likely to answer more difficult questions correctly than would be expected based on his/her ability level. As a result, the cheating examinee might answer easier items incorrectly and more difficult items correctly. In other words, the person does not fit the model. There are many statistics that evaluate person-fit in general (for a review, see Meijer & Sijtsma, 2001). An often-used person-fit index is the l_z statistic (Drasgow, Levine, & Williams, 1985).

Cheating Indices

There are also indices available that are specifically created to detect a particular type of cheating behavior – like copying behavior. Often used copying detection indices are the *B*- and *H*-index (Angoff, 1974), *K, K1* and *K*2 (Sotaridona & Meijer, 2003), *PAIR1* and *PAIR2* (Hanson, Harris, & Brennan, 1987), *S1* and *S2* (Sotaridona & Meijer, 2003), g_2 (Frary, Tideman, & Watts, 1977) and ω (Wollack, 1997).

In contrast to *B, H,* the *K*-indices and *S1*, which only use information of the incorrect responses, the *PAIR1, PAIR2, S2,* g_2 and ω statistic make use of both the incorrect and correct responses. The advantage of this is that all the responses, thus more information, is used. The idea behind the first group of indices is that a suspect who has a larger number of matching incorrect answers than most of the other examinees in the same subgroup (with the same number of wrong answers) should be the one who copied answers from the source (Shu, 2010). In the second group, a theoretical distribution to compute the likelihood of observing the similarity by chance is used.

Operationalization Cheating

Previous studies have investigated the performance of these indices in simulation studies. Table 1 shows different simulation studies that examine cheating detection. As can be seen there, cheating is operationalized in different ways. Some studies changed the probability of answering particular - usually difficult - items correctly to 1.00 (De la Torre & Deng, 2008; Emons, 2008; Karabatsos, 2003; Meijer, 1994; Meijer, Molenaar, & Sijtsma, 2004) or increased this probability (Dimitrov & Smith, 2006; Jurich, Demars, & Goodman, 2012). This can represent many different ways of cheating, for instance, item exposure, prior item knowledge and answer copying. Others focused on answer copying only and changed some answers of one examinee into the answers of another examinee (Hanson, Harris, & Brennan, 1987; Sotaridona, & Meijer, 2002, 2003; Wollack, 1996, 1997, 2003, 2006; Wollack & Cohen, 1998; Zhang & Walker, 2008; Zopluoglu & Davenport, 2011, 2012). We believe this is more realistic as an operationalization of answer copying than 1-scores for the copied items because copiers can also occasionally copy an incorrect item.

The articles also differ in which items are affected. In most studies, the difficult items were copied, but there were also many studies in which random items were chosen. Some studies, e.g., Hanson, Harris, and Brennan (1987) and Wollack (1996; 1997) also included string-based copying, in which consecutive items were copied.

Person-Fit Measures vs. Copying Indices

Hanson, Harris and Brennan (1987) concluded that indices that specifically model examinees who copy strings of consecutive items (e.g., *H, PAIR1* and *PAIR2*) perform better – have a higher true positive rate – in the string-based conditions than more general copying indices (*B,* g_2*, P* and *CP*). We believe this makes sense because the indices are specifically created to detect copying of strings. With the same line of reasoning, we believe that indices that specifically model examinees who copy answers perform better in detecting copying than

more general person-fit indices. The copying-indices are specifically created to detect copying, in contrast to the general person-fit statistics, which identify unusual response patterns, without specifying the cause.

In this study, we will compare the performance in copying detection of the ω copying-index and the l_z* person-fit statistic. Different studies show that the omega performs best (Cizek, 1999; Sotaridona & Meijer, 2002; Wollack, 2003). However, Sotaridona and Meijer (2003) showed that the S_1 and S_2 yield higher power under certain conditions. We have chosen to use the omega statistic in our study because it is widely used as a copying-index and performs very well generally. The g_2 index is structurally identical to omega, except the estimation of the answer match probability (Wollack, 2006). The g_2 index uses classical test theory to compute this probability where the ω index uses the nominal response model. Previous studies have shown that g_2 has inflated Type-I error rates (Hanson, Harris, & Brennan, 1987; Wollack, 1997, 2003). Reise and Due (1991) and Drasgow and Levine (1986) have shown that the l_z index is among the best indices to detect misfit response patterns caused by test cheating. As we will explain later, the l_z* index is the improved version of the l_z index. We will investigate whether the ω index, which is especially designed to detect copying, performs better in copying detection than this statistic. Before we explain our expectations more thoroughly, we will describe the l_z* person-fit statistic and the ω copying-index in more detail.

Table 1

Item- and Person Characteristics of Simulation Studies with Answer Copying Detection.

		Item Characteristics				Person Characteristics			
Source	Operationalization cheating	Total	# Affected	% Affected	Kind of item	Total	# Affected	% Affected	Ability
De la Torre & Deng (2008)	Correct probability 1.00	10-30-50		10-30	Difficult	5,000	-	-	Random
Dimitrov & Smith (2006)	Correct probability .90	10-20-30	10: 2-4 20: 4-8 30: 6-12	20-40	Difficult	9,000	2,430	27	Low
Emons (2003) chapter 2	Correct probability 1.00	20-40	5-8-10	12.5-20-25-40-50	Difficult	1,000	1,000	100	Random
Emons (2003) chapter 4	Answer copying from note: correct probability 1.00	20-40	20: 5-8 40: 5-10	20: 25-40 40: 12.5-25	Difficult	1,000	1,000	100	Random
Hanson, Harris & Brennan (1987)	Answer copying: answers copier same as source	100	10-20-30-40-50	10-20-30-40-50	Random	9,143	500	5.5	Random
Jurich, Demars, Goodman (2012)	Prior item knowledge: adding .5 to correct probability	100	25-100	25-100	Random	3,000	150-300-750-1,500	5-10-25-50	Random
Karabatsos (2003)	Correct response probability of 1.00	17-33-65	3-6-12	18	Difficult	500	25-50-125-250	5-10-25-50	Low

Table 1 Continued

Meijer (1994); Meijer, Molenaar, & Sijtsma (2004)	0-scores changed to 1-scores	17-33	17: 3 33: 6	18	Difficult	450	25-50	5.5-11	Random
Shu (2010)	Prior item knowledge	40-80	40: 12-20-28 80: 24-40-56	30-50-70	Random	2,000	1,000-1,600/ 1,600-2,000	50-80/ 80-100	60% low
Sotaridona & Meijer (2002)	Answer copying: answers copier same as source	40-80	40: 4, 8, 12, 16 80: 8, 16, 24, 32	10-20-30-40	Random	2002: 100-500-2,000	5-25-100	5	Lower than source
Sotaridona & Meijer (2003)	Answer copying: answers copier same as source	40-80	40: 4, 8, 12, 16 80: 8, 16, 24, 32	10-20-30-40	Random	100-500	5-25-100	5	Lower than source
Wollack (1996; 1997)	Answer copying: answers copier same as source	40-80	40: 4-8-12-16 80: 8-16-24-32	10-20-30-40	Random, difficult, string	100-500	5-25	5	Lower than source
Wollack (2003)	Answer copying: answers copier same as source	20-40-80	20: 2-4-6-8 40: 4-8-12-16 80: 8-16-24-32	10-20-30-40	Random	20,000: 50-100-250-500-1,000-2,000-5,000-10,000	4-8-20-40-80-160-400-800	8	Random
Wollack (2006)	Answer copying: answers copier same as source	40-80	40: 4-8-12-16 80: 8-16-24-32	10-20-30-40[1]	Random, strings, mixed	20,000: 25-50-100-250-500-1,000-2,000-5,000-10,000	2-4-8-20-40-80-160-400-800	8	Random

[1] All these percentages of items copied in each condition.

Table 1 Continued

Wollack & Cohen (1998)	Answer copying: answers copier same as source	40-80	40: 4-8-12-16 80: 8-16-24-32	10-20-30-40	random-strings, difficulty	100-500	5-25	5	Lower than source
Zhang & Walker (2008)	Responses to items with difficulty > 1.2 correct	10-20-40	2-4-8	20	Difficult	1,000[2]	100	10	Low
Zopluoglu & Davenport (2011)	Answer copying: answers copier same as source	40	4-8-…-40	10-20-…-100	Random	500 pairs	500 pairs	100	Random
Zopluoglu & Davenport (2012)	Answer copying: answers copier same as source	40	4-8-…-40	10-20-…-100	Random, difficulty, string	500 pairs	500 pairs	100	Random

[2] At first, Zhang and Whalker (2008) also included sample sizes of 500 and 2,000, but the results were collapsed because they were very similar for the three sample sizes.

l_z* Person-Fit Statistic

The l_z* statistic is the modified version of the l_z statistic (Drasgow, Levine, & Williams, 1985), which is the standardized version of the l_0 statistic. In the log-likelihood function (Levine & Rubin, 1979),

$$l_0 = \sum_{i=1}^{k} \{X_i \log P_i(\theta) + (1 - X_i) \log [Q_i(\theta)]\},$$

the likelihood of the observed responses is compared to the expected value for the population. $P_i(\theta)$ denotes the probability of a correct score on item i (i = 1, 2, …, k) and $Q_i(\theta) = 1 - P_i(\theta)$ the probability of an incorrect score on item i.

Unfortunately, l_0 is not standardized. Hence, it depends on θ whether an item-score pattern is classified as model-fitting or misfitting. Furthermore, the null distribution of fitting item scores is unknown for l_0, which makes it impossible to classify an item-score pattern as misfitting or not (Meijer & Sijtsma, 2001). Drasgow, Levine and Williams (1985) proposed a standardized version of the l_0 statistic, the l_z statistic:

$$l_z = \frac{l_0(\theta) - E[l_0(\theta)]}{V[l_0(\theta)]^{1/2}},$$

where $E[l_0(\theta)]$ is the expectation and $V[l_0(\theta)]$ the variance of l_0:

$$E[l_0(\theta)] = \sum_{i=1}^{k} \{P_i(\theta) \log P_i(\theta) + Q_i(\theta) \log Q_i(\theta)\}, \text{ and}$$

$$V[l_0(\theta)] = \sum_{i=1}^{k} P_i(\theta)\, Q_i(\theta)\, w_i(\theta)^2,$$

where $w_i(\theta)$ – the weights – represent $log \frac{P_i(\theta)}{Q_i(\theta)}$.

If a respondent's l_z is below the critical value, the response pattern is classified as aberrant - or misfitting. Otherwise, it is classified as "normal", which means that it is consistent with the model predictions (Lee, 2013).

In practice, the true ability (θ) is unknown and that is why it is usually estimated. Unfortunately, l_z does not have an asymptotically standard normal distribution when the true ability levels are replaced by sample ability estimates (Molenaar & Hoijtink, 1990; Nering, 1995, 1997; Reise, 1995). Several researchers have found that the null distributions are not even asymptotically standard normal when the true ability is used. The distributions are negatively skewed, and often leptokurtic, which results in a conservative classification of item score patterns as misfitting (Meijer & Sijtsma, 2001; Molenaar & Hoijtink, 1990; Nering, 1995). Snijders (2001) proposed a correction for this index such that it is asymptotically standard normal distributed when sample ability estimates are used.

l_z* is derived from l_z by adding $c_n(\hat{\theta})r_0(\hat{\theta})$ to the numerator and replacing the weights $w_i(\theta)$ by the modified weights $\tilde{w}_i(\hat{\theta})$ in the denominator (Magis, Raîche, & Béland, 2012). These new statistics will be discussed further on. l_z* is obtained by

$$l_z^* = \frac{l_0(\hat{\theta}) - E[l_0(\hat{\theta})] + c_n(\hat{\theta})r_0(\hat{\theta})}{\tilde{V}[l_0(\hat{\theta})]^{1/2}},$$

where the modified variance is

$$\tilde{V}[l_0(\hat{\theta})] = \sum_{i=1}^{k} P_i(\theta)\, Q_i(\theta)\, \tilde{w}_i(\theta)^2.$$

The modified weights $\tilde{w}_i(\hat{\theta})$ are obtained by

$$\tilde{w}_i(\hat{\theta}) = w_i(\hat{\theta}) - c_n(\hat{\theta})\, r_i(\hat{\theta}).$$

In this article we will focus on the 2PL-IRT model. In this model, $c_n(\theta)$ and $r_0(\theta)$ can be calculated with

$$c_n(\theta) = \frac{\sum_{i=1}^{k} a_i P_i(\theta)\, Q_i(\theta)\, w_i(\theta)}{\sum_{i=1}^{k} a_i^2\, P_i(\theta)\, Q_i(\theta)}$$

and

$$r_0(\theta) = \frac{\sum_{i=1}^{k} a_i{}^3 \, P_i(\theta) \, Q_i(\theta)[P_i(\theta) - \, Q_i(\theta)]}{2 \, \sum_{i=1}^{k} a_i{}^2 \, P_i(\theta) \, Q_i(\theta)}.$$

For obtaining $\hat{\theta}$, Magis, Raîche, & Béland (2012) discuss the use of 3 estimators. For reasons of conciseness, we used Warm's weighted likelihood estimator $\hat{\theta}_{\mathrm{WLE}}$ (Warm, 1989) as ability estimator in $r_0(\theta)$.

The ω Index

The ω index (Wollack, 1997) measures the standardized difference between the number of answer matches – both correct and incorrect – between a pair of examinees, and the number predicted by chance. An examinee – the copier (c) – is suspected of copying answers from another examinee – the source (s). H_{cs} is equal to the number of items where the response of the copier matches the response of the source. $P_{ik}(\theta_c)$ denotes the probability of the copier selecting the same answer *k* on item *i* as the source. This probability is calculated by Wollack (1997) with the nominal response model (Bock, 1972),

$$P_{ik}(\theta_c) = \frac{\exp\left(\zeta_{ik} + \lambda_{ik}\theta_c\right)}{\sum_{v_i=1}^{V_i} \exp\left(\zeta_{iv} + \lambda_{iv}\theta_c\right)},$$

where ζ_{ik} and λ_{ik} denote the intercept and slope parameters, respectively, V_i denotes options for the multiple choice item i, and θ_c denotes the ability level of the copier.

The expected value of answer matches is conditional on θ_c, the item response vector of the source (U_s), and the item parameters $\xi = (\xi_1, \ldots, \xi_I)$ with $\xi_i = (\zeta_{i1}, \ldots \zeta_{iV}, \lambda_{i1}, \ldots, \lambda_{iV})$:

$$E\left(h_{cs} \middle| \theta_c, U_s, \xi\right) = \sum_{i=1}^{I} P_{ik}(\theta_c).$$

The standard deviation of h_{cs} is

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

The previous statistics combined make it possible to calculate the ω statistic with

$$\omega_{cs}{}^{3} = \frac{h_{cs} - E\left(h_{cs}\middle|\theta_{c,}U_s, \xi\right)}{\sigma_{h_{cs}}}.$$

The ω index is asymptotically standard normally distributed (Wollack, 1997). The larger its value, the higher the indication that c copied from s.

Wollack (1997) concluded that ω had good power to detect copiers if at least 20% of the items was copied on an 80-item test and at least 30% on a 40-item test. Furthermore, the Type I error rate of ω was below alpha in virtually all conditions. We expect to replicate this result.

Hypothesis

Our hypothesis is that the copying-index ω performs better in the detection of answer copying than the general person-fit index l_z*. More specifically, we predict that ω has an overall higher detection rate. Furthermore, we expect the empirical type I error rates (false positives) to be close to the nominal type I error rates (alpha levels) for l_z* and – as in Wollack's (1997) study – below its nominal level for ω.

Previous studies (e.g., Wollack, 1996; Wollack & Cohen, 1998; Sotaridona & Meijer, 2002, 2003; Wollack, 2003, see Table 1) manipulated the test length, amount of copying, and sample size. Zopluoglu and Davenport (2012) argued that, in general, the power of the indices increased as test length and amount of copying increased. Longer tests and higher amounts of copying provide more information about the answer copying, but only if it exists.

[3] The subscript of ω here is chosen to make clear that ω says something about the source-copier pair. However, it is not used in the rest of the paper because it is not used in the paper of Wollack (1997) or in other previous papers.

Furthermore, larger sample sizes provide more accurate parameter estimation, and thus, the power is more reliably estimated.

Method

Data Generation

A simulation study was performed to evaluate the Type I error rate and power of the l_z* person-fit index and Wollack's ω copying-index to detect copying behavior. Data were generated under the two-parameter logistic IRT model (2PLM; Birnbaum, 1968). We did not want to include a guessing parameter because this makes it more difficult to estimate the item parameters. The difficulty parameter values, as well as the theta-values of examinees, were generated from a standard normal distribution, $N(0,1)$. Theta-values were independently generated for copiers and sources. The discrimination parameter values were generated from a uniform distribution ranging from +0.5 to +1.5. Fox and Meijer (2008) used normally distributed discrimination parameters, $N(1, 0.2)$. However, a uniform distribution does not allow for negative values of the discrimination parameter, which are not plausible in practice. New examinees were generated for each replication. A program in "R" was written that performed the required simulations.

Simulation of Cheaters

Emons (2003) fixed the probability of answering correctly to 1.00 for items with the highest difficulty parameter. In this way, it is simulated how examinees bring notes illegally and use these answers. Meijer (1994) and Meijer, Molenaar, and Sijtsma (2004) simulated cheating by changing some examinees' 0-scores for the most difficult items to 1-scores. These two ways of simulating cheating result in the same scores for the cheaters, namely correct scores on cheated items. However, as mentioned before, we are specifically interested in answer copying from another examinee because this is a popular way of cheating (Fox & Meijer, 2008). So, to implement cheating, the first 5% of the lowest theta-value copiers were

matched with 5% randomly chosen sources. We believe that examinees with an average or high ability rely on their own knowledge to respond to the items. On the other hand, the examinees with a low ability will be more likely to copy someone else's answers. They are more likely to obtain a score gain by answer copying than the other examinees.

As can be seen in Table 1, different answer copying studies included 5% copiers (e.g., Hanson, Harris, & Brennan, 1987; Wollack, 1997; Sotaridona & Meijer, 2002, 2003). Wollack's study (2003) included 8% copiers, but the author noted that that number "is probably slightly higher than the average percentage of students copying during any given test administration" (p. 194). This number – 5% copiers – is also consistent with the studies of Bellezza and Bellezza (1989) and Frary and Olson (1985) mentioned before.

Zopluoglu and Davenport (2012) noted that random copying may not be correct in real life, and that other types of copying – like difficulty-weighted copying – are probably more correct. We think that examinees first try to answer an item themselves – especially in high-stake tests – and only look at their neighbor's answer if they do not know it themselves. Hence, we believe that the copying of difficult items happens more in practice than the copying of consecutive strings. The items are not ordered on difficulty, which makes it unlikely that respondents copy strings of items. Even if they see multiple answers of the source at once, they probably only copy if they have no idea themselves. That is why we made the copiers copy only the most difficult items. The percentage of items copied differs per condition. As a result, the copiers have more 1-scores than expected based on their low theta-values. However, in our study, also incorrect answers will be copied. Especially since the most difficult items are copied, not all sources will know the correct answer.

Wollack (1997) randomly selected 5% of the examinees as copier and made them copy answers from a more able examinee within copying distance. In our study, it is theoretically possible that a copier copies from a source with lower theta-value. Even though

the copiers have low ability levels, it is possible that the randomly chosen source has an even lower ability level. However, we believe that it is possible that a low-ability copier looks at the answers of his or her neighbor, independent of his or her ability level. For instance, the copier may not know the other examinee. Furthermore, we did not use a seating chart to identify source-copier pairs within copying distance. We assumed that the pairs were already suspected of copying.

Independent Variables

Test length. Following Wollack (1996, 1997, 2003, 2006), Wollack and Cohen (1998), Shu (2011), Sotaridona and Meijer (2002, 2003), and Sijtsma and Meijer (2001), data were generated for two levels of test length: $k = 40$ and $k = 80$. Test length is known to affect the accuracy of item parameter estimation, which in turn affects person trait estimation (Zhang & Walker, 2008). Moreover, test length is found to increase ω's power (Wollack, 1997). Even though shorter tests might be used in practice, detecting person misfit for shorter tests is statistically almost impossible regardless of which statistical model is used (Rubb, 2013).

Amount of items copied. Following Sotaridona and Meijer (2002, 2003), Wollack (1996, 1997, 2003) and Wollack and Cohen (1998), the proportion of items copied (m) was 0, .1, .2, .3 or .4. Thus, in the 40-item conditions, 0, 4, 8, 12 or 16 items are copied, and in the 80 item conditions, 0, 8, 16, 24 or 32 items. The situation without answer copying was included to estimate empirical type I error rates. Wollack (1997) concluded that ω's power to detect answer copying was higher when more items were copied.

Sample size. Both 1,000 and 10,000 copier – source pairs were generated. According to Zhang and Walker (2008), sample size is related to the statistical power of fit statistics. However, Wollack (1997) did not find an effect of sample size on empirical power.

Significance level. Following Zopluoglu and Davenport (2011, 2012), the theoretical Type I error rates (significance levels) used were .05, .01 and .001. Meijer (2003) noted that relatively large alpha levels, e.g., .05 and .10, are preferable because most person-fit statistics have relatively low power at low alpha levels because of limited test length. Furthermore, he mentioned that extreme person-fit values will only alert the researcher that the behavior is unexpected and worth studying more closely. However, Sotaridona, Van der Linden, and Meijer (2006) and Shu (2010) argued that conservatism is often more desirable in answer copying detection. False positives can have serious consequences for particular examinees, especially when the results of the test are not used for screening but as cheating test. It is usually less bad to fail to detect a cheater (false negative) than to incorrectly accuse someone of cheating (false positive). To study the effect of these different possible choices for alpha, we compared the results for different levels of alpha.

The interaction of conditions resulted in a 2 (test length) x 5 (amount of items copied) x 2 (sample size) x 3 (significance level) design, for a total of 60 testing conditions. New data were generated for 20 different conditions (2 test length x 5 amount of copying x 2 sample size). The dataset in the 1,000 copier-source pairs conditions was replicated 1,000 times per condition and – due to increasing computation time - 100 times in the 10,000 copier-source pairs conditions. The higher the number of replications, the smaller the standard errors of the false positives and false negatives – up to a point of diminishing returns (Rubb, 2013).

Dependent Variables

The performance of the ω copying-index and the l_z* person-fit statistic in detecting copying behavior is determined by the trade-off between the Type I error rate and the detection rate. An optimal index should be powerful enough to detect true answer copiers, but at the same time should not pick out the non-copying examinees too often. The empirical Type I error rate is the proportion of non-copiers incorrectly identified as copier (false

positives). Hence, the proportion of falsely detected pairs of both indices was computed based on 1,000 or 10,000 non-copying pairs for each level of test length and theoretical alpha level. The detection rate – or power – is the proportion of copiers correctly identified as copier. The detection rate for each statistic was computed as the proportion of simulated copiers who were identified as copying by the indices.

For the l_z* statistic, a pair of examinees was identified as copying if the l_z*-value was below the one-tailed critical value corresponding to the alpha level (i.e., critical values of -1.64, -2.32, and -3.09 for the alpha level of .05, .01, and .001, respectively). For the ω index, on the other hand, a pair is identified as copying if the ω-value is above the one-tailed critical value corresponding to the alpha level (i.e., critical values of 1.64, 2.32, and 3.09 for the alpha level of .05, .01, and .001, respectively).

Results

Performance of the Indices

Detection rates. For the three significance levels, Table 2 shows the detection rates of l_z* and ω for detecting answer copying under 20 combinations of test length, sample size and number of items copied. As expected, the detection rates increased – for both l_z* and ω – as the theoretical alpha level, test length and amount of copying increased. However, the detection rates were almost equal for both sample sizes.

In general, the detection rate was quite low. The detection rate was extremely low for small alpha levels (.001 and .01) and low amounts of items copied (m = .1 and m = .2). However, better detection rates were found for higher amounts of items copied (m=.40) and high alpha levels. For instance, the detection rate for ω was .777 for n = 1,000, 80 items, an alpha level of .05 and 40% items copied. The detection rate of l_z* in this condition was .521.

When we look at the relative performance of the indices, we see that l_z* performs better than ω in almost all conditions. The conditions in which ω performs better than l_z* are

bolded in table 2. When 30% of the items were copied, ω performs better than l_z in the 80 items and α = .05 conditions. When 40% of the items were copied, ω performs better than l_z* in the conditions with α = .05 and in the conditions with α = .01 and 80 items. Importantly, the conditions in which ω performs better are the conditions with the highest absolute levels of detection for both indices.

Table 2

Detection Rates for Answer Copying

	n = 1,000[a]				n = 10,000[a]			
	k = 40		k = 80		k = 40		k = 80	
α	l_z*	ω	l_z*	ω	l_z*	ω	l_z*	ω
				m = .1				
.05	.061	.044	.106	.064	.062	.047	.108	.060
0.01	.025	.006	.055	.011	.027	.007	.056	.010
.001	.008	<.001	.024	.001	.010	<.001	.024	.001
				m = .2				
.05	.174	.106	.278	.191	.176	.108	.285	.201
.01	.105	.020	.194	.046	.107	.021	.199	.051
.001	.055	.002	.124	.005	.055	.002	.126	.006
				m = .3				
.05	.288	.231	.419	**.474**	.291	.234	.417	**.465**
.01	.199	.059	.328	.180	.204	.061	.324	.176
.001	.124	.007	.238	.033	.130	.008	.238	.031
				m = .4				
.05	.390	**.450**	.521	**.777**	.390	**.455**	.525	**.776**
.01	.290	.163	.429	**.464**	.292	.166	.434	**.465**
.001	.196	.028	.336	.149	.198	.029	.337	.153

Note. k denotes the test length, m denotes the proportion of items copied by the copiers, and α denotes the theoretical alpha level. The bolded detection rates denote the conditions in which ω performs better than l_z*.

[a] 1,000 or 10,000 source-copier pairs

Empirical type I error rates. Table 3 shows the empirical Type I error rates for both indices within each amount of copying for different theoretical alpha levels. Test length and sample size did not appear to have an effect on the error rates of both indices. Hence, Type I error rates were averaged across the two test lengths and the two sample sizes (see Appendix

A for the Type I error rates in each condition). The empirical type I error rates for ω never exceed the nominal level in any of the conditions. Rather, as expected, they were much smaller than their nominal levels. The empirical type I error rate for l_z* was equal to its nominal level for α = .001 and 40% of items copied, and slightly below its nominal value for α = .01 and no items copied[4]. However, it (slightly) exceeds its nominal level in most conditions with an alpha level of .001, with a maximum empirical type I error rate of .0017. In the remaining conditions, the empirical type I error rates for l_z* were much smaller than their nominal levels, as with the type I error rates for ω.

In the conditions without answer copying, the ω index was more succesful to hold the Type I error rate around its nominal level at an alpha level of .05 than l_z*, but ω was less succesful to do this than l_z* at alpha levels of .01 and .001. In the conditions with answer copying, the ω index was less succesful than l_z* to hold the Type I error rate around its nominal level at alpha levels of .05 and .01. At an alpha level of .001, the Type I error rate of the ω index was about equally far below its nominal level as that of the l_z* index above its nominal level. An increase in amount of copying appeared to decrease the empirical type I error rate for l_z* and slightly increase ω's empirical type I error rate.

[4] Inspection of the Type I error rates for each condition (see Appendix A) reveals that the Type I error rate for l_z* in the condition with α = .01, m = 0, k = 40, and n = 10,000 was 0.0102, while l_z*'s average Type I error rate for α = .01 and m = 0 was only 0.0098. So, although the difference between these two values is very small, this made the difference between a type I error rate slightly above and slightly below its nominal value.

Table 3

Empirical Type I Error Rates

α	l_z*	ω
	$m = 0$	
.05	.0380	.0346
.01	.0098	.0052
.001	.0017	.0003
	$m = 0.1$	
.05	.0333	.0341
.01	.0082	.0050
.001	.0013	.0003
	$m = 0.2$	
.05	.0303	.0347
.01	.0073	.0054
.001	.0012	.0004
	$m = 0.3$	
.05	.0285	.0356
.01	.0069	.0055
.001	.0011	.0004
	$m = 0.4$	
.05	.0270	.0361
.01	.0065	.0057
.001	.0010	.0004

Note. m denotes the proportion of items copied by the copiers, and α denotes the theoretical alpha level.

Discussion

This was the first article to compare the l_z* person-fit statistic and ω copying-index directly. The hypothesis stated that the ω copying-index would perform better in the detection of answer copying (i.e., have higher detection rates than the l_z* person-fit statistic). Against our expectations, l_z* performed better than ω in most conditions. Nevertheless, the conditions in which ω performed better are the conditions with the highest absolute levels of detection for both indices.

In contrast to Wollack's (1997) findings, the detection rate for ω was only high (about .8) for large amounts of items copied (m = .4), a test length of 80 items and an alpha level of .05. The detection rate for l_z* did not reach this detection rate level in any of the conditions.

In real-life testing situations, the test length and alpha level can be controlled by the test administrator. That is why it is possible to make a trade-off beforehand between the detection rate and empirical type I error rate. You want to detect cheating, but not at the expense of innocent examinees. An ideal index maintains the empirical type I error on or slightly below the nominal level, but not too far below. If it is far below its nominal level, the type I error rate is lower than you decided it to be beforehand, which decreases the detection rate. The reasons for this is that the detection rate is lower for smaller alpha levels, and thus, the indices have lower power than if the empirical type I error rate would have been equal to alpha.

The results revealed that the empirical type I error rate was (far) below its nominal value in all conditions for ω and in most of the conditions with alpha levels .05 and .01 for l_z*. However, l_z*'s empirical type I error rates were equal to or (slightly) above the nominal value in the conditions with α = .001 and in the conditions with α = .01 and no items copied. Hence, l_z* is slightly liberal for most conditions with α = .001, but conservative for most conditions with alpha levels of .05 and .01. The ω index is conservative in all conditions.

Test length did not appear to influence the empirical type I error rates. Because the test of 80 items had higher detection rates than the test of 40 items, we recommend using more items whenever possible. Unfortunately, the amount of items copied is usually not known beforehand. Because the detection rates of both l_z* and ω are extremely low for a small amount of items copied, but are larger when more items are copied, it is hard to say something about the power of the indices in practice. Nevertheless, you might argue that it is

less bad to fail to detect an answer copier if the examinee only copied a few items than when he/she copied many items.

Besides the detection rates and empirical type I error, there are other differences between the l_z* and ω index. First of all, the ω index is created to test for already suspected source-copier pairs. More specifically, the index says something about the likelihood of response matches in a source-copier pair. On the other hand, l_z* says something about the likelihood of the responses of one examinee under the model. So, l_z* does not require designating a source. Secondly, l_z* is a general person-fit index that detects general aberrant behavior. Hence, it might detect other kinds of misfit – like guessing – as well. So, examinees may become falsely accused of cheating, while they actually perform other – allowed or not – aberrant behavior. The ω copying-index is probably less affected by these other behaviors because it focuses specifically on answer copying.

We computed the l_z* and ω index for all 1,000 (or 10,000) examinee pairs. However, we must note that test administrators should be careful by using these indices as general screening for answer copying. Some examinees – a proportion equal to the type I error rate - will be falsely indicated as cheater, and these false positives can have serious consequences for the examinees concerned. Because the ω index is specifically designed for the detection of already suspected source-copier pairs, it should not be used for all possible pairs of examinees.

This study has some limitations. First of all, Snijders (2001) proposed a modification of the l_z such that it follows an asymptotically standard normal distribution when the ability level is estimated. Unfortunately, the mean of l_z* was not close to zero in our study. Rather, the mean l_z* was about 0.3 in conditions with 40 items and even about 0.4 in conditions with 80 items. The standard deviation of l_z* ranged from 1.07 to 1.30. We mainly looked at the conditions without copying, but the means and standard deviations of the other conditions

were quite similar. Van Krimpen-Stoop and Meijer (1999) did also find a mean of l_z* above zero, with a mean of about 0.1 for P&P (paper-and-pencil) tests, and even reaching levels of 0.4 for a Computerized Adaptive Test (CAT). The relatively high mean of l_z* leads to a more conservative index. As a result, fewer examinees are falsely identified as answer copier, but also fewer answer copiers are correctly identified.

We had to compute the l_z statistic in order to compute l_z*, which made it possible to compare the distribution of these two statistics. This reveals that although the l_z statistic has a mean closer to zero (ranging from 0.189 to 0.255) than l_z*, l_z* is less negatively skewed than l_z. As an illustration, figure 1 and 2 show the distributions of both l_z and l_z*, respectively, for one sample of 1,000 source-copier pairs, without copying and $k = 40$. We are not sure what caused the relatively high mean of l_z*. One possibility is that Snijders (2001) corrected for the fact that theta is estimated, but he does not correct for the fact that the item parameters are estimated. However, this does not explain why l_z* has a higher mean than l_z because l_z does not take into account this either.

Fortunately, the ω index appeared to follow a standard normal distribution. Its mean was about zero and its standard deviation about, but slightly below, one. See Appendix B for the average means and standard deviations in each condition for l_z*, l_z and ω.

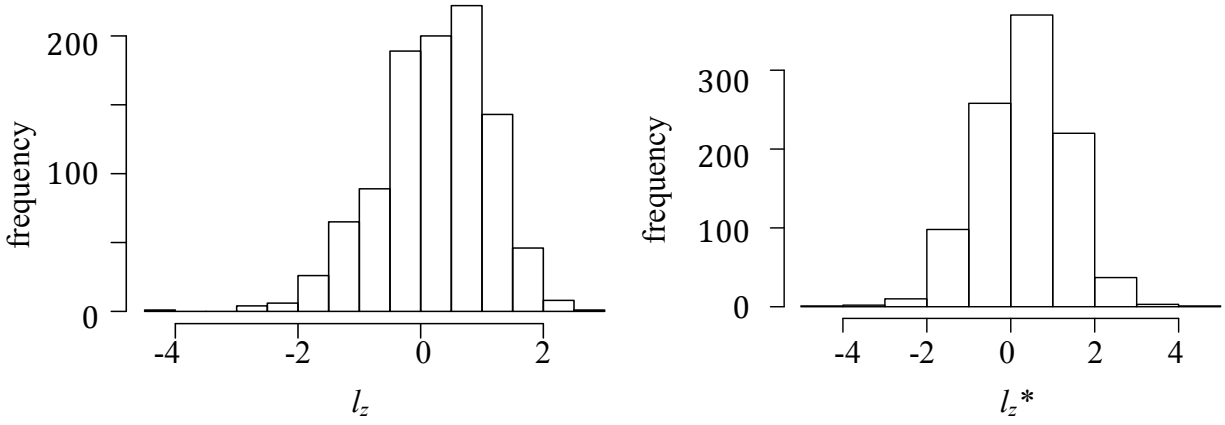

Figure 1 and 2. Distribution of the l_z and l_z* statistic.

Secondly, the copiers in our study were the 5% examinees with the lowest theta-values. We argued before that people only copy answers if they have no idea themselves. However, the copiers in our study have extremely low theta-values, and thus, they probably do not know the answers of most items. Both the copiers' theta-values and the difficulty parameter values were generated from a standard normal distribution, so the copiers have a correct probability of less than 50% on at least 95% of the items. As a result, it would be realistic if these people copy even more than 40% of the items. Nevertheless, it might be not possible to copy more than 40% of the items because of measures taken to reduce answer copying – like examination supervisors. Furthermore, the detection rates are likely to be lower when the copiers are not the examinees with the lowest theta-values, but have higher theta-values. In that case, the differences in theta between the source and copier are lower, which results in less deviation from their actual score without answer copying, and thus, less misfit. Nevertheless, we believe that examinees with the lowest theta-values are most likely to engage in answer copying because they are most likely to obtain a score gain and are least able to complete the items on their own.

Another possible limitation is that we replicated the conditions with 10,000 source-copier pairs only 100 times instead of 1,000 due to increasing computation time. As a result, the 10,000 source-copier pairs conditions are more prone to sampling fluctuations. However, the differences in detection rate and empirical type I error rate between the two sample sizes were uniform and so small that they gave no reason to increase the replication rate.

As we argued before, answer copying threatens the validity of tests, which in turn affects high-stake decisions. Hence, it is important that well-performing copying detection statistics are created. These can be used in practice to increase test validity. Unfortunately, we believe that copying detection methods are not often enough used in practice. For instance, teachers in colleges could use these methods to detect suspected source-copier pairs. Software

is available that makes it possible to apply these copying indices, for instance "R" packages (e.g., CopyDetect, Zopluoglu, 2013), but this software is not easy-to-use. Therefore, the creation of user-friendly software can highly facilitate the practical use of these methods. In this way, the validity of tests can be increased. But first, it is important that copying detection methods are found that have high detection rates, while keeping the empirical type I error rate under control. In this study, we compared a copying-index and a general person-fit statistic. We were positively surprised by the performance of l_z* compared to ω, even though its detection rates were low in general. This provides information about which index to use in which condition.

We have several recommendations for future research. First of all, it is interesting to compare more different person-fit and copying-indices. The l_z index – and thus probably also l_z* – was among the best person-fit indices to detect cheating (Reise & Due, 1991; Drasgow & Levine, 1986). Hence, it is possible that l_z* was only an exception, and that ω performs consistently better than other person-fit statistics. It is also possible that other copying-indices than ω perform better than l_z* or even better than both l_z* and ω. The l_z* and ω index do not distinguish between the copier's true ability and the gain in ability due to cheating. When copiers copy answers from a high ability source, they are likely to copy correct answers and increase the observed ability estimates (Zopluoglu & Davenport, 2012). In our study, the ability estimates of the low-theta copiers are likely to be biased because they gain in ability due to answer copying. As illustration, for one situation with 1,000 source-copier pairs, 40 items and 10% of the items copied, the theta-value of the examinee with the lowest ability was -3.8, while it was estimated to be only -2.4. Hence, methods that do take into account this gain in ability due to cheating might be better.

Wollack (2006) already combined different copying-indices. It may be interesting to combine person-fit and copying-indices and see how these combinations perform in the

detection of copying. As mentioned before, the disadvantage of general person-fit indices is that they also detect other types of misfit. Hence, a combination of general and specific indices with more specific copying-indices might lead to better copying-detection.

Furthermore, Van der Linden and Sotaridona (2006) concluded that the power to detect answer copying was greater when the pairs of examinees had larger differences in estimated abilities. In the study of Van der Linden and Sotaridona (2006), best power was obtained for cases with low ability for copier and higher ability for the source. Lewis and Thayer (1998) noted that the power in detecting true answer copying pairs is very likely to decrease as the ability of the source examinee increase. Moreover, Zopluoglu and Davenport (2012) concluded that the empirical type I error rates were highest when both examinee's ability levels were below zero, and were smallest when both examinee's ability levels were above zero. In our study, we did not compare the power and empirical type I error rate at different levels of the copier's and source's ability. We would like to include this in our future studies.

In our simulation study, all examinees responded to all items. However, this is not likely to happen in practice. Zhang and Walker (2008) examined the effect of missing data on person-model fit and person trait estimation. They concluded that the higher the proportion of missing data, the larger the number of persons incorrectly diagnosed. The pairwise deletion method led to the best recovery of person-model fit and person trait level. To make the simulation study even more realistic, it might be interesting to include non-response.

References

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

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16*(3), 151-155.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability.* In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. London: Routledge.

De La Torre, J., & Deng, W. (2008). Improving Person-Fit Assessment by Correcting the Ability Estimate and Its Reference Distribution. *Journal of Educational Measurement, 45*(2), 159-177.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied measurement, 7*(2), 170-183.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*(1), 59-79.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86.

Emons, W. H. M. (2003). *Detection and diagnosis of misfitting item-score vectors.* Amsterdam: Dutch University Press

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied* Psychological *Measurement*, *32*(3), 224-247.

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(8), 595-610.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational and Behavioral Statistics, 2*(4), 235-256.

Frary, R. B., & Olson, G. H. (1985, March). *Detection of Coaching and Answer Copying on Standardized Tests*. Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of* copying (Research Report Series No. 87-15). Iowa City, IA: American College Testing Program.

Jurich, D. P., DeMars, C. E., & Goodman, J. T. (2012). Investigating the impact of compromised anchor items on IRT equating under the nonequivalent anchor test design. *Applied Psychological Measurement*, *36*(4), 291-308.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–290.

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

Magis, D., Raîche, G., & Béland, S. (2012). A Didactic presentation of Snijders's lz* index of person fit with emphasis on response model selection and ability estimation. Journal of Educational and Behavioral Statistics, 37(1), 57-81.

McCabe, D. L. (1993). Faculty responses to academic dishonesty: The influence of student honor codes. *Research in Higher Education, 34*(5), 647-658.

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics &Behavior, 11*(3), 219-232.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311-314.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21*, 99–113.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*(1), 72-87.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*(2), 111-120.

Meijer, R. R., & Tendeiro, J. N. (2012). The use of the lz and lz* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, *37*(6), 758-766.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75-106.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121–129.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, *21*(2), 115-127.

Rakovski, C. C., & Levy, E. S. (2007). Academic Dishonesty: Perceptions of Business

Students. *College Student Journal, 41*(2), 466-481.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*(3), 213-229.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*(3), 217-226.

Rubb, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*(1), 3-38.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*(1), 41-53.

Shu, Z. (2010). *Detecting test cheating using a deterministic, gated item response theory model* (Unpublished dissertation). The University of North Carolina at Greensboro, Greensboro, NC.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*(2), 191-207.

Snijders, T. A. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331-342.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical Properties of the K-Index for Detecting Answer Copying. *Journal of Educational Measurement*, *39*(2), 115-132.

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, *40*(1), 53-69.

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement*, *30*(5), 412-431.

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*(3), 283-304.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*(4), 307-320.

Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of educational measurement, 40*(3), 189-205.

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, *19*(4), 265-288.

Wollack, J. A. (2007). Computer Software Review: Cheating Detection at Your Fingertips A Review of INTEGRITY. *Applied Psychological Measurement*, *31*(3), 233-239.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement*, *22*(2), 144-152.

Zhang, B., & Walker, C. M. (2008). Impact of Missing Data on Person—Model Fit and Person Trait Estimation. *Applied Psychological Measurement, 32*(6), 466-479.

Zopluoglu, C. (2013). CopyDetect An R Package for Computing Statistical Indices to Detect Answer Copying on Multiple-Choice Examinations. *Applied psychological measurement*, *37*(1), 93-95.

Zopluoglu, C., & Davenport Jr, E. C. (2011). The Effects of Answer Copying on the Ability Level Estimates of Cheater Examinees in Answer Copying Pairs. *Online Submission*.

Zopluoglu, C., & Davenport, E. C. (2012). The Empirical Power and Type I Error Rates of the GBT and ω Indices in Detecting Answer Copying on Multiple-Choice Tests. *Educational and Psychological Measurement*, *72*(6), 975-1000.

Appendix A

Empirical Type I error rates for all conditions.

	n = 1,000[a]				n = 10,000[a]			
	k = 40		k = 80		k = 40		k = 80	
α	l_z*	ω	l_z*	ω	l_z*	ω	l_z*	ω
				m = 0				
.05	.0361	.0343	.0392	.0353	.0377	.0336	.0388	.0350
0.01	.0097	.0051	.0098	.0054	.0102	.0050	.0097	.0055
.001	.0017	.0003	.0015	.0004	.0019	.0003	.0015	.0004
				m = .10				
.05	.0322	.0337	.0335	.0347	.0335	.0333	.0340	.0347
.01	.0083	.0051	.0080	.0056	.0087	.0048	.0080	.0054
.001	.0014	.0003	.0013	.0004	.0015	.0003	.0012	.0004
				m = .20				
.05	.0294	.0338	.0309	.0358	.0305	.0340	.0303	.0354
.01	.0074	.0051	.0072	.0057	.0076	.0051	.0071	.0056
.001	.0013	.0003	.0011	.0004	.0013	.0003	.0010	.0004
				m = .30				
.05	.0280	.0349	.0282	.0366	.0290	.0345	.0288	.0366
.01	.0071	.0053	.0066	.0060	.0074	.0050	.0065	.0058
.001	.0011	.0003	.0010	.0004	.0013	.0003	.0010	.0004
				m = .40				
.05	.0264	.0347	.0270	.0374	.0276	.0349	.0269	.0375
.01	.0066	.0053	.0063	.0060	.0068	.0053	.0064	.0061
.001	.0011	.0003	.0009	.0005	.0011	.0003	.0010	.0005

Note. k denotes the test length, m denotes the proportion of items copied by the copiers, α denotes the theoretical alpha level, and n denotes the sample size.

[a] 1,000 or 10,000 source-copier pairs

Appendix B

Average means and standard deviations for the l_z, l_z* and ω index.

	n = 1,000[a]					
	k = 40			k = 80		
m	l_z	l_z*	ω	l_z	l_z*	ω
0	0.238	0.404	0.019	0.187	0.318	-0.016
	(0.87)	(1.11)	(0.92)	(0.87)	(1.08)	(0.92)
.1	0.245	0.405	0.004	0.197	0.313	0.015
	(0.86)	(1.10)	(0.92)	(0.86)	(1.07)	(0.92)
.2	0.252	0.407	0.032	0.204	0.309	0.056
	(0.88)	(1.12)	(0.92)	(0.90)	(1.13)	(0.94)
.3	0.253	0.404	0.062	0.207	0.310	0.099
	(0.92)	(1.16)	(0.94)	(0.97)	(1.21)	(0.98)
.4	0.253	0.407	0.091	0.204	0.309	0.141
	(0.97)	(1.21)	(0.97)	(1.07)	(1.30)	(1.04)
	n = 10,000[a]					
	k = 40			k = 80		
m	l_z	l_z*	ω	l_z	l_z*	ω
0	0.239	0.401	0.019	0.189	0.323	-0.018
	(0.87)	(1.11)	(0.92)	(0.86)	(1.08)	(0.92)
.1	0.246	0.407	0.004	0.198	0.310	0.014
	(0.86)	(1.10)	(0.92)	(0.86)	(1.07)	(0.92)
.2	0.253	0.410	0.032	0.207	0.317	0.057
	(0.87)	(1.12)	(0.92)	(0.89)	(1.13)	(0.94)
.3	0.255	0.407	0.062	0.208	0.306	0.097
	(0.91)	(1.16)	(0.94)	(0.97)	(1.21)	(0.98)
.4	0.254	0.410	0.091	0.204	0.308	0.140
	(0.97)	(1.21)	(0.97)	(1.07)	(1.30)	(1.04)

Note. k denotes the test length, m denotes the proportion of items copied by the copiers and n denotes the sample size. Standard deviations are in parentheses.

[a] 1,000 or 10,000 source-copier pairs