ACT WORKING PAPER 2015-08

Estimating the Probability of Traditional Copying, Conditional on Answer-Copying Statistics

Jeff Allen, PhD
Andrew Ghattas

December, 2015

ACT Working Paper Series

ACT working papers document preliminary research. The papers are intended to promote discussion and feedback before formal publication. The research does not necessarily reflect the views of ACT.

Jeff Allen is a statistician in the Research division at ACT. He specializes in longitudinal research linking test scores to educational outcomes and student growth models.

Andrew Ghattas is a statistician finishing his PhD in Biostatistics at the University of Iowa. He works for the UI Policy Center on policy evaluation and does freelance Statistical Consulting through Logos Analytics LLC. His main interests are Bayesian hierarchical modeling, simulation experiments, and model selection.

Acknowledgements:

The authors would like to thank Ruitao Liu, Richard Sawyer, Jim Sconing of ACT, Inc., and two anonymous reviewers for their suggestions on previous versions of this paper. We also thank Jim Wollack of The University of Wisconsin-Madison for providing the item parameters used for simulation. A previous version of this paper was presented at the 2013 National Council on Measurement in Education National Conference. The final version of this paper will appear in Applied Psychological Measurement.

Estimating the Probability of Traditional Copying, Conditional on Answer-Copying Statistics

Abstract

Statistics for detecting copying on multiple-choice tests produce p-values measuring the probability of a value at least as large as that observed, under the null hypothesis of no copying. The posterior probability of copying is arguably more relevant than the p-value, but cannot be derived from Bayes' Theorem unless the population probability of copying and probability distribution of the answer-copying statistic under copying are known. In this paper, we develop an estimator for the posterior probability of copying that is based on estimable quantities and can be used with any answer-copying statistic. The performance of the estimator is evaluated via simulation and we demonstrate how to apply the formula using actual data. Potential uses, generalizability to other types of cheating, and limitations of the approach are discussed.

Introduction

Cheating on standardized tests is a well-documented problem and there have been several high-profile incidents across the U.S. in recent years (NCES, 2013). One of the most basic and oldest forms of cheating occurs when one examinee copies the responses of another. Traditional copying occurs when sight, sound, or touch (e.g., tapping) is used to obtain responses, without the aid of electronics.[1] For traditional copying to occur, students must take the same test at the same time and in the same physical location. Efforts to prevent traditional copying include assigned seating at an appropriate distance (NCME, 2012), spiraling of test forms, physical barriers that prevent visual copying, and test proctors who look for wandering eyes and listen for unnecessary noises. Despite these preventive efforts, traditional copying can still occur for students in the same test center. When cases of suspected copying are reported, answer-copying statistics can be used to obtain additional evidence of copying. Or, the statistics can be used to trigger investigations of misconduct. These statistics are often designed for multiple-choice tests and measure level of item response similarity (Angoff, 1974). Examples of answer-copying statistics include the *K*-Index (Holland, 1996), the Omega Index (ω; Wollack, 1996, 1997), the S-check (Weslowsky, 2000), the *VM*-Index (Belov, 2011), and the generalized binomial test (van der Linden & Sotaridona, 2006; Zopluoglu & Davenport, 2012).

Answer-copying statistics often yield a p-value representing the probability that the statistic could exceed a critical value, under the hypothesis of no cheating. The p-value can be expressed as $\text{pr}(X \geq a | Y = 0,\ C = 1)$ where X is a random variable representing the answer-copying statistic, Y is a dichotomous random variable where Y=1 if copying occurred and Y=0 otherwise, and C=1 indicates that an examinee pair came from the same test center (C=0 for

[1] An example of nontraditional copying would be examinees in different physical locations texting each other with exam responses.

different-center pairs). In the broader field of test security data forensics, answer-copying statistics are not alone in their use of the p-value as evidence of fraud. Other applications of the p-value include the detection of aberrations in erasure behavior (van der Linden & Jeon, 2012), as well as test score gains and group improvement over the prior year.

While computationally feasible and grounded in statistical theory, the p-value does not directly address the question: "What is the probability that test fraud occurred, given the available evidence?" For individuals charged with detecting test fraud, this question is more relevant than "What is the probability of the available evidence, given that test fraud did not occur?" Using the notation introduced earlier, this probability of interest can be written as $\text{pr}(Y = 1|X = a, C = 1)$.

If investigators knew that the probability of fraud was above a certain threshold (say, 0.50), they would have an objective means to decide whether to continue the investigation. The p-value, by itself, does not provide this. In fact, depending on the power of the statistic and the population probability of test fraud, the probability that test fraud occurred could be quite small, even if the p-value is very small. For example, suppose that an answer-copying statistic has power of 0.60 (assuming an α=0.01 test) and that copying occurs among 1 in 1,000 of all same-center examinee pairs. By Bayes' Theorem (Equation 1, forthcoming), the probability that copying occurred, given an observed p-value less than 0.01, is just 0.057. In this case, 0.057 is the posterior probability of copying (PPC).

Individuals that understand laws of probability will recognize that a significant p-value does not necessarily suggest that copying occurred. However, other individuals involved with investigations of test fraud (e.g., investigators, education officials, case panelists, arbitrators, and examinees) cannot be expected to understand the complex relationship between the p-value,

power, rate of copying in the population, and the PPC. Hence, there is ample opportunity for the p-value to be misinterpreted or judged out of context. For example, individuals might incorrectly interpret the p-value as the "probability that copying did not occur" (e.g., the complement of the PPC). If the PPC could be reasonably estimated, it could be reported alongside p-values and facilitate a greater understanding of the evidence (or lack thereof) of copying.

To use Bayes' Theorem to arrive at the PPC, we must know the probability distribution function (PDF) of the answer-copying statistic when copying occurs, as well as the population rate of copying among same-center examinee pairs. While the PDF can be estimated through simulation after specifying certain levels of copying, neither it nor the rate of copying in the population is known in practice. In this paper, we develop a formula for the PPC which does not involve these unknown quantities and demonstrate how to estimate the components of the formula.

Methods

The formula for the PPC is derived using laws of probability – namely the multiplication rule and law of total probability (Hogg and Craig, 1995, pp. 21-23). In the derivation, an assumption must be made, and later we discuss the legitimacy of the assumption. We argue that the formula for the PPC can be employed with any answer-copying statistic, and we illustrate how it can be used with a variant of the Omega Index (denoted ω; Wollack, 1996), which has been shown to have strong power and sensitivity (Sotaridona & Meijer, 2003; Zopluoglu & Davenport, 2012). To study how well the PPC approximates the true probability of copying, simulation is used where the PDF of an answer-copying statistic is used to estimate the PPC. To illustrate a real-world application of the formula, we estimate the PPC for various levels of the

answer-copying statistic using data from a large assessment program. We now explain these steps in greater detail.

Developing a formula for the posterior probability of copying

We define the population of interest as examinee pairs who test at the *same center*. Our initial expression of the PPC, $\text{pr}(Y = 1|X = a, C = 1)$, is appropriate when *X* is a discrete random variable. More generally, the PPC can be written as $\text{pr}(Y = 1|\ |X - a| < \delta, C = 1\ )$. In practice, $\delta = 0.5$ should be used when the answer-copying statistic is discrete. If the answer-copying statistic is continuous, smaller values of δ can be used. For example, if a=5 and $\delta = 0.125$, the PPC is the probability that copying occurred, given that the answer-copying statistic (*X*) was between 4.875 and 5.125. To simplify notation, we write the PPC as $\text{pr}(Y = 1|\ X \approx a, C = 1)$.

Our goal is to establish an expression for the PPC that is a function of estimable quantities. Using Bayes' Theorem, the PPC can be written as:

Equation 1: Posterior Probability of Copying Based on Bayes' Theorem

$$\text{pr}(Y = 1|\ X \approx a, C = 1) = \frac{\text{pr}(X \approx a|\ Y = 1, C = 1) \times \text{pr}(Y = 1|C = 1)}{\text{pr}(X \approx a|\ Y = 1, C = 1) \times \text{pr}(Y = 1|C = 1) + \text{pr}(X \approx a|\ Y = 0, C = 1) \times \big(1 - \text{pr}(Y = 1|C = 1)\big)}$$

This expression has three unknowns: (a) the PDF of *X* for same-center examinee pairs that copied (*Y*=1, *C*=1), (b) the population probability of copying among same-center examinee pairs, and (c) the PDF of *X* for same-center examinee pairs that did not copy (*Y*=0, *C*=1), which can be estimated using traditional answer-copying statistics that estimate the PDF of *X* under the assumption of no copying. Because *Y* is unobserved, neither (a) nor (b) are known or directly estimable. Thus, we sought a different formula for the PPC. We applied laws of probability to

derive a formula for the lower bound of the PPC. In the appendix, a step-by-step derivation of the formula is provided, resulting in Inequality 1.

Inequality 1: Lower Bound of Posterior Probability of Copying for Same-Center Pairs

$$\text{PPC} \geq 1 - \frac{\text{pr}(X \approx a | C = 0)}{\text{pr}(X \approx a | C = 1)}$$

Let P_0=pr$(X \approx a | C = 0)$ and P_C=pr$(X \approx a | C = 1)$ represent the PDFs of X for different-center pairs and same-center pairs, respectively. We use the P_0 and P_C notation for simplicity going forward. The lower bound of the PPC is thus based on the ratio of the PDFs of X for different-center and same-center examinee pairs. As the ratio decreases, the lower bound of the PPC increases. For example, if an extreme value is twice as likely to be observed for same-center pairs, the lower bound of the PPC is 0.50 for examinee pairs at that value. Intuitively, this means that evidence of copying is stronger when a large value of X is more likely to occur for same-center pairs (where traditional copying is possible) than for different-center pairs (where traditional copying is not possible).

To arrive at Inequality 1, we assumed that copying could not occur for different-center pairs (consistent with our definition of traditional copying) and refer to this as Assumption 1. We also assumed that the PDF of X for large values of X is the same for examinee pairs from different test centers ($C = 0$) and for examinee pairs from the same center for whom copying did not occur ($C = 1, Y = 0$); we refer to this as Assumption 2. Later, we discuss the legitimacy of Assumption 2 and consequences of its violation.

Because Inequality 1 does not involve the unobserved random variable Y (the indicator for whether copying occurred), we have expressed the lower bound of the probability of interest as a function of quantities that can be estimated with large samples of same-center and different-center examinee pairs.

An alternative form of the PPC can be expressed as $\text{pr}(Y = 1|X \geq a, C = 1)$ by conditioning on *X* being greater than or equal to *a* instead of approximately equal to *a*. This form of the PPC leads to an inequality of the same form as Inequality 1:

Inequality 2: Alternative Form of Lower Bound of Posterior Probability of Copying for Same Center Pairs

$$\text{Alternative form of PPC} \;\geq\; 1 - \frac{\text{pr}(X \geq a|\, C = 0)}{\text{pr}(X \geq a|\, C = 1)}$$

In practice, the alternative form of the PPC is not useful for estimating the probability of copying for an examinee pair. Instead of estimating the probability of copying for an observed value of the answer-copying statistic, it estimates the probability of copying for all values of the answer-copying statistic greater than or equal to the observed value. So, for example, for an examinee pair with an answer-copying z-score of 3.0, it would estimate the probability of copying among examinee pairs with a z-score of 3.0 or larger, leading to an inflated estimate of the probability of copying for the examinee pair. Therefore, we restrict our attention to the form of the PPC given in Inequality 1.

The formula for the lower bound of the PPC can be applied with any answer-copying statistic *X*, as long as it's expressed as a test statistic or p-value under the null hypothesis of no copying. In the case of the Omega Index, *X* can be a function of the p-value (e.g., $X = |\log_{10}(p)|$, such that a p-value of .0001 converts to *X*=4) or *X* can be the z-score used to derive the p-value. In this study, we let *X* be the z-score of an Omega Index variant. Next, we describe a simulation study designed to evaluate the performance of the PPC estimator.

Simulation study

The purpose of the simulation was to study the performance of the PPC estimator when used with a specific answer-copying statistic (an Omega Index variant) and under realistic

simulation scenarios. Data were generated using the nominal response IRT model (Bock, 1972). Under this model, the probability of selecting response alternative *k* of item *i* for person *j* is given by p_{ijk} in Equation 2.

Equation 2: Nominal Response Model

$$p_{ijk} = \frac{\exp(\alpha_{ik} + \beta_{ik}\theta_j)}{\sum_{k=1}^{m} \exp(\alpha_{ik} + \beta_{ik}\theta_j)}$$

where α_{ik} is the threshold parameter for response option *k* of item *i*, $\beta_{ik}$is the slope parameter for response option *k* of item *i*, and θ_j is the ability level of examinee *j*. The nominal response model item parameters were the same as those used in two previous studies (Wollack, 1996 and Zopluoglu & Davenport, 2012) for a 40 item multiple-choice test with 5 response options for each item. For each item, the five threshold parameters were constrained to sum to 0; similarly the five slope parameters were constrained to sum to 0. The nominal response model provides a framework for generating item response data that reflects the complexity of each response probability depending on ability.

Data were simulated with the following steps for 5,000,000 same-center examinee pairs and 5,000,000 different-center examinee pairs. **First**, the ability level for one examinee (the potential copier) was drawn from a standard normal distribution, and the ability level of the other examinee (the potential source of copying) was set equal to the ability of the first examinee, plus a random number drawn from a uniform [0,1] distribution. Therefore, the ability of the potential source is greater than or equal to the ability of the potential copier, reflecting the situation where a copier seeks a source of greater ability from whom to copy.[2] **Second**, item responses were generated using the nominal response model. **Third**, among 5% of the same-center examinee

[2] In reality, examinees do not know the ability level of other examinees. However, examinees may be able to guess which other examinees are likely to perform well on the test because of their performance on prior tests or because of familiarity with their academic achievements in school.

pairs, level of copying was randomly assigned, with number of items copied specified as *m*=4, 8, 12, …,40 out of the 40 total items (the same copying levels studied by Zopluoglu & Davenport, 2012). For copied items, the source's responses replaced the suspect's responses. Similar to Zopluoglu & Davenport (2012), a random copying mechanism was used, whereby the suspect randomly chooses the *m* items to copy among all 40 items on the test. **Fourth**, each examinee's raw score (number of correct items) was calculated. **Fifth**, a nonparametric variant of the Omega Index was calculated for each of the 10,000,000 examinee pairs. The Omega Index variant is calculated by converting the total number of identical responses to a z-score, and then converting the z-score to a p-value assuming normality. Using the actual Omega Index, the mean and the standard deviation used in the z-score calculation are based on the nominal response model (for full details on how the Omega Index is calculated, please see Wollack, 1997). For the Omega Index variant, instead of using the nominal response model conditioned on latent ability, the item response probabilities were conditioned on raw score (ranging from 0 to 40). Response probabilities were estimated as the simple proportion of examinees (for each possible raw score) that selected each response option. This approach is feasible with large samples, as are used in this simulation.

When implementing the PPC estimator in practice, one must choose the minimum z-score for which to calculate the PPC. PPC estimates will be negative when $P_0 > P_C$. In practice, to avoid a probability estimate that is negative, the PPC estimate should be set to 0 when the formula gives a negative result. For this study, we chose z=1.50 as the minimum z-score for which to calculate the PPC because z-scores less than 1.50 would not typically be considered evidence of copying.

Each examinee pair's z-score was rounded to the nearest 0.25. Z-scores less than 1.50 were considered not suggestive of copying, and z-scores greater than 6.00 were grouped with those near 6.00 (as we will see later, the PPC estimate does not change much for z-scores greater than 6). This yields z-score levels of <1.50, 1.50, 1.75, …, 5.75, and 6+. These z-score levels are observed in practice and represent a plausible range of answer-copying statistic values. At one extreme, z<1.5 represents a case where no additional evidence of copying arises from the answer-copying statistic. At the other extreme, $z \geq 6$ represents a case where evidence of copying would seem very strong.

Step 3 of the simulation specified the population rate of copying at 5% among same-center pairs. In practice, this rate is unknown and may vary considerably by examinee population, test stakes, and test security procedures. Because the population rate of copying has a large influence on the PPC (Equation 1), it is important to study the performance of the PPC under different rate scenarios. Note that the rate of copying among examinee *pairs* must be distinguished from the rate of copying among *examinees*. For example, suppose that 10 of 50 examinees in a test center partake in copying, and that each copier has his or her own source. In this case, there are 1,225 examinee pairs and the rate of copying is 10/1,225 = 0.8% among examinee pairs, but 20% among examinees. Population rates of copying among pairs can be very small, so we repeated the simulation using population rates of copying of 0.05% and 0.5%.

We used the simulated data to estimate P_0 (based on different-center pairs who could not have copied) and P_C (based on same-center pairs of whom 100p% copied), resulting in estimates of the lower bound for the PPC (from Inequality 1). The lower bound of the PPC was estimated separately for each z-score level. The lower bound of the PPC should increase with larger z-scores. We calculated the actual proportion of same-center examinees who copied, conditional

on z-score. If the PPC estimator is performing well, the PPC estimate should be slightly smaller than the actual proportion who copied.

Example estimation of the posterior probability of copying

We applied the PPC estimator to actual data from the ACT Explore Mathematics test which consists of 30 multiple-choice items, each with five response options (ACT, 2013). ACT Explore is typically administered during a regular school day, and so test center is the same as school. We used data from three testing years and the most common test form used in each year. We removed examinees that had guessing-patterned responses, defined as 10 or more consecutive items with the same response. From these records, we formed 83.4 million same-center (same-school) pairs and 182.6 million different-center pairs who were enrolled in the same school district. Because we are interested in rare events (extreme values of X), large samples of examinee pairs are needed to estimate P_0 and P_C. Later, we discuss how statistical modeling has potential to reduce the large sample size burden.

For each examinee pair, the Omega Index variant was calculated. Again, instead of using the nominal response model conditioned on latent ability, the item response probabilities were conditioned on raw score (ranging from 0 to 30). While the ACT Explore Mathematics test has five response options, it is also possible for examinees to not respond, so we used six response options including missing. Similar to the simulation, each examinee pair's z-score was rounded to the nearest 0.25, resulting in z-score levels of <1.5, 1.50, 1.75, …, 5.75, 6+.

Results

Simulation study results

Using Inequality 1, the PPC estimates are based on P_0 (the PDF of the Omega Index z-score for different-center examinee pairs) and P_C (the PDF of the Omega Index for same-center examinee pairs). While PDFs can be estimated using statistical models, we used the simple proportion of examinee pairs at each z-score level to estimate the PDFs. Figure 1 shows the estimated PDFs for Omega Index z-score levels between 1.50 and 6.00 from the simulation study when the population rate of copying was 0.05. In Figure 1, the two PDFs do not have equal area under the curves (and do not integrate to 1) because only a portion of the function's domains are shown ($1.5 \leq X$). The PDF for same-center pairs (P_C) is close to the PDF for different-center pairs (P_0) for z-score levels between 1.5 and 2.0, and then the difference becomes larger. For z-scores between 3.50 and 5.75, P_0 approaches 0 as the z-score increases while P_C holds relatively steady. We see a spike in P_C at the z-score=6 level because z-scores greater than 6 were included.

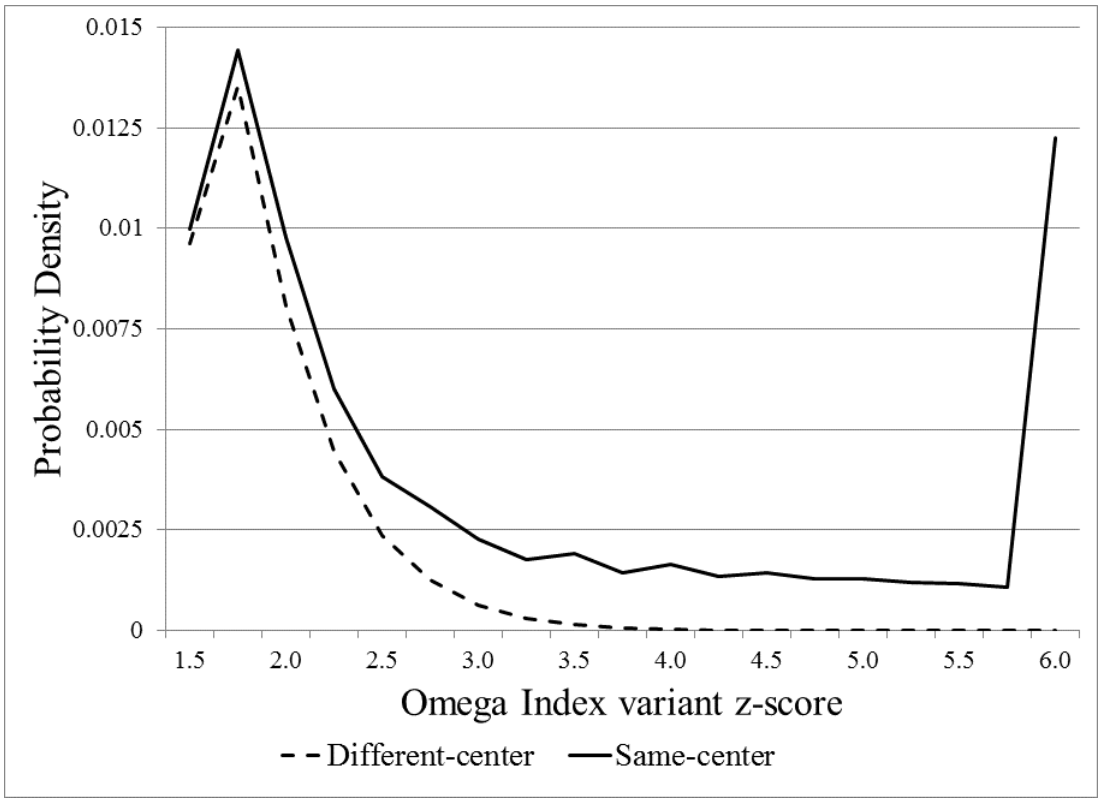

Figure 1: Right-hand tails of Omega Index variant z-score distributions when population rate of copying is 0.05

The PPC estimates follow from the P_0 and P_C estimates. For example, consider the case of $z \approx 3.50$ when the population rate of copying is 0.05 (Figure 1). P_0 is estimated as 0.00013 and P_C is estimated as 0.00191. Therefore, the PPC is estimated as 0.93 (1 - 0.00013 / 0.00191). Figure 2 shows the PPC estimates, by z-score level, resulting from the simulation study for the three levels of *p* (population rate of copying): 0.05, 0.005, and 0.0005. As expected, the estimated PPC increases sharply with z-score level. For example, when p=0.005 (the middle set of curves), the PPC estimate is 0.02 for $z \approx 2$, 0.23 for $z \approx 3$, 0.88 for $z \approx 4$, and 0.99 for $z \approx 5$. The probability of copying increases very sharply after z-scores of 2.50, and begins to plateau at $z \approx 4.00$.

The performance of the PPC estimator can be evaluated by comparing the PPC estimates to the actual proportion who copied, which is also plotted in Figure 2. For example, when p=0.0005, among examinee pairs with z≈4.25, the proportion who copied was 0.68 and the corresponding PPC estimate was 0.63. For z≈5.75 and z≈6.00, the PPC estimates (1.00) matched the true proportion.

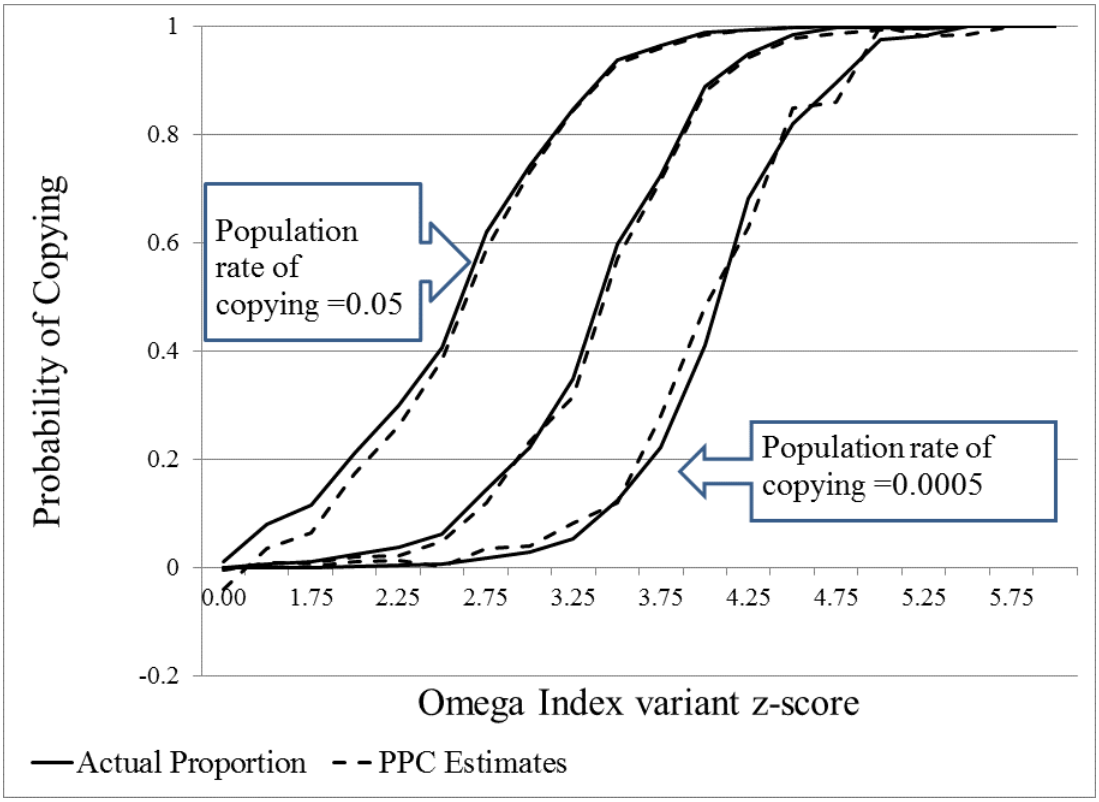

Figure 2: Comparing the PPC estimates to the actual proportion who copied

From Figure 2, we see that the PPC tends to closely estimate the true proportion who copied. As the population rate of copying increases, larger PPC values will result for the same z-score level. Also as the population rate of copying increases (e.g., when the population rate of copying is 0.05), the PPC tends to underestimate the true proportion for lower z-score values. As the population rate of copying becomes smaller, there are fewer extreme cases of the answer-

copying statistic z-scores. This, in turn, causes the estimates of P_C and the PPC to be less precise. In Figure 2, the curves for p=0.0005 are not as smooth as the other curves, reflecting the drop in precision.

Example application of PPC estimator: ACT Explore Mathematics data

The PPC estimator was applied to large samples of examinee pairs who took the ACT Explore Mathematics test as 8th graders. In Figure 3, the PPC estimates are plotted by z-score, again using z-score intervals of width 0.25. The figure shows general consistency across testing years in the relationship of z-score and PPC estimates. The PPC is relatively stable between z-scores of 1.50 and 3.75, but then increases sharply between z-scores of 3.75 and 6.00.

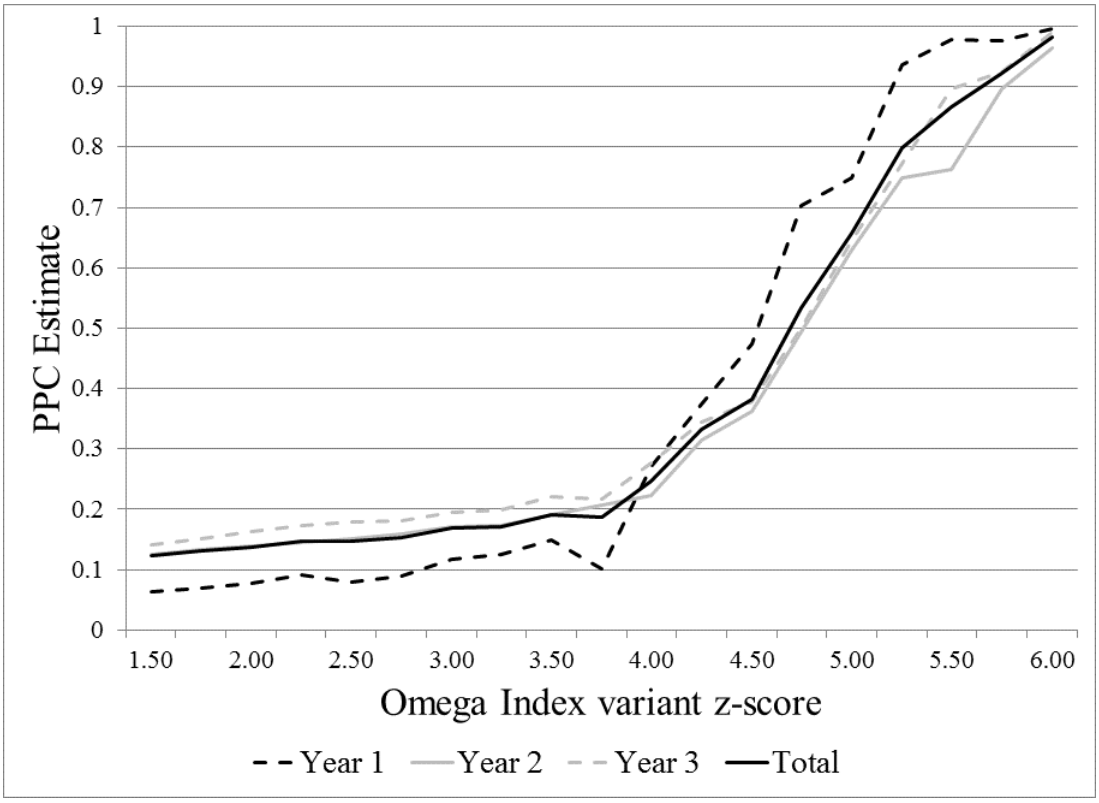

Figure 3: PPC Estimates, by Omega Index variant z-score and ACT Explore test year/form

Contrasting Figure 2 (simulated data) to Figure 3 (real data), we find that the PPC curves are smoother for the simulated data, despite the larger sample sizes used for the real data. For $z>3.75$, the PPC estimates are considerably smaller for the real data. This suggests that either 1) the population rate of high levels of copying in the ACT Explore population is less than the simulated scenarios, 2) the power of the answer-copying statistic is greater in the simulated data sets, or both 2) and 3). For smaller z-score levels (e.g., $z\approx 2$), the PPC estimates based on the real data are larger than those based on the simulated data.

Discussion

Like other measures used to detect cases of potential test fraud, answer-copying statistics have traditionally used "p-values" that measures the cumulative probability distribution of the statistic, under the null hypothesis of no copying. Responding to calls for estimating the probability of test fraud (Wainer, 2012), we developed a formula for the probability of copying, conditional on observed answer-copying statistics. We applied the formula to a simulated data set, as well as an actual data set of same-center and different-center examinee pairs who took the ACT Explore Mathematics test.

Proposed uses of the PPC estimator

Through simulation, we found that the PPC estimates closely matched the actual proportion of examinee pairs who copied, especially when the answer-copying statistic (Omega Index variant z-score) was very large. Therefore, the PPC estimator can be used to approximate the probability that copying occurred. This finding has practical implications for how parties involved with investigations of test misconduct interpret answer-copying statistics. For example, the PPC estimate could be used by an investigative team to better understand the risk of making

type 1 (false positive) and type 2 (false negative) errors when acting upon observed answer-copying statistics. For cases of extremely large statistics, PPC estimates may approach 1, in which case the estimate could be used to inform others of the overwhelming weight of the evidence in favor of copying.

The relationship between answer-copying statistic value (e.g., Omega Index z-score) and the PPC is affected by the population rate of copying among same-center examinee pairs. From Bayes' Theorem (Equation 1) and the simulation study, we see that larger PPC values result when the population rate of copying increases. The relationship between the answer-copying statistic value and the PPC is also affected by the power of the test: As power increases, so too does the PPC (Equation 1). In this study, we only estimated the PPC using one answer-copying statistic (the Omega Index variant z-score). If multiple answer-copying statistics are used, the PPC estimator provides a simulation-free method to compare the power of the competing statistics.

We applied the PPC estimator to actual data from the ACT Explore Mathematics test, administered to 8th graders. We found that the estimated probability of copying increased only slightly between z-scores of 1.50 and 4.00, but then increased sharply for z-scores between 4.00 and 6.00. For z-scores of 6 and larger, the estimated probability of copying ranged from 0.937 to 0.990 across three testing years/forms, suggesting a strong likelihood that copying occurred. Combining data across the three years, the PPC estimate exceeded 0.5 for z-scores of approximately 4.75 (between 4.625 and 4.875). This result is somewhat surprising as it suggests that among examinee pairs with a z-score of 4.50, copying occurred for less than half of the cases.

The PPC estimator can also be used to set cutoffs for answer-copying statistic values that trigger investigations. For example, based on the results in Figure 3 for the ACT Explore Mathematics test, an Omega Index variant z-score of approximately 4.50 has a PPC estimate of 0.38 and a z-score of approximately 4.75 has a PPC estimate of 0.53. If one wished to choose a z-score cutoff related to a 50% chance of copying, the cutoff should lie somewhere between 4.50 and 4.75.

By applying the method to three independent samples of ACT Explore Mathematics test administrations, we were able to observe the stability across test forms/years of the PPC estimates. Generally, the relationship of z-score to PPC estimate was quite similar across the three administrations (Figure 3). The PPC estimator can be applied to any test with a multiple choice format and large samples of same-center and different-center examinee pairs. It would be prudent to estimate the z-score-to-PPC relationship for each test administration, and to examine whether the pattern varies from the patterns observed in previous administrations (see Figure 3 for example).

Revisiting Assumption 2

In the development of the formula for the lower bound of the PPC, we made an assumption which allowed us to express the formula in terms of estimable quantities. We assumed that, for large values of X (the answer-copying statistic), the probability distribution of X is the same for different-center pairs relative to same-center pairs who did not copy. If the assumption fails, Inequality 1 does not hold. If the right hand tail of the probability distribution of X were actually larger for same-center examinee pairs (who did not copy), relative to different-center examinee pairs, it could reverse the direction of the inequality.

Logic and prior research (Allen, 2012) suggest that students with shared educational and life experiences have slightly higher levels of item response similarity. Thus, the legitimacy of Assumption 2 may depend on the extent that same-center examinee pairs share educational and life experiences, relative to the different-center examinee pairs. In the simulation study, we generated the data in a way that did not violate the assumption: Specifically, the answer-copying statistic was a function of examinee ability level and chance and so same-center examinees who did not copy were expected to have the same distribution of X as different-center examinees.

When applied to actual data, we sampled same-center examinee pairs as well as different-center *same-school district* pairs. Prior research suggests that examinee pairs who attend the same school have higher similarity levels than those who do not (Allen, 2012). By requiring the different-center pairs to have attended school in the same district, we reduced the chance that the key assumption is seriously violated. When the PPC estimator is used in practice, efforts should be made to construct the different-center pair samples in such a way that they are as likely as the same-center pairs to have shared educational and life experiences. For example, "same-center" pairs could be students from the same school who tested in the same classroom, and "different-center" pairs could be students from the same school who tested in different classrooms. In this case, there is still a chance that Assumption 2 is violated because same-classroom students may have received the same instruction, making them more prone to higher similarity levels.

In practice, investigators are most concerned with large values of X. Assumption 2 therefore actually has two parts of practical importance: 1) the probability distribution of X is the same for different-center pairs relative to same-center pairs who did not copy, and 2) X is large enough for the first part to hold. From Figure 2, we see that the PPC underestimates the actual rate of copying when X is smaller (e.g., less than 3), especially when the true rate of copying is

high. For larger values of X and when the actual rate of copying is low, Assumption 2 may be more likely to hold.

Assumption 2 can be expressed as $\text{pr}(X \approx a | C = 1, Y = 0) = \text{pr}(X \approx a | C = 0) = P_0$. Because same-center pairs may be more likely to share educational and life experiences (and thus have larger values of X), we are concerned that $\text{pr}(X \approx a | C = 1, Y = 0) > \text{pr}(X \approx a | C = 0)$, or $\frac{\text{pr}(X \approx a | C=1, Y=0)}{\text{pr}(X \approx a | C=0)} = \partial > 1$. In this case, the lower bound of the PPC is $1 - \frac{\partial P_0}{P_C}$. To illustrate the consequences of violating Assumption 2, we can examine what would happen to PPC estimates for different values of ∂. In Figure 4, we plot the same data from Figure 3 (PPC estimates for ACT Explore Mathematics data) for ∂=1.0 (no violation), 1.1 (10% violation), 1.2 (20% violation), and 1.3 (30% violation).

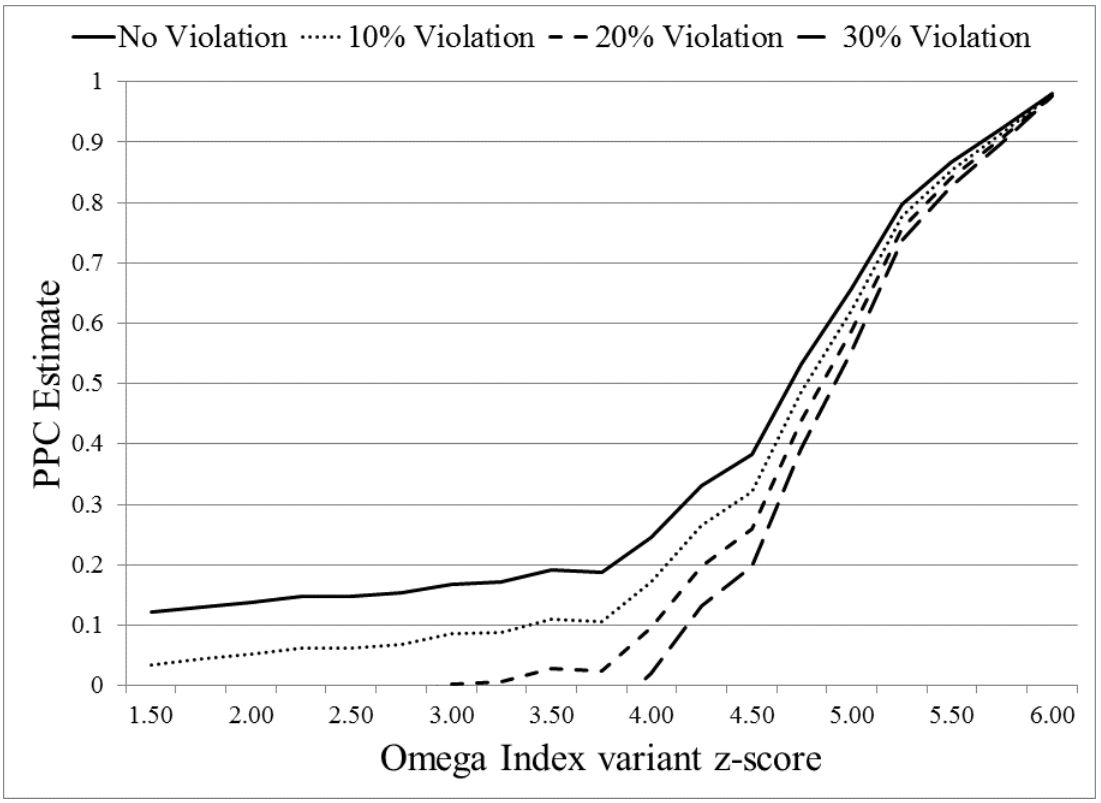

Figure 4: Effect of Assumption 2 violations on PPC estimates

From Figure 4, we see that violations to Assumption 2 have lesser consequence as X increases. As ∂ increases, the PPC estimate should decrease. For the ACT Explore example, the estimate of the lower bound of the PPC is 0.66 when X=5 when there is no violation of the assumption, 0.62 when there is a 10% violation, 0.59 when there is a 20% violation, and 0.55 when there is a 30% violation. When X=6, there is no practical consequence of the violations as all of the PPC estimates are 0.98. For $X < 4.5$, the consequences of violating Assumption 2 are profound.

Generalizing the approach for other types of test fraud

The lower bound of the PPC was expressed in terms of Y (an indicator of whether copying occurred), X (an answer-copying statistic), and C (an indicator of whether the examinee pair tested at the same center). However, Inequality 1 could potentially be applied to other modes of test fraud. For example, Y can more generally represent an indicator of whether test fraud occurred, X can represent the statistic designed to detect test fraud (defined such that larger values of X are more suggestive of fraud), and C can represent an indicator of whether the form of test fraud was possible. In all cases, X and C must be observable, but Y need not be. Generalizing Inequality 1 to a test fraud context other than traditional copying also means that the key assumption (Assumption 2) must hold in the new context.

Limitations and future research

We only considered one answer-copying statistic based on an Omega Index variant. Other procedures, such as the generalized binomial test, may offer greater power (Zopluoglu & Davenport, 2012). Additional research should be done to compare the performance of the PPC estimator for different answer-copying statistics.

Our results apply only to traditional copying; that is, copying that occurs when examinees test in the same location at the same time, and use sight, sound, or touch to communicate. With new technologies, nontraditional forms of copying may also be problematic. It may be difficult to generalize Inequality 1 to nontraditional forms of copying because of the lack of an observable variable that indicates whether copying was possible (*C* indicates shared test center for our case of traditional copying). Moreover, nontraditional forms of copying can lead to overestimation of P_0. Even with overestimation of P_0 due to nontraditional forms of copying, Inequality 1 would still hold because the effect of the overestimation would be consistent with the direction of the inequality. If copying does occur for different-center pairs, it can result in an underestimate of the PPC.

In our demonstration of the estimation of the lower bound of the PPC, we limited the scope to one multiple-choice assessment, the ACT Explore Mathematics test. The resulting estimates of the PPC for different levels of answer-copying statistics would likely be different for other assessments. In particular, results could vary by subject area, by examinee population, by test administration protocols designed to prevent copying, and by the stakes of the test. Future research could examine this by repeating the methodology used here with additional assessments and different populations of examinees.

One advantage to using data from the ACT Explore Mathematics test was the abundant sample sizes of examinee pairs that are afforded by a large national testing program. With the large sample sizes, we could estimate P_C (the rate of an extreme answer-copying statistic for same-center pairs) and P_0 (the rate of an extreme answer-copying statistic for different-center pairs) as simple proportions. In practice, investigators may need to estimate the probability of copying based on much smaller sample sizes. Another area of additional research would be to

examine model-based methods (e.g., Poisson, exponential, or other extreme value models) for estimating P_C and P_0 that may offer efficient estimation without needing such large samples. Even with a very large data set, the rates of extreme answer-copying statistics are noisy and less reliable when very low levels of copying are present.

Even with extremely large samples, there is still uncertainty in the PPC estimates. In this study, we did not attempt to calculate standard errors of the PPC estimates. While the PPC estimator is a very simple expression, the calculation of the standard error is confounded by violations of independence across examinee pairs. Because individual examinees are members of multiple pairs, there is autocorrelation and the independence assumption is violated. Additional research is needed to estimate confidence intervals for the lower bound of the PPC, accounting for the autocorrelation.

This paper demonstrated a systematic approach for estimating the probability of traditional copying. While it is useful to have an estimate of the probability of copying, such measures should not be used in place of preventive efforts, such as adherence to test administration protocols and training personnel (e.g., students, teachers, proctors) on acceptable and unacceptable behaviors (NCME, 2012). Investigations of test fraud should carefully evaluate the reasonableness of the assumptions needed to estimate the probability of copying, and consider alternative explanations for extreme values of an answer-copying statistic. For example, examinees employing the same guessing strategy can cause a shift in the distribution of the answer-copying statistic, and lead to an overestimate of the PPC. The probability of copying should be considered along with other forms of available evidence, including analyses of changes in test scores from previous years or test administrations, erasures, scratch work needed to solve problems, and comparisons with course performance.

References

ACT. (2013). *ACT Explore technical manual.* Iowa City: Author. Retrieved from http://www.act.org/explore/pdf/TechManual.pdf.

Allen, J. (2012). *Relationships of examinee pair characteristics and item response similarity*. (ACT Research Report 2012-8). Iowa City, IA: ACT, Inc.

Angoff, W.H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69*(345), 44-49.

Belov, D.I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement, 35*(7), 495-517.

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

Hogg, R.V. & Craig, A.T. (1995). *Introduction to Mathematical Statistics*. Prentice-Hall: Englewood Cliffs, NJ.

Holland, P.W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (ETS Technical Report No. 96-4). Princeton, NJ: Educational Testing Service.

National Council on Measurement in Education (NCME). (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Washington, D.C.: Author.

National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education (2013). *Testing integrity: Issues and recommendations for best practice*.

Sheu, C.F., Chen, C.T., Su, Y.H., Wang, W.C. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods, 37(2)*: 202-218.

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

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. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*, 283-304.

Wainer, H. (2012). Cheating: Some ways to detect it badly. *Chance, 25*(3), 54-57.

Weslowsky, G.O. (2000). Detection of excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics, 27*(7), 909-921.

Wollack, J.A. (1996). Detection of answer copying using item response theory. *Dissertation Abstracts International*, 57/05, 2015.

Wollack, J.A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, 22, 144-152.

Zopluoglu, C., & Davenport, E.C. (2012). The empirical power and type 1 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: Proof of Inequality 1

Expression Note: Highlighted text shows part of expression that changed at each step	**Rationale for step**
$\text{pr}(Y = 1 \mid X \approx a, C = 1) = \frac{\text{pr}(X \approx a, C = 1, Y = 1)}{\text{pr}(X \approx a, C = 1)}$	Multiplication rule
$= \frac{\text{pr}(X \approx a) \ - \ \text{pr}(X \approx a, C = 1, Y = 0) \ - \ \text{pr}(X \approx a, C = 0, Y = 0) \ - \ \text{pr}(X \approx a, C = 0, Y = 1)}{\text{pr}(X \approx a, \ C = 1)}$	Law of total probability
$= \frac{\text{pr}(X \approx a) \ - \ \text{pr}(X \approx a, C = 1, Y = 0) \ - \ \text{pr}(X \approx a, C = 0, Y = 0) - 0}{\text{pr}(X \approx a, C = 1)}$	$\text{pr}(X \approx a, C = 0, Y = 1) = 0$ Traditional copying cannot occur for different-center pairs (Assumption 1)
Let $P_0 = \text{pr}(X \approx a \mid C = 0)$, let $P_C = \text{pr}(X \approx a \mid C = 1)$, and let $Q \ = \ \text{pr}(X \approx a) = P_0 \times \text{pr}(C = 0) + P_C \times \text{pr}(C = 1)$	
$= \frac{Q \ - \ \text{pr}(X \approx a, C = 1, Y = 0) \ - \ \text{pr}(X \approx a, C = 0, Y = 0)}{P_C \ \times \ \text{pr}(\ C = 1)}$	Law of total probability
$= \frac{Q \ - \ \text{pr}(X \approx a \mid C = 1, Y = 0) \ \times \ \text{pr}(C = 1, Y = 0) \ - \ \text{pr}(X \approx a \mid C = 0, Y = 0) \ \times \ \text{pr}(C = 0, Y = 0)}{P_C \ \times \ \text{pr}(\ C = 1)}$	Multiplication rule
$= \frac{Q \ - \ \text{pr}(X \approx a \mid C = 1, Y = 0) \ \times \ \text{pr}(C = 1, Y = 0) \ - \ P_0 \ \times \ \text{pr}(C = 0, Y = 0)}{P_C \ \times \ \text{pr}(\ C = 1)}$	$\text{pr}(X \approx a \mid C = 0, Y = 0) = P_0$ because all different-center pairs have Y=0
$= \frac{Q \ - \ \text{pr}(X \approx a \mid C = 1, Y = 0) \ \times \ \text{pr}(C = 1, Y = 0) \ - \ P_0 \ \times \ \text{pr}(C = 0)}{P_C \ \times \ \text{pr}(\ C = 1)}$	$\text{pr}(C = 0, Y = 0) = \text{pr}(C = 0)$ because traditional copying cannot occur for different-center pairs
$= \frac{Q \ - \ P_0 \ \times \ \text{pr}(\ C = 1, Y = 0) \ - \ P_0 \ \times \ \text{pr}(C = 0)}{P_C \ \times \ \text{pr}(\ C = 1)}$	$\text{pr}(X \approx a \mid C = 1, Y = 0) \ = P_0$ Assumption 2
$= \frac{P_C \ \times \ \text{pr}(\ C = 1) \ - \ P_0 \ \times \ \text{pr}(\ C = 1, Y = 0)}{P_C \ \times \ \text{pr}(\ C = 1)}$	Substitution for Q, algebra
$= 1 - \frac{P_0 \ \times \ \text{pr}(\ Y = 0 \mid C = 1)}{P_C} \geq \ 1 - \frac{P_0}{P_C}$	Multiplication rule, algebra, $\text{pr}(Y = 0 \mid C = 1) \leq 1$