

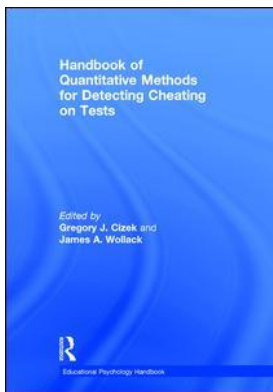
This article was downloaded by: 10.3.98.104

On: 29 Sep 2020

Access details: *subscription number*

Publisher: *Routledge*

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: 5 Howick Place, London SW1P 1WG, UK



## **Handbook of Quantitative Methods for Detecting Cheating on Tests**

Gregory J. Cizek, James A. Wollack

### **Detecting Potential Collusion among Individual Examinees using Similarity Analysis**

Publication details

<https://www.routledgehandbooks.com/doi/10.4324/9781315743097.ch3>

Dennis D. Maynes

**Published online on: 14 Oct 2016**

**How to cite :-** Dennis D. Maynes. 14 Oct 2016, *Detecting Potential Collusion among Individual Examinees using Similarity Analysis* from: *Handbook of Quantitative Methods for Detecting Cheating on Tests* Routledge

Accessed on: 29 Sep 2020

<https://www.routledgehandbooks.com/doi/10.4324/9781315743097.ch3>

**PLEASE SCROLL DOWN FOR DOCUMENT**

Full terms and conditions of use: <https://www.routledgehandbooks.com/legal-notices/terms>

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# 3

## DETECTING POTENTIAL COLLUSION AMONG INDIVIDUAL EXAMINEES USING SIMILARITY ANALYSIS

*Dennis D. Maynes*

### INTRODUCTION

This chapter demonstrates how potential collusion among individual examinees using similarity statistics may be detected. Most of the research in this area has been conducted using answer-copying statistics (Frery, Tideman, & Watts, 1977; Wollack, 1997). These statistics have generally been used to determine whether a person, known as the copier, potentially copied answers from another person, known as the source. However, this approach is limited in practice because it does not account for two or more test takers working together to breach the security of the exam by communicating and sharing test content and/or answers to questions during the exam. The source-copier approach also does not provide a mechanism to detect collusion that resulted from communication external to the testing session. In exchange for setting aside the assignment of a potential source and copier, similarity statistics are able to detect potential collusion in these and other situations. Although answer copying and similarity statistics are related, the focus in this chapter will be entirely on answer similarity statistics. A more complete discussion of answer copying statistics can be found in Zopluoglu (this volume).

With the pervasive use of technology for testing and for cheating, cheating by copying answers is no longer the primary threat that it once was. Indeed, electronic devices have been used by test takers to communicate and share answers in real time across great distances. Hence, similarity statistics provide a means of detecting general forms of collusion that are not as easily detected by answer-copying statistics. In recent years, researchers have been exploring these more general forms of potential test fraud (Zhang, Searcy, & Horn, 2011; Belov, 2013).

### PROPERTIES OF SIMILARITY STATISTICS

In published research for answer-copying and similarity statistics, most researchers have restricted their work to pairs of students where potential answer-copying was

present. The more general test security threat of collusion has not received much attention by academic researchers. Collusion describes a testing scenario where two or more examinees are working together for the personal gain of at least one of those examinees. Answer copying is a special case of collusion for which the source examinee may not realize his or her own involvement. However, collusion also extends to cover instances where examinees are deliberately sharing their answers; communicating during the exam, verbally or nonverbally; working together before the exam to share content (i.e., acquire preknowledge); using surrogate or proxy test takers; or receiving disclosed answers to test questions by instructors, teachers, and trainers. This chapter focuses on the role of similarity statistics in helping to detect test collusion. Wollack and Maynes (this volume) describe an application of the approach discussed here for purposes of detecting and extracting groups of examinees among whom collusion was potentially present.

Answer copying statistics require specifying both a copier and a source, and because the statistic is conditional on the responses provided by the source examinee, the value of the statistic is different depending on which examinee is treated as the copier and which as the source. In contrast, similarity statistics have the property of symmetry; that is, they produce the same values regardless of whether one examinee is a source and another is an information beneficiary. Symmetry is a desirable property because a copier does not need to be hypothesized and because it allows for clustering techniques to be used to extract a group structure.

Another desirable property of similarity statistics is that their probability distributions are based on the assumption of independence, hence allowing for a specific statistical hypothesis of independent test taking to be tested. The term *nonindependent test taking*, instead of *cheating* or *collusion*, is used in this chapter describing the scenario when the hypothesis of independent test taking is rejected, even though similarity statistics are designed to detect potential cheating and/or collusion.

One point of debate among similarity (or answer copying) researchers relates to which items should be considered (see Wollack & Maynes, 2011). The heart of this debate is whether or not matches on correct items should be considered as providing evidence of misconduct. Clearly, each identical incorrect answer provides some evidence of wrongdoing, and many indexes focus exclusively on matches of this variety (Holland, 1996; Bellezza & Bellezza, 1989). However, others have argued that valuable information can also be extracted from correct answers, even if not as much, and so have designed indexes that consider all items (Wesolowsky, 2000; Wollack, 1997; van der Linden & Sotaridona, 2006).

It is the opinion of this author that using the entire set of test responses is preferred because the outcome of a response may be viewed as a random event (under assumptions of local independence within IRT models). When the similarity statistic is based on identical incorrect responses only, a critical piece of evidence is ignored. The purpose of cheating on tests is to get a high score. This is only possible when identical correct responses are present. For example, suppose two or more test takers have gained access to a disclosed answer key containing most, but not all, of the answers. Statistics that focus exclusively on identical incorrect responses will likely not detect this situation (assuming that the incorrect responses were from items with nondisclosed answers). Statistically, the probability distribution of identical incorrect responses depends upon the number of identical correct responses (see Table 3.2 in this chapter for this exposition). The number of identical correct responses may be extreme, but when it is used to condition the probability distribution (i.e., when it is treated as an

variable not providing information about potential test fraud), the extreme improbability of observed similarities will not be accurately portrayed.

## MAKING STATISTICAL INFERENCES WITH SIMILARITY STATISTICS

There is some debate among practitioners concerning the specific statistical inferences that should be made with respect to collusion on tests. Statistical inferences can and should be made about test score validity (i.e., can you trust the score?). Statistical inferences can also be made about test-taker behavior (i.e., did the person cheat?). There seems to be some confusion among researchers about which inferences can properly be made and how they should be made. Regardless of disagreements among researchers, proper statistical inferences require modeling the statistical distribution of the similarity statistic under the assumption that tests were taken normally (i.e., no cheating or testing irregularity occurred). Thus, the distribution of the similarity statistic needs to be modeled before statistical inferences may be made.

### *Discussion of Some Similarity Statistics*

Because this chapter is concerned with similarity statistics, common answer-copying statistics such as  $\omega$  (Wollack, 1997),  $g_2$  (Frery et al., 1977), and the K index (Holland, 1996) are not discussed. Instead, a few salient similarity statistics are now discussed in chronological order.

Hanson, Harris & Brennan (1987) described two similarity statistics, PAIR1 and PAIR2. These are bivariate statistics, with PAIR1 equal to the number of identical incorrect responses *and* length of the longest string of identical responses, and PAIR2 equal to the number of identical incorrect responses in the longest string of identical responses *and* the ratio of the number of identical incorrect responses to the sum of the nonmatching responses and the number of identical incorrect responses. A theoretical, model-based method for estimating the distributions for these statistics was not provided by the authors. Instead, they suggested that empirical sampling could be used to assess the extremity of a particular pair of similar tests. As a result, probability estimates of observed similarities for these two statistics are not currently computable. Because tail probabilities are not computable for these statistics, they are not suitable to use for detection because appropriate Type I error control cannot be imposed. However, they have been and will likely continue to be used for purposes of supporting allegations of testing improprieties.

Bellezza & Bellezza (1989) introduced error similarity analysis (ESA). The statistic in this analysis is the number of incorrect matching responses between a pair of test takers. The distribution is based on a binomial distribution where the conditional probability of providing a matching wrong answer was estimated by counting the total number of matching wrong answers and dividing by the total number of shared wrong answers in the data set. The same authors published an update (1995) in which they suggested that the single conditional probability estimate for conditionally matching incorrect answers could be replaced with estimates for each item on the test, if desired.

Wesolowsky (2000) introduced a statistic with the probability calculated by a computer program, S-Check. This statistic counts the number of matching answers for a pair of test takers. The probability of a matching answer is computed using the assumption of independence and the performance level of the two takers, where probabilities

of correct answers are estimated by a smooth function “suggested by  $l_p$  distance iso-contours from location theory” (p. 912). The probability that student  $j$  will correctly answer question  $i$  is

$$p_{ij} = [1 - (1 - r_i)^{a_j}]^{1/a_j},$$

where  $a_j$  is a parameter that estimates student ability (similar to the function of  $\theta$  in item response theory) and  $r_i$  is equal to the proportion of students that answered the item correctly (i.e., the item’s  $p$ -value). Probabilities of incorrect matching answers are computed using the product rule (i.e., assuming independence), and probabilities of individual incorrect responses conditioned upon an incorrect response (see Wesolowsky’s paper for details). Wesolowsky states that the number of matching answers follows a generalized binomial distribution, but for sake of computational efficiency he approximated the tail probability using the normal distribution.

Van der Linden & Sotaridona (2006) introduced a similarity statistic, GBT, based on the generalized binomial distribution. This statistic counts the total number of matching answers. The probability of a matching answer is computed using the assumption of independence and the performance level of the two takers, where response probabilities are estimated using the nominal response model (Bock, 1972).

Maynes (2014) described a bivariate statistic, M4, which consists of the number of identical correct and the number of identical incorrect answers. Following van der Linden and Sotaridona, Maynes uses the nominal response model to estimate item response probabilities conditioned upon test taker performance. The probability distribution for this statistic is postulated to follow a generalized trinomial distribution, where probabilities of matching responses are estimated using the assumption of statistical independence and item response probabilities from the nominal response model.

### *Violations of Assumptions*

The derivations of the distributions of these statistics assume (1) responses between test takers are stochastically independent; (2) response probabilities depend upon test-taker performance, which can be modeled using a mathematical model; and (3) item responses are “locally” or conditionally independent and only depend upon test-taker performance. As seen in the previous section, several models have been used by researchers, with the nominal response model being referenced and used most recently. The choice of model will depend upon the data and the computational tools that are available for estimating the model. Model suitability is always a question that should be asked. When possible, the author prefers to verify model appropriateness through goodness of fit tests or by empirically sampling from live data that demonstrate the extremity of pairs detected by the similarity statistic (e.g., similar to the sampling procedures recommended by Hanson et al., 1987).

Extreme values of similarity statistics provide evidence of nonindependent test taking (i.e., some common characteristic or occurrence shared by two individuals has influenced their answer selections). Observed similarity between the test responses (or nonindependent test taking) can result from several situations or factors (including a few nonfraudulent behaviors), most of which are described in this section. Even though the similarity statistics are designed to detect nonindependent test taking, additional information (e.g., obtained by examining seating charts or interviews) may be required to determine the nature of the nonindependent test taking that occurred.

While the statistics may not tell us what happened or the behavior that led to nonindependent test taking, they may provide information for making inferences concerning unobserved behaviors and the trustworthiness of test scores. The degree or amount of nonindependence (i.e., deviation from typical or expected similarity), the scope of nonindependence (i.e., the size of the detected clusters), and the pattern of nonindependence among the responses (i.e., the alignment of matching and nonmatching responses) provide important clues as to why the null hypothesis of independent test taking was rejected.

An understanding of the assumptions of the statistical procedures is critical to the practitioner, because a violation of the assumptions may result in observing extreme values of the similarity statistic for two or more test instances. Understanding causes and effects that can be responsible for nonindependent test taking helps the practitioner evaluate possible explanations for an extreme value of the statistic.

It is often the case that similarity statistics demonstrate robustness when the assumptions do not strictly hold. For example, it is seldom true that all test takers in the population are independent. Indeed, many test takers offer “studying together” as a defense against statistical detection of potential test fraud. In offering this “defense,” they seem to overlook the fact that nearly all test takers study together.

Beyond the property of robustness, another important statistical concept is the idea of effect size. When a statistical assumption is violated, one should ask whether the violation would result in a large or small effect. Research suggests that studying together and sharing a common environment will only result in small, not large, effects (Allen, 2014).

Some ways in which assumptions may be violated are now listed. Because robustness of the statistics to violations and the effect sizes of violations are not known, it would be improper to place undue emphasis on them. Also, it is important to remember that these statistics are employed to detect potentially fraudulent test-taking behaviors.

1. The test responses were not independent.
  - a. **Instructional bias:** This can be especially problematic when students are taught to solve problems incorrectly or when the teaching is wrong.
  - b. **Collaborative learning:** There is a definite line between instruction and collusion, but at times that line blurs when students are encouraged to “help” each other.
  - c. **Repeated test taking by the same test taker:** Sometimes test takers are administered the same test again. Such an administration may generate excess similarity if the test taker remembers previous answers to the same question.
  - d. **Test fraud:** The following behaviors, which are generally acknowledged as some form of cheating, violate the assumption of independence:
    - i. Test takers obtain preknowledge of the test questions and/or answers (e.g., through the internet or some other media).
    - ii. Test takers copy from each other during the test.
    - iii. Test takers communicate with each other during the test.
    - iv. Test takers receive assistance or answers to questions from the same helper (e.g., an instructor or teacher).
    - v. Test takers use the same stand-in or proxy to take the test for them.
    - vi. A person responsible for administering the test (e.g., an educator or teacher) changes answers that were provided by the test takers (e.g., by erasing and marking the correct answer).



2. The selected mathematical model does not adequately estimate or approximate the match probabilities.
  - a. **Partial or incomplete tests:** Occasionally test takers omit responses. In general, shared omitted responses do not constitute evidence of potential test fraud. But, a large number of omitted responses can bias the probability estimates, leading to errors in computing the probability of the similarity statistic.
  - b. **Mislabeled data:** Errors in the data may result in apparent nonindependent test taking. For example, suppose the exam is given with two forms and two test takers are given Form A but their data is labeled as Form B. The extreme unusualness of the response patterns when scored under Form B contributes to extremely improbable alignments, which are perfectly reasonable under Form A.
  - c. **Item exposure or item drift:** When a subset of the items has become exposed to the extent that most test takers are aware of the item content, the items may be easier than previously estimated. This can cause errors in the probability computations.
  - d. **Negative item discriminations:** Items that have negative correlations with the total test score present a unique challenge for models that incorporate test-taker performance. Usually, only very small number of these items are present on exams, but a large number of these items can negatively affect probability computations.
  - e. **Nonmeasurement items:** Sometimes the testing instrument is designed and used to collect information besides the answers to the test questions (e.g., demographic questions or acceptance of a nondisclosure agreement). Unless removed, these items can create unwanted noise in the models and the probability computations.
  - f. **Differential item functioning:** If subgroups exist where the items perform in a distinctly different way than they do for the main population, improbable alignments may occur. If this type of problem is suspected, it may be appropriate to conduct the similarity analysis using subpopulation-specific models.
3. Item responses are not “locally” or conditionally independent.
  - a. **Constant answering strategies and patterned responses:** Two or more test takers use the same answering strategy, which results in them answering the questions in nearly the same way. For example, one of these strategies is “Guess ‘C’ if you don’t know the answer.”
  - b. **Multipart scenario items:** Responses for these items are generally provided sequentially, and responses may depend upon previous responses. This violation can inflate the value of the similarity statistic, yielding false positives.

When anomalous pairs of similar tests are found, it is important to determine the reason. The above discussion has listed potential explanations for extreme values of the similarity statistic. Analysis of live and simulated data indicates that inappropriate increases in test scores and pass rates also may be observed when the similarity is due to nonindependent test taking.

Although care has been taken to list how the assumptions might be violated, it should be remembered that similarity statistics are computed from pairs of data. As a result, in order for a pair to be detected by a similarity statistic, the assumptions must

be violated in the same way for both members of the pair. Because of this, similarity statistics are quite robust to violations of assumptions that do not involve fraudulent manipulation of the test responses.

### *Discussion of Exploratory and Confirmatory Analyses*

An exploratory analysis is conducted when the similarity statistic is computed for all pairs of test instances within the data set, without any preconceived hypotheses about which individuals may have been involved in cheating. Exploratory analyses are synonymous with data mining or data dredging. These are usually done to monitor the test administration for the existence of potential collusion. Because all the pairs have been computed and examined, unless a multiple-comparisons correction is used, a spuriously high number of detections will be observed. The Bonferroni correction or adjustment provides one way to account for conducting multiple related tests (i.e., that an examinee did not work independently on his or her exam).

Typically, the Bonferroni correction is used to establish a critical value for the test of hypothesis. The critical value is that value from the distribution that has a tail probability equal to the desired alpha level divided by the number of elements in the population that were examined (or the number of hypotheses that were statistically evaluated). When exploring all possible pairs, Wesolowsky (2000) recommended using  $N \times (N - 1)/2$  as the denominator for the Bonferroni adjustment, because that is how many comparisons were performed.

However, the simple adjustment to the critical value of the similarity statistic suggested by Wesolowsky is not appropriate when multiple data sets of varying sizes (e.g., schools or test sites) are analyzed. In fact, Wesolowsky's suggestion would result in critical values which vary from data set to data set. This is an unsatisfactory solution because all test takers would not be measured against the same standard or threshold.

It is difficult to maintain the same standard for evaluating the pairs in an exploratory analysis, because the recommendation is to perform an adjustment that depends upon  $N$ , which varies. If observations are added to a data set, which is then reanalyzed, the standard will change. If, as mentioned above, the number of comparisons is restricted using a proximity measure (e.g., the test takers were tested at the same test site or on the same day), then the standard will vary. Hence, a reasonable approach is needed to acknowledge this situation and at the same time apply a consistent rule for extracting pairs. This can and should be done by ensuring that the multiple comparisons adjustment is applied in a way that takes into account the number of pairwise comparisons and the number of test takers, even when the analysis is restricted to subgroups of varying sizes (e.g., different numbers of test takers were tested at the same test sites or on the same days).

The approach which has confirmed to be satisfactory through simulation is to factor the Bonferroni equation<sup>1</sup> using two terms. The procedure converts the similarity statistic into an "index value" using the relationship,  $p = 10^{-I}$ , where  $p$  is the upper tail probability value of the similarity statistic and  $I$  is the index value (Maynes, 2009). In the inverted formulation of the problem, the probability value associated with the Bonferroni critical value becomes  $p_{c,s} = p/(N \times (N_s - 1)/2) = \{10^{-I} \times 10^{-\log_{10}((N_s - 1)/2)}\} \times 10^{-\log_{10}(N)}$ . Instead of adjusting critical values, this formulation adjusts the similarity statistic (or its probability), thereby taking into account differences in data set sizes and allowing use of the same critical value for every single test taker in the population. The first factor takes into account the number of comparisons for a particular test



taker or student. It determines the index value for the maximum observed similarity statistic (as determined by the smallest tail probability), using the relation  $I_{max,s} = I + \log_{10}((N_s-1)/2)$ , where  $N_s$  is the number of comparisons for a particular student or test taker. At this point, all the index values,  $I_{max,s}$  are comparable from student to student. If desired, the second factor can be applied to take into account the number of test takers that were analyzed,  $I_s = I_{max,s} + \log_{10}(N)$ , where  $N$  is the number of students in the analysis.

A confirmatory analysis, on the other hand, is usually conducted on those response strings where independent evidence (e.g., a proctor's testing irregularity report) suggests that test security might have been violated. When performed this way, the Bonferroni adjustment should not be used, unless as suggested by Wollack, Cohen & Serlin (2001) for dealing with specific individuals who are being investigated multiple times, as might be the case if an individual is believed to have copied from multiple neighboring examinees.

There has been some debate among researchers and practitioners concerning the critical value that should be used for a confirmatory analysis. It appears that Angoff (1974) was the first author to report use of a specific critical value. He reported that ETS requires 3.72 standard deviations or greater before taking action, which corresponds with an upper tail probability value of 1 in 10,000 if the distribution of the statistic were normal. Angoff further states that the distribution of the similarity statistic will most likely be skewed and not normal. When asked for a recommended critical value to invalidate a test score, many practitioners will follow Angoff's critical value because (1) it is conservative, (2) similarity may be caused by nonfraudulent factors, and (3) current methods only approximate actual statistical distributions.

However, Angoff's critical value should not be taken as the de facto industry standard. On the contrary, selection of the critical value is a policy decision that is the sole responsibility of each testing organization. The decision should adhere to accepted scientific practice and abide by the organization's goals and responsibilities. Decisions concerning scores must take into account (1) the totality of evidence for individual situations, (2) the effect of Type I errors (i.e., false positives) on test takers balanced against the harm from Type II errors (i.e., failure to take action when nonqualified individuals receiving passing scores), and (3) the organization's ability to implement its policy decisions (e.g., if the number of flagged pairs is excessively burdensome, fewer pairs should be handled). Recommendation of a single critical value is somewhat simplistic because it cannot adequately address these factors.

### *Limitations of Similarity Statistics and Models*

Similarity statistics have the potential to detect and/or expose significant security risks to the exams. Even so, the statistics are subject to some limitations, in addition to situations that may violate the assumptions. These are:

1. A sufficiently large sample size must be used to estimate the match probabilities adequately. The question of adequateness may be answered by goodness-of-fit analyses. The question "How large of a sample is needed?" depends upon model assumptions. For example, a three-parameter logistic model requires a larger sample for estimation than a Rasch model.
2. Similarity statistics cannot determine responsibility and/or directionality. For example, some answer-copying statistics are computed under the hypothesis that

the “source” and the “copier” are known. A similarity statistic is unable to make an attribution of a source or a copier. Additional information is needed to make such an inference.

3. Similarity statistics cannot detect cheating that occurs through some other means than nonindependent test taking. For example, if a surrogate or proxy test taker is employed by only one test taker, the similarity statistic will not be able to detect this.
4. As discussed above, similarity statistics, when used to compare each test with every other test (i.e., data mining), must utilize an approach to adjust the index (or critical) value to reflect multiple comparisons and keep the Type I error rate controlled. However, in doing so, it is important to recognize that the power of the statistical procedure will be lowered.
5. Similarity statistics are sensitive to the number of items in common between two test instances. If two examinees share very few items (e.g., as with a CAT test), the power of similarity statistics to detect collusion between those individuals will generally be low.
6. The distribution of similarity statistics depends upon test-taker performance. Power decreases as the test scores increase. Thus, similarity statistics cannot detect cheating between tests when performance levels are very high (e.g., nearly every question is answered correctly).

### DIFFICULTIES IN MODELING THE DISTRIBUTIONS OF SIMILARITY STATISTICS

When Angoff (1974) published his analysis, he dismissed the notion that the distribution of the similarity statistic could be modeled theoretically. He stated, “However, even a brief consideration of this [problem] makes it clear that the complexities in making theoretical estimates of such a distribution are far too great to make it practical” (p. 44). His primary objection was that there was no known way to model the correlations between correct and incorrect responses.

Using assumptions of statistical independence conditioned upon test-taker performance, Item Response Theory (Lord, 1980) is able to model the correlations between the item responses for each pair of test takers. The NRM (Bock, 1972) allows correlations to be modeled for each distinct response. Thus, modern IRT addresses Angoff’s concerns and provides the framework for computing the statistical distribution of the similarity statistics.

*It should be emphasized that these distributions cannot be modeled without conditioning upon examinee performance (i.e.,  $\theta$ ) and without recognizing that items (and item alternatives) vary in difficulty.* For example, the Error Similarity Analysis by Bellezza and Bellezza (1989) uses the observed number of identical incorrect responses between two examinees. This analysis conditions upon the total number of errors in common between any two students, but it does not recognize that items vary in difficulty. The other difficulty with the analysis is that the observed number of identical incorrect responses depends upon the observed number of identical correct responses, which is a random variable and is not modeled in the analysis. Thus, this example supports Angoff’s position that the distribution of similarity statistics is not easily modeled theoretically. Building upon partially successful attempts by earlier researchers, distributions of more recently published similarity statistics (i.e., GBT, S-Check, and M4) appear to be well approximated using generalized binomial and trinomial distributions.

As discussed in the preceding paragraphs, some researchers have developed similarity statistics under the assumption that items are equally difficult or that wrong answers are equally attractive. For example, this assumption was made explicitly by Bellezza and Bellezza (1989) and implicitly by Angoff (1974). In fact, the author has attempted to develop computationally efficient algorithms for similarity statistics using the assumption that items are equally difficult (Maynes, 2013). He reported that the assumption results in overdetecting pairs of similar test responses (i.e., Type I error is inflated) because the expected value of the number of matching responses is underestimated (see Maynes, 2013 for mathematical details). The assumption that all items are equally difficult or that all wrong answers are equally attractive, while appealing and appearing to be trivial, is not supportable in practice. Because of this, the test response data should *not* be pooled across items to estimate matching probabilities. Doing so will result in approximating distributions that are biased toward the lower tail, which will spuriously raise false positive rates (i.e., the tail probabilities will be smaller than the actual probabilities).

### *Region of Permissible Values for Similarity Statistics*

Even though the similarity statistics described in this chapter use examinee performance and item difficulty to estimate the statistical distributions, they ignore the fact that the distributions are bounded (especially when the normal approximation is used, as recommended by Wesolowsky). Maynes (2013) used the term “region of permissible values” to describe how the distributions of similarity statistics are bounded.

After conditioning upon test-taker performance (i.e., the number of correct answers), the distribution of the number of identical correct and incorrect responses is confined to a region of permissible values (Maynes, 2014). Values outside of this region are impossible. For example, once it is known that both test takers answered every question correctly, even one identical incorrect response would be impossible. Holland (1996) discusses the relationship between the number of matching (identical) incorrect answers and the number of answers where the two test takers disagreed. He emphasized that once the raw test scores are known (i.e., the total number of items answered correctly by each test taker), the *marginal totals* are fixed and constrain the possibilities for the two-by-two table of agreement between items answered correctly and incorrectly. An example from a 60-item test is shown in Table 3.1.

The marginal totals in Table 3.1 are provided in bold font because they are fixed. Given the data in Table 3.1, the greatest number of questions on which both Test Takers  $T_1$  and  $T_2$  could answer correctly is 42, and the lowest number is 27. In other words, the test takers **MUST** answer at least 27 of the same questions correctly and **CAN-NOT** answer more than 42 of the same questions correctly. Conversely, if 42 questions are answered correctly by both test takers, it **MUST** be the case that they answered the same 12 questions incorrectly and they disagreed upon the answers for the three remaining questions.

**Table 3.1** Example Agreement Between Test Takers  $T_1$  and  $T_2$  with Scores of 42 and 45

$T_1 / T_2$	Correct	Incorrect	Total
Correct	42 to 27	0 to 15	<b>42</b>
Incorrect	3 to 18	15 to 0	<b>18</b>
Total	<b>45</b>	<b>15</b>	<b>60</b>

**Table 3.2** Agreement Between Test Takers  $T_1$  and  $T_2$  With Scores of  $Y_1$  and  $Y_2$

$T_1 / T_2$	Correct	Incorrect	Total
Correct	$N_{11} = R$	$N_{12} = Y_1 - R$	$Y_1$
Incorrect	$N_{21} = Y_2 - R$	$N_{22} = N + R - (Y_1 + Y_2)$	$N - Y_1$
Total	$Y_2$	$N - Y_2$	$N$

Given the total number of questions ( $N$ ) and the number of correct answers for the two test takers ( $Y_1$  and  $Y_2$ ), all of the cell counts in the two-by-two table of agreement will be determined when one other count has been established. For convenience sake, it is suitable to use the number of correctly answered questions ( $R$ ) shared by the two test takers for this quantity. These quantities are shown in Table 3.2.

In Table 3.2,  $R$  (the number of identical correct answers) cannot exceed the minimum value of  $Y_1$  and  $Y_2$  and not less than the maximum value of 0 and  $(Y_1 + Y_2 - N)$ . Likewise, the value of  $N_{22}$  (the maximum number of identical incorrect responses) must be between the values of 0 and  $N - \max(Y_1, Y_2)$ . If there is only one correct answer for each item,  $R$  is the number of identical correct answers. If there is only one incorrect answer for each item (i.e., True/False question),  $N + R - (Y_1 + Y_2)$  is the number of identical incorrect answers.

In summary, the following relationships define the region of permissible values:

1. If each question has only one correct answer and one incorrect answers (e.g., True/False), the region of permissible values is defined on an interval between  $\max(0, N - Y_1 - Y_2)$  and  $N + \min(Y_1, Y_2) - \max(Y_1, Y_2)$ .
2. If each question has only one correct answer and multiple incorrect answers (e.g., typical multiple-choice question), the region of permissible values is defined by a triangular area with the number of identical correct answers,  $R$ , lying between  $\max(0, Y_1 + Y_2 - N)$  and  $\min(Y_1, Y_2)$ , and with the number of identical incorrect answers lying between 0 and the value  $R - \max(0, Y_1 + Y_2 - N)$ . The total number of identical answers is defined on the interval between  $\max(0, Y_1 + Y_2 - N)$  and  $N + \min(Y_1, Y_2) - \max(Y_1, Y_2)$ .
3. If each question has multiple correct answers and multiple incorrect answers (e.g., a math problem where several answer variants are correct), the region of permissible values is defined by a rectangular area with the number of identical correct answers,  $R$ , lying between 0 and  $\min(Y_1, Y_2)$ , and with the number of identical incorrect answers lying between 0 and the value  $\min(Y_1, Y_2) - \max(0, Y_1 + Y_2 - N)$ . The total number of identical answers is defined on the interval between 0 and  $2[\min(Y_1, Y_2)] - \max(0, Y_1 + Y_2 - N)$ .

While the above exercise in establishing the region of permissible values for the number of identical correct responses,  $R$ , and the number of identical incorrect responses,  $W$ , may seem trivial, it is of critical importance. None of the distributions of similarity statistics published to date have accounted for this restriction. Instead, mathematical formulae have been used to compute probabilities for tail values that are impossible to observe. Thus, the computations published for the distributions of GBT, S-Check, and M4 only approximate actual probabilities.

To understand the approximation errors that result from ignoring the region of permissible values, it is useful to overlay the region of permissible values onto the bivariate

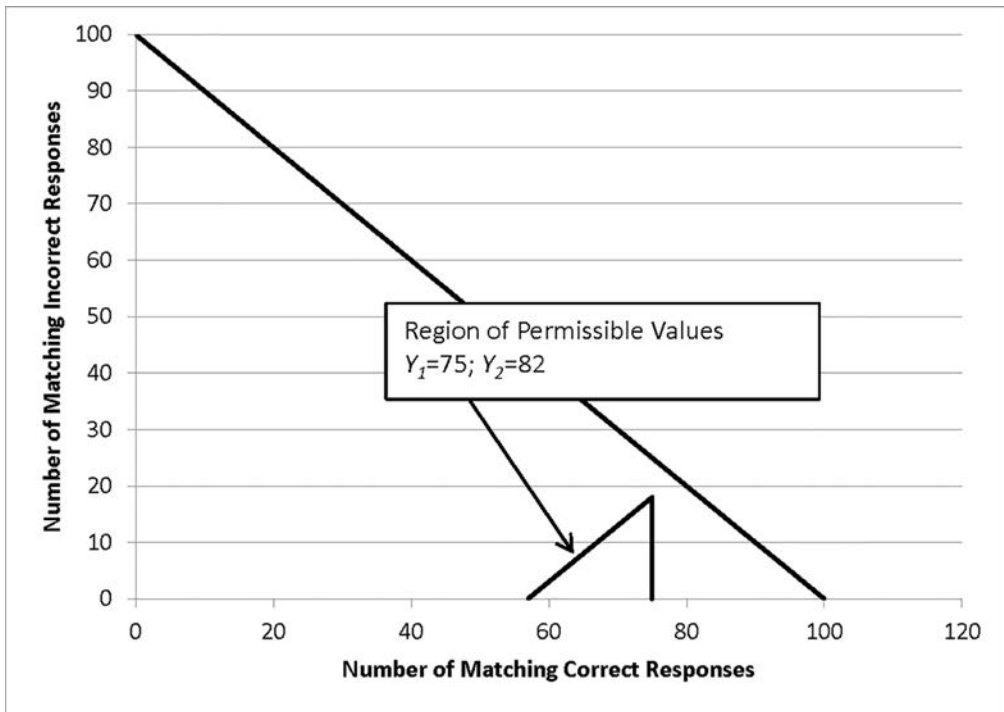


Figure 3.1 Illustration of region of permissible values; 100 items

sample space of matching correct and matching incorrect answers. This has been done in Figure 3.1.

In Figure 3.1, the upper line that begins at the point (0, 100) and ends at (100, 0) represents the boundary of the number matches that could possibly be observed. It defines the probability space that is modeled by the GBT and M4 statistics. Probabilities are explicitly or implicitly computed for every single one of the points within this triangle. The region of permissible values for two test takers with scores of 75 and 82 is shown in the smaller triangle with vertices at (57, 0), (75, 0), and (75, 18). Each of the similarity statistics computes and adds up probability values that are impossible to observe, being outside the region of permissible values. The question remains whether this neglect is important. For the generalized binomial statistic, GBT, probabilities will be assigned to impossible lower tail values. Probabilities will also be assigned to impossible upper tail values. In general, computing probabilities for impossible upper tail values will result in computing tail probabilities that are too large. Thus, the statistical procedures will tend to overstate Type I errors. The actual Type I error rate will be lower than the reported probability. This results in loss of statistical power. This effect is more pronounced as the difference between  $Y_1$  and  $Y_2$  increases. A similar analysis applies to M4, with the exception that the probabilities are summed using a curved contour, which results in closer approximations to actual probabilities. Computed tail probabilities will still be larger than the actual probabilities (this result is seen in the simulation of M4 Type I errors in this chapter).

The S-Check statistic uses a normal approximation which has an infinite tail. Thus, tail probabilities computed by this statistic are likely to be larger than the same probabilities computed by the GBT and M4 statistics, meaning this statistic will lose more power than the other two statistics.

### THE M4 SIMILARITY STATISTIC

This section discusses the M4 similarity statistic because it appears to provide the best approximation for computing the probability of observed similarity (i.e., it is least impacted by noncompliance with the region of permissible values) and because it is more appealing to use two separate pieces of evidence (i.e., the number of identical correct and identical incorrect responses) that have varying power than to combine them and possibly dilute the strength of the evidence.

The M4 similarity statistic is a bivariate statistic. It consists of the number of identical correct responses and the number of identical incorrect responses. It has been generally acknowledged that incorrect matching answers provide more evidence of nonindependent test taking than correct matching answers, because these are low-probability events. Thus, a bivariate statistic should be able to take advantage of this fact statistically.

The M4 similarity statistic is more appealing than the GBT and S-Check statistics because it allows the incorrect matching answers and the correct matching answers to be evaluated jointly, but separately.

#### *Probability Density Function and Tail Probabilities for the M4 Similarity Statistic*

The probability density function of the M4 similarity statistic may be approximated by assuming statistical and local independence of matching responses. Using these assumptions, the statistic follows a generalized trinomial distribution. This distribution is a special case of the generalized multinomial distribution. It does not have a closed form, but its generating function (Feller, 1968) is

$$G(x, y) = \prod_i (r_i + p_i x + q_i y), \tag{Equation 1}$$

where  $p_i$  is the probability of a matching correct response for item  $i$ ,  $q_i$  is the probability of a matching incorrect response for item  $i$ ,  $r_i$  is the probability of a nonmatching response for item  $i$ ,  $x$  is the count of observed matching correct responses, and  $y$  is the count of observed matching incorrect responses. The reader should note that  $p_i + q_i + r_i = 1$ .

Using the generating function of the generalized trinomial distribution (Equation 1), the joint probability distribution of  $x$  and  $y$  may be computed using a recurrence relation (Tucker, 1980, see also Graham, Knuth, & Patashnik, 1994). For a pair of responses, there are three possibilities: a matching correct, a matching incorrect, and a nonmatching response. Let these possibilities be summarized using three triplets:  $(1,0,0)$ ,  $(0,1,0)$ , and  $(0,0,1)$ , where the first value of the triplet is a binary value indicating whether there was a matching correct answer, the second value indicates whether there was a matching incorrect answer, and the third value indicates whether there was a nonmatching number. Because the third value is equal to one minus the sum of the other two values, it can be discarded. Hence, the trinomial distribution can be formed using bivariate pairs of  $(1,0)$ ,  $(0,1)$ , and  $(0,0)$ .

Using the above notation, the trinomial distribution for M4 can now be expressed mathematically. The joint probability distribution for M4 is written as

$$\begin{aligned} T_{k+1}(x, y) &= p_{k+1}(1,0)T_k(x-1, y) \\ &\quad + q_{k+1}(0,1)T_k(x, y-1) \\ &\quad + (1 - p_{k+1}(1,0) - q_{k+1}(0,1))T_k(x, y), \text{ with boundary condition} \\ T_0(0,0) &= 1 \text{ and } T_0(x, y) = 0 \forall (x, y) \neq (0,0), \end{aligned} \tag{Equation 2}$$



where the values of  $T(x,y)$  are computed successively for the matches of each item response pair represented by the subscript  $k+1$ ,  $x$  is the number of matching correct responses, and  $y$  is the number of matching incorrect responses. The value of  $k+1$  begins with 1 and ends with  $n$  (the number of items answered by both test takers). In Equation 2, the value of  $r_{k+1}$  is implicitly computed by subtraction (i.e.,  $1-p_{k+1}(1,0)-q_{k+1}(0,1)$ ). The bivariate pairs of (1,0), and (0,1) are explicitly shown in Equation 2 to emphasize that the probabilities are associated with matching correct, matching incorrect and nonmatching responses. When the values of  $p_i$  and  $q_i$  are constant, the generalized trinomial distribution becomes the trinomial distribution.

Recurrence equations (such as Equation 2) can be a bit difficult to understand in symbolic notation. An illustration shows how the recurrence is computed.

Table 3.3 illustrates four subtables. The first subtable corresponds to the initial condition, when with probability one there are no matching correct and no matching incorrect responses. The succeeding subtables correspond to the joint probability distribution that results after adding an item to the recurrence. For the first item, the probability of a correct match is 0.30 and the probability of an incorrect match is 0.20, likewise the two corresponding probabilities for the second item are 0.35 and 0.15, and for the third item they are 0.45 and 0.25. The rows in each subtable correspond to the number of correct matches. The columns correspond to the number of incorrect matches. The subtables are triangular shaped because the number of correct identical matches added to the number of incorrect identical matches cannot exceed the number of items.

Equation 2 specifies how to compute the values for each cell in the table. For example, the value of the cell for  $T_3(1,1)$  is computed by substituting  $p_3$  and  $q_3$  into the formula to obtain  $0.45 \times T_2(0,1) + 0.25 \times T_2(1,0) + 0.30 \times T_2(1,1)$ .

**Table 3.3** Recurrence Pattern for the Generalized Trinomial Distribution

$T_0: 0/0$ (initial)	0
	0 1

$T_1: 0.3/0.2$ (first item)	0	1
	0 0.5	0.2
	1 0.3	

$T_2: 0.35/0.15$ (second item)	0	1	2
	0 0.25	0.175	0.03
	1 0.325	0.115	
	2 0.105		

$T_3: 0.45/0.25$ (third item)	0	1	2	3
	0 0.075	0.115	0.05275	0.0075
	1 0.21	0.1945	0.04225	
	2 0.17775	0.078		
	3 0.04725			

It is important to realize that the trinomial distribution has three tails, as illustrated in Table 3.3 and Figure 3.1. As a result, standard probability computations that are used for univariate distributions are not applicable. There is no obvious direction for computing the tail probability because there is no upper tail and no lower tail. Because there is not an obvious upper tail, the desired tail probability must be computed by using a subordering principle as recommended by Barnett & Lewis (1994).

The subordering principle is implemented using a two-step procedure. First, an “upper” probability is computed for each point  $(x,y)$  in the M4 distribution. This is done by adding the probabilities for all bivariate points  $(t,v)$  where  $t$  is greater than or equal to  $x$  or  $v$  is greater than or equal to  $y$ . This quantity is named  $D_{x,y}$ . Second, the subordering principle stipulates that the desired probability associated with the point  $(x,y)$  is the sum of all values of  $T(j,k)$  where  $D_{j,k} \geq D_{x,y}$ . The subordering principle for the probability computation is illustrated in Figure 3.2.

In Figure 3.2, the diamond provides the location of the expected value of the distribution and the square provides the observed value for a pair of extremely similar test instances. Step one denotes addition of the bivariate probabilities from the observed data to point to the triangular boundary in the direction of upward and to the right. If the probability computation stops with step one, *the tail probability will be an underestimate of the actual probability* (i.e., the probability of rejection will be inflated) because the subordering principle has not yet been applied completely. The values from step one are used in defining the subordering principle. The data points in the bivariate distribution are ordered using the values computed in step one. The bivariate probabilities associated with the ordered points are summed to compute the tail probability. This is represented as step two in Figure 3.2.

Thus, the probability of the observed value is computed by using a directional ordering relation that defines the direction of extremeness as being upwards (a greater number of identical incorrect answers) and to the right (a greater number of identical correct answers) of the observation. After having computed the probability using the directional ordering relation, the definition of extremeness is a direct interpretation of

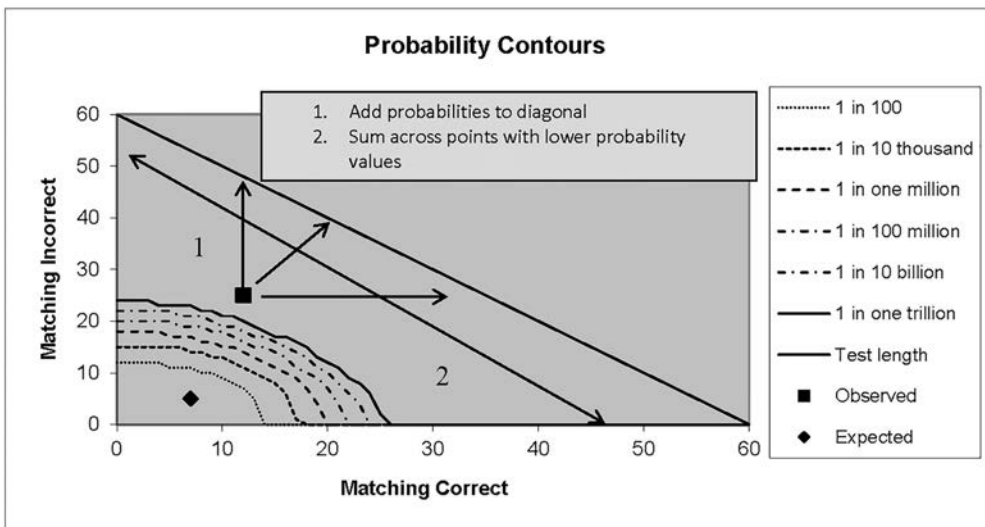


Figure 3.2 Illustration of the M4 Probability Computation

the probability. Observations with smaller probabilities are more extreme than other observations.

### *Analysis of Licensure Data Set*

The M4 statistic can be used both to detect potential cases of nonindependent test taking and to confirm or reject propositions that particular individuals may have committed test fraud. Detection of potential test fraud involves data mining and examining all pairs of interest in the data set. Because a large number of pairs will be evaluated, the likelihood of committing a Type I error (i.e., incorrectly stating that a test instance was not taken independently) will be high unless an appropriate error control is used. The multiple comparisons used in the analysis of the licensure data set was based on the maximum order statistic (Maynes, 2009). This correction is written as

$$I_{max,s} = -\log_{10}(1-(1-p_{min})^{(N_s-1)/2}), \quad (\text{Equation 3})$$

where  $I_{max,s}$  is the index for the examinee computed using  $N_s$ , the number of test takers that were compared, and  $p_{min}$ , the smallest observed probability value. Additionally, because M4 is computationally intense, the licensure data set was split in two subsets, each subset corresponding to one of the forms. As was done with the simulation, the responses to the pretest items were discarded. There were a total of 2,687,976 pairwise comparisons performed.

The detection threshold was set using the procedure described in the “Discussion of Exploratory and Confirmatory Analyses” section. That procedure stated that the probability value should be adjusted using  $(N_s-1)/2$  ( $N_s$  was set at 1,636 and 1,644 for Form 1 and Form 2, respectively). This was done using Equation 3. It further stated that after comparing the test response for each test taker, the probability value should then be adjusted using  $N$  (3,280). The second adjustment means that the detection threshold was set at an index value of 4.82, after adjusting for the number of test takers per form. Detection using this threshold maintains a low false positive rate of 0.05 while providing a good possibility of detecting true positives. The Bonferroni calculation for the second adjustment was  $p = 0.05/2687976$ , which is approximately 1 in 53 million. The simulation results in Table 3.5 in this chapter indicate that a significant number of nonindependently answered items (e.g., perhaps 30% to 50%) would be necessary in order for a pair to be detected using this procedure. Table 3.4 lists the pairs that were detected in the analysis of the licensure data.

In Table 3.4, the Flagged column indicates whether the examinee is a known cheater. The next three columns indicate the Examinee’s country, state in which the examinee applied for licensure, and the school or institution where the examinee was trained. The Test Site indicates where the test was administered. The Pass/Fail Outcome indicates whether the examinee had a passing score on the exam. And the M4 Similarity Index is the observed probability for the pair adjusted using Equation 3.

From among 2,687,976 pairs that were compared, M4 detected seven pairs. If this procedure were performed on this many tests which were taken independently, on average only one pair would be detected every 20 analyses. In other words, the expected number of detected pairs using the Bonferroni adjustment was 0.05.

The reader will notice in Table 3.4 that some individuals were detected multiple times. Detections that involve more than two individuals are known as clusters and are discussed by Wollack and Maynes (this volume).

**Table 3.4** Detections of Potentially Nonindependent Pairs in the Licensure Data

Examinee ID	Flagged	Country	State	School	Exam Form	Test Site	Site State	Pass/Fail Outcome	Raw Score	M4
										Similarity Index
e100624	1	India	28	5530	Form1	2305	28	1	140	9.5
e100505	1	India	28	8152	Form1	2305	28	1	132	9.5
e100505	1	India	28	8152	Form1	2305	28	1	132	7.7
e100498	1	India	42	8119	Form1	5880	42	1	136	7.7
e100505	1	India	28	8152	Form1	2305	28	1	132	5.9
e100452	1	India	28	8172	Form1	2305	28	1	128	5.9
e100624	1	India	28	5530	Form1	2305	28	1	140	5.5
e100452	1	India	28	8172	Form1	2305	28	1	128	5.5
e100226	1	India	42	8155	Form1	5303	54	0	112	5.2
e100191	1	India	42	5530	Form1	2305	28	0	107	5.2
e100505	1	India	28	8152	Form1	2305	28	1	132	5.0
e100494	1	India	42	8198	Form1	5302	54	1	131	5.0
e200294	0	India	42	8198	Form2	2305	28	0	113	6.5
e200448	1	India	28	8198	Form2	2305	28	1	126	6.5

*Presentation of Results from M4 Similarity Analysis*

To plot and evaluate M4, two statistically dependent values are computed when comparing two test instances: (1) the number of identical correct responses and (2) the number of identical incorrect responses. These values may be plotted using a pie chart with three slices. Examples of these data are shown in Figure 3.3, using the first pair listed in Table 3.4.

The panel on the left of Figure 3.3 provides the counts of observed similarities between the two tests, and the panel on the right of Figure 3.3 provides the expected agreement on the tests. There were 170 scored questions analyzed in this comparison.

As a means of comparison, two other test takers, e100553 and e100654, with exactly the same scores as the pair illustrated in Figure 3.3, were selected from the licensure data. For this new pair, the M4 similarity index was equal to 0.3 (the median value of the distribution). These data are shown in Figure 3.4.

The format and labels of the data in Figure 3.4 are the same as those in Figure 3.3. The observed numbers and expected numbers of agreement in Figure 3.4 are nearly equal, which is expected.

The counts of observed and expected agreement may be more readily visualized when the selected responses of the common items are aligned. For the sake of simplicity in the presentation and exam security, the actual item responses are not shown. Instead, each item response is shown as being the same correct response, the same incorrect response, or differing responses between the two exams. The data for Figure 3.3 and Figure 3.4 are shown in Figure 3.5.

There are two panels in Figure 3.5. The upper panel depicts the alignment data for the 170 items for the two test takers whose data were shown in Figure 3.3 (i.e., the extremely similar pair—index value of 12.5). The lower panel depicts the alignment data for the same items for the two test takers whose data were shown in Figure 3.4

Extreme Similarity – Index = 12.5

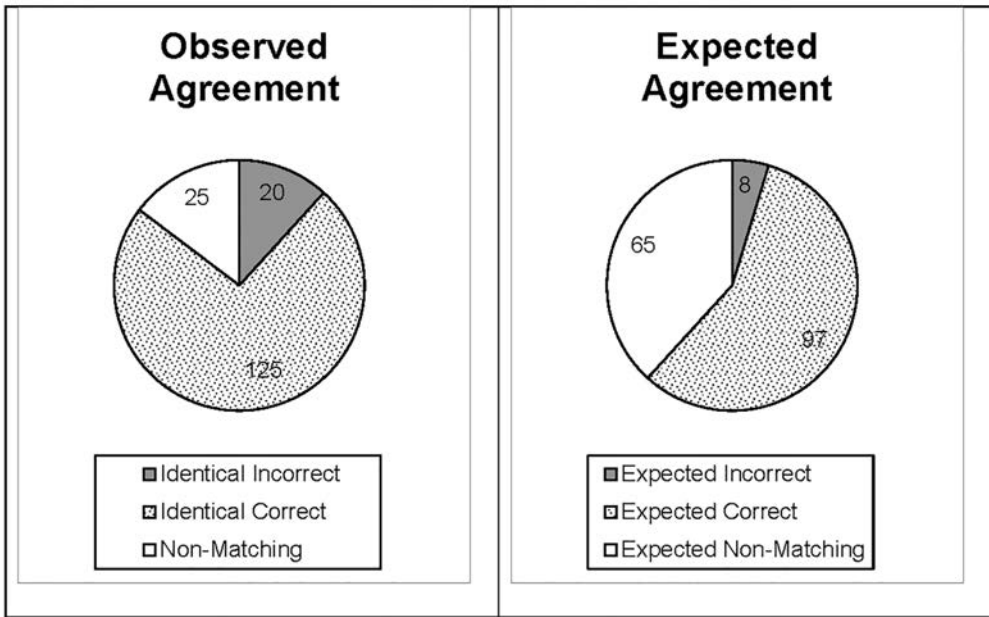


Figure 3.3 Illustration of the M4 Similarity Statistic: Extreme Similarity Index = 12.5

Median Similarity – Index = 0.3

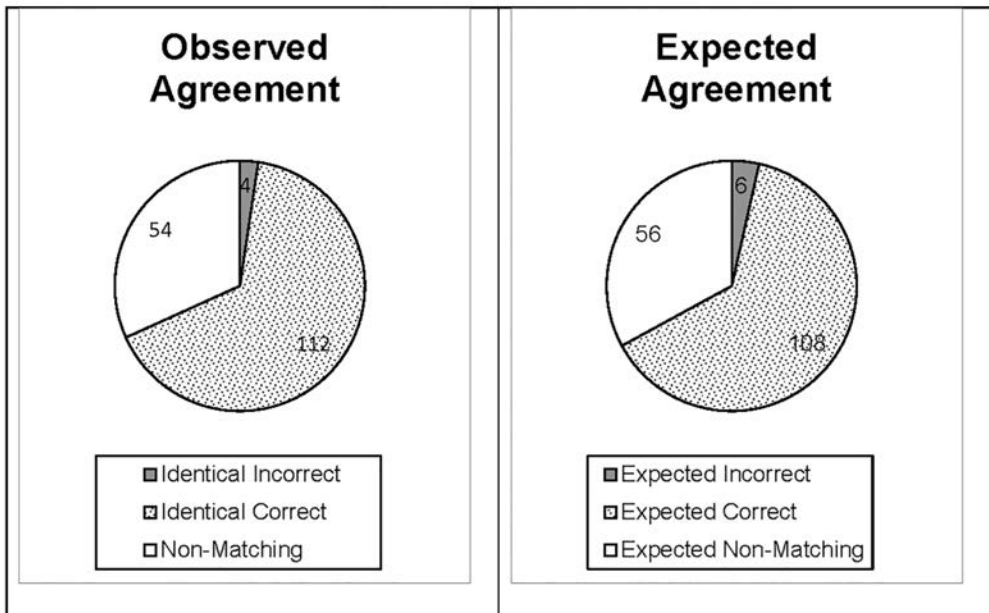
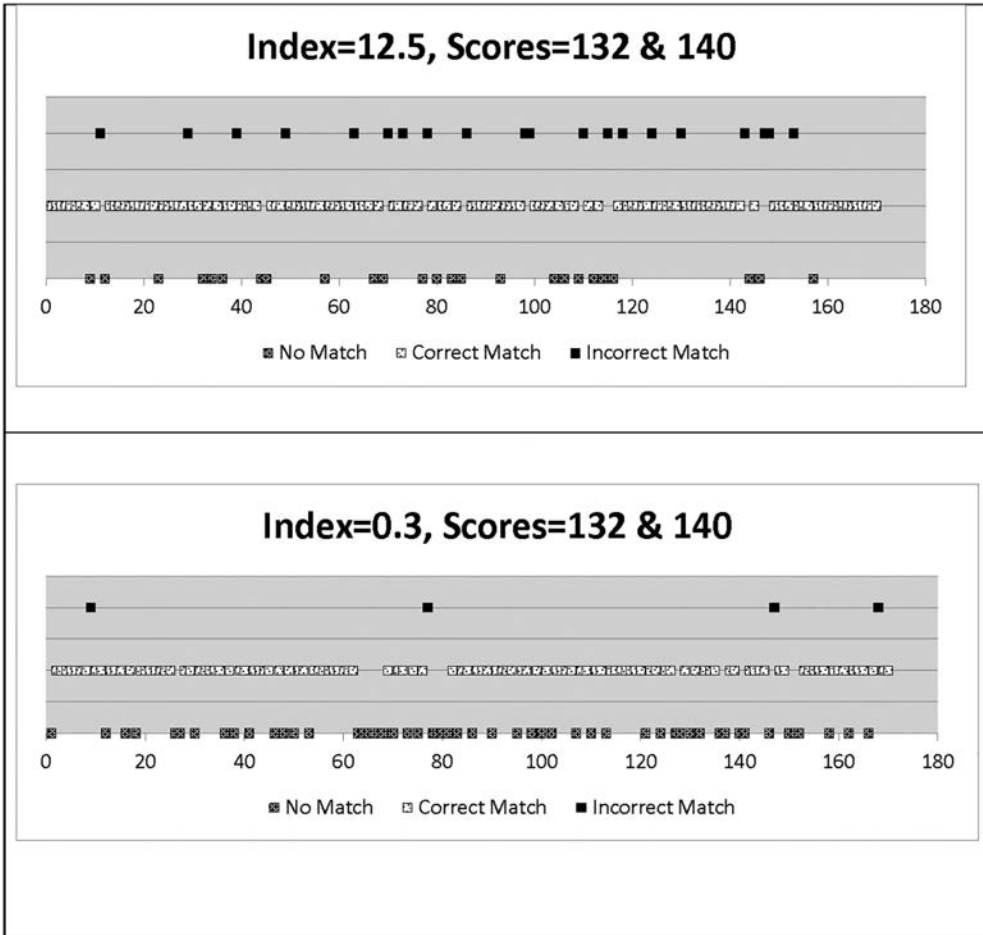


Figure 3.4 Illustration of the M4 Similarity Statistic: Median Similarity Index = 0.3



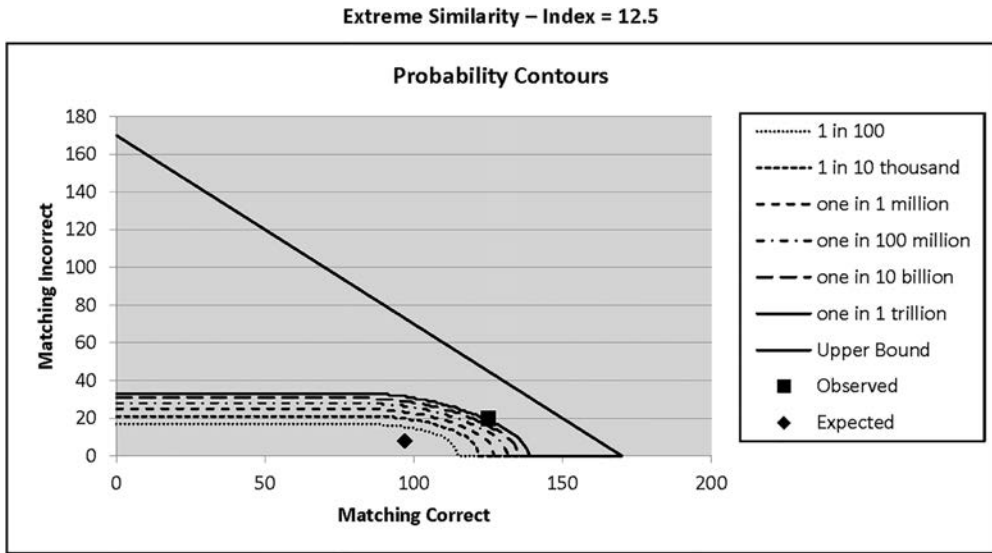
**Figure 3.5** Illustration of Aligned Responses

(i.e., the typical similarity pair with an index value of 0.3). As was shown in Figures 3.2 and 3.3, the numbers of agreed upon incorrect and correct responses for the extremely similar pair were much higher than for the typical similarity pair.

The extremity of an observed value of  $M_4$  may be shown using a contour plot of the bivariate probability distribution for the number of identical correct and incorrect responses. The contour plot for the extremely similar pair (i.e., the test takers from Figure 3.1 with an index of 12.5) is shown as Figure 3.6.

Figure 3.6 illustrates the region of the probability distribution for the bivariate distribution of the matching correct and matching incorrect response counts. The observed data from Figure 3.3 (i.e., 20 matching incorrect and 125 matching correct responses) have been plotted using a large square, and the expected values of the distribution (i.e., 8 matching incorrect and 97 matching correct responses) have been plotted using a diamond. The curved lines (labeled 1 in 100, 1 in 10 thousand, etc.) between the expected and observed values represent diminishing probability contours. Each contour line is drawn at a power of 100. The contour with the lowest probability represents a probability level of 1 chance in  $10^{12}$ , or one in a trillion.





**Figure 3.6** Contour Plot of the M4 Similarity Statistic: Extreme Similarity Index = 12.5

The direction of probability computation is chosen to maximize evidence for the alternative hypothesis of nonindependent test taking when the data are extreme and due to collusion. For example, the point (0, 0) is very far from the expected value of the distribution, but it does not provide evidence that the tests are similar. The upper bound is the limit of the distribution where the number of identical correct and identical incorrect answers equals the number of items (meaning that the number of nonmatching answers is zero, because the total number of items is the sum of identical correct, identical incorrect, and nonmatching responses).

#### *Analysis of Type I and Type II Errors for M4 Similarity*

A simulation was conducted to evaluate Type I and Type II error rates for M4. The simulation was performed using the licensure data set provided to chapter authors in this book. The licensure data set was prepared by removing the responses for the pre-test items, leaving 170 scored items on each form. Even though the M4 statistic can be used to create clusters of nonindependently taken exams, in this section the simulation was conducted using pairs. Parameters for the NRM were estimated using the provided licensure data. These parameters were then used to simulate item responses.

Type I errors result from incorrectly rejecting the null hypotheses. Type II errors result from failure to reject the null hypotheses when the null hypotheses are false. Type I errors will potentially produce inappropriate score invalidations. Type II errors will result in acceptance of inappropriate test scores. Thus, the practitioner must balance the costs of Type II errors against the costs of Type I errors and the benefits of correctly detecting nonindependent test taking against the benefits of correctly determining that the test was taken appropriately. The primary mode by which these costs and benefits may be balanced is through selection of the flagging threshold for the similarity statistic. Conservative thresholds reduce Type I errors at the expense of increased Type II errors. Statistical decision theory can provide guidance for selecting the threshold, but ultimately determination of the threshold is a policy decision.

### *Simulation Methodology*

Twelve data sets were generated using simulation under different levels of dependence, varying from 0% copying to 55% copying in increments of 5%. Each of the 12 data sets was generated in the following manner:

1. Each test response vector in the live data set was selected and used as a “reference” vector 31 times, thereby allowing (a) source data to be actual test data and (b) a total of over 100,000 pairs of examinees to be simulated (i.e.,  $3,280 \times 31$  equals 101,680).
2. The value of  $\theta$  was computed for the reference vector. Two values of  $\theta$ ,  $\theta_1$  and  $\theta_2$  were sampled from a normal distribution using  $\theta$  as the mean and a variance of one.
3. Using a prespecified proportion of copying (i.e., varying from 0% to 55%), items were randomly sampled without replacement from the reference vector and selected for copying.
4. Two item response vectors (IRV),  $\nu_1$  and  $\nu_2$ , were then generated using  $\theta_1$  and  $\theta_2$ , respectively. Item responses were copied from the reference vector of those items that were selected for copying and were generated using  $\theta_1$  and  $\theta_2$  from the NRM for the remaining items. Even though a source-copier model was not simulated, this is functionally equivalent to simulating the entire IRV for two copiers and then changing responses to match the source (i.e., the reference vector) for the appropriate proportion of copied items.

The above approach allows for simulating the standard source-copier setup. It also allows for simulating general collusion where the “copied” items become “disclosed” or “shared” items. It also may be extended to simulate other collusion scenarios. However, in the present case, only simulated pairs of IRV’s are needed.

### *Simulation Results—Null Condition*

The null condition corresponds to the 0% copying level. This is the condition when tests are taken independently and is the customary null hypothesis for the M4 statistic. When M4 is used for detection and/or confirmation of tests taken nonindependently, the upper tail value is evaluated. If the NRM is appropriate for the test response data, the index value of M4 follows an exponential distribution. The index value is mathematically equivalent to the upper tail probability ( $p = 10^{-\text{index}}$ ). Thus, an index value of 1.301 corresponds with an upper tail probability of 0.05. Because each pair was examined only once, the Bonferroni correction was not used in the simulations. Table 3.5 summarizes the tail probabilities estimated from the simulated condition of 0% copying.

**Table 3.5** Type I Error Rates When Tests Are Taken Independently

Index	Theoretical Rate	Number of Detections	Observed Rate
1.301	0.05	3,155	0.031
2.0	0.01	520	0.0051
2.301	0.005	268	0.0026
3.0	0.001	44	0.00043
3.301	0.0005	17	0.00017
4.0	0.0001	1	0.00001

**Table 3.6** Type II Error Rates for Copying Levels From 5% to 55%

Copying Condition	Detection Thresholds Specified as the M4 Similarity Index Value					
	1.301	2.0	2.301	3.0	3.301	4.0
5%	0.87891	0.96790	0.98213	0.99547	0.99752	0.99931
10%	0.65582	0.86510	0.91226	0.96929	0.98140	0.99445
15%	0.38748	0.65773	0.74347	0.87983	0.91487	0.96408
20%	0.15572	0.37209	0.46875	0.66862	0.73634	0.85555
25%	0.05381	0.16902	0.23685	0.41233	0.49002	0.65446
30%	0.01296	0.05371	0.08386	0.18462	0.23885	0.37953
35%	0.00306	0.01681	0.02887	0.07522	0.10291	0.19037
40%	0.00077	0.00400	0.00747	0.02237	0.03319	0.07023
45%	0.00013	0.00095	0.00199	0.00690	0.01061	0.02540
50%	0.00004	0.00022	0.00033	0.00160	0.00255	0.00662
55%	0.00000	0.00001	0.00010	0.00044	0.00068	0.00218

Table 3.5 shows that the simulation probabilities for the M4 Similarity statistic are close to, but less than, the theoretical probabilities, showing that M4 has good control of the nominal Type I error rate. As mentioned in the discussion of the region of permissible values, the theoretical rates for M4 are greater than observed rates. The difference is due to the way in which M4 approximates the actual probability value.

### *Simulation Results—Copying Conditions*

Table 3.6 provides the false negative or Type II error rates for selected index values of M4.

As expected, the Type II error rate for M4 decreases as the copying rate increases or, conversely, statistical power<sup>2</sup> increases as the copying rate increases. At copying levels of 50% or greater, the statistic has very high power. Additionally, the Type II error rate increases as the detection threshold increases.

## SUMMARY

This chapter has provided an exposition for a method of detecting nonindependent test taking by comparing pairs of test instances using similarity statistics. A simulation was performed that showed (1) the Type I error rate is not inflated using the M4 similarity statistic, and (2) a significant number of identical responses are needed before a pair of test takers will be detected as potentially not having taken their tests independently. Future papers are needed to describe and explore clustering approaches, computing cluster-based probability evidence for nonindependence, or evaluating performance changes that may have resulted through the behavior that produced the cluster of similar test instances. The problem of cluster extraction is very important, and Wollack and Maynes (this volume) explore the accuracy of extraction for a simple clustering algorithm.

## NOTES

1. The approximation of the Bonferroni adjustment is best when  $p$  is small. When  $p$  is large, the distribution of the maximum order statistic provides a better approximation (Maynes, 2009).
2. Statistical power is the ability to detect true positives. It is equal to 1 minus the Type II error rate.

## REFERENCES

- Allen, J. (2014). Relationships of examinee pair characteristics and item response similarity. In N. M. Kingston and A. K. Clark (Eds.) *Test fraud: Statistical detection and methodology* (pp. 23–37). Routledge: New York, NY.
- Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69, 44–49.
- Barnett, V. & Lewis T. (1994). *Outliers in statistical data*, 3rd Edition, pp. 269–270. John Wiley and Sons: Chichester, UK and New York, NY.
- Bellezza, F. S. & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, 16, 151–155.
- Bellezza, F. S. & Bellezza, S. F. (1995). Detection of copying on multiple choice test: An update. *Teaching of Psychology*, 22, 180–182.
- Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, 50(2), 141–163.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 46, 443–459.
- Feller, W. (1968). *An introduction to probability theory and its applications*. Volume I, 3rd Edition (Revised Printing). John Wiley & Sons, Inc.: New York, London, Sydney. p. 279.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 6, 152–165.
- Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete mathematics*, 2nd Edition. Addison-Wesley Publishing Company: Reading, MA. Chapter 7.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying*. ACT Research Report Series. 87–15. September 1987.
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index*. ETS Program Statistics Research. Technical Report No. 96–4.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates: Hillsdale, NJ.
- Maynes, D. D. (2009, April). *Combining statistical evidence for increased power in detecting cheating*. Presented at the annual conference of the National Council on Measurement in Education, San Diego, CA.
- Maynes, D. D. (2013, October). *A probability model for the study of similarities between test response vectors*. Presented at the Conference of Statistical Detection of Potential Test Fraud, Madison, WI.
- Maynes, D. D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston and A. K. Clark (Eds.) *Test fraud: Statistical detection and methodology* (pp. 53–82). Routledge: New York, NY.
- Tucker, A. (1980). *Applied combinatorics* (pp. 111–122). John Wiley & Sons, Inc.: New York, Brisbane, Chichester, Toronto.
- van der Linden, W. J. & Sotaridona, L. S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31(3), 283–304.
- Wesolowsky, G. O. (2000). Detection excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, 27(7), 909–921.
- Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21(4), 307–320.
- Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement*, 25(4), 385–404.
- Wollack, J. A., & Maynes, D. D. (2011, February). *Data forensics: What works, what doesn't*. Presentation at the annual conference for the Association of Test Publishers, Phoenix, AZ.
- Zhang, Y., Searcy, C. A., & Horn, L. (2011, April). *Mapping clusters of aberrant patterns of item responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.