A Probability Model for the Study of Similarities between Test Response Vectors

Presented at the Conference on Statistical Detection of Potential Test Fraud, Madison, WI, 10/18-10/19, 2013

October 8, 2013 (updated November 14, 2013)

Dennis Maynes

Abstract

This paper proposes a model for evaluating the similarity between two test response vectors as a means of testing the hypothesis of independent test taking. Previously published approaches have ignored the fact that the distribution of the similarity statistic is constrained, which can weaken the power of a statistical test. The models in this paper illustrate, more clearly than any previously published paper, the mechanisms which underlie the probability distributions of similarity statistics.

Introduction

In reference to the detection of cheating by copying or sharing answers on tests, William Angoff wrote (1974), "The obvious method is to compare the responses on the suspect answer sheet with responses on the answer sheets of examinees seated nearby and to look for greater-than-normal similarities." He then stated, "Even a brief consideration of this solution makes it clear that the complexities in making theoretical estimates of such a distribution are far too great to make it practical." This paper provides a solution to this impractical problem. Yes, there is difficulty in estimating the distribution of the similarities between two test responses. In fact, most researchers have attempted to find approximations which could be implemented practically. Because today's computers are able to generate millions of random test response vectors in minutes or even seconds, it is reasonable to use simulation techniques for studying the distribution of the similarities between two test responses.

Summary of Similarity Statistics Research

Since the 1920's, researchers have published and proposed statistical methods for detecting answer copying between students on exams (Frary, 1993). Nearly all statistical methods that have been used for detecting answer copying have relied upon the number of answers that were the same between two test takers. The earliest researchers used the number of incorrect answers that were the same when two test response vectors were compared. The trend to use only identical incorrect answers has continued to the present day (for example, see Belleza & Belleza, 1989). The reason for using only identical incorrect answers was summarized by Frary (1993) when he stated that Angoff favored using only identical incorrect answers because "nonstatisticians might attribute right-answer correspondences

to jointly held knowledge no matter how small the probability that an actual instance occurred in the absence of copying."

The debate between restricting the analysis to use only identical *incorrect* responses or to use *all* the identical responses has been ongoing in the literature for several years. For example, ETS has adopted the position that only identical incorrect responses should be used (Holland, 1996). Holland justifies this position with the following quotation: "If two examinees answer the same item correctly, this in itself would constitute evidence of only a very weak nature that some collaboration or copying took place, for it might be entirely reasonable to presume simply that both examinees knew the correct answer to the item (Buss & Novick, 1980)." However, in the same paragraph, Holland stated, "This common-sense view flies in the face of strong empirical evidence that close agreement on correct answers is unusual, rather than common, for examinees with similar scores."

Despite Angoff's concern, cited by Frary, and the position taken by ETS, other researchers have pressed forward and analyzed both identical correct and identical incorrect responses between a pair of test takers (Frary, 1993; Wesolowsky, 2000; van der Linden & Sotaridona, 2006). These researchers have used the total number identical responses (i.e., using correct and incorrect responses) as the statistic of interest. They have modeled the distribution of the number identical responses probabilistically using a generalized binomial distribution where the probability of a match depends upon the performance of each test taker. They have also assumed statistical independence so that the match probabilities could be computed by summing products of response probabilities.

Another line of research that has been followed is that of "source-copier" analysis (for examples see Frary, Tideman, and Watts, 1977; and Wollack, 1997). The basic idea of this approach is to condition the response distribution of one test taker (usually referred to as the "copier") upon the observed set of responses for the other test taker (usually known as the "source"). With this approach, the statistical distributions differ (which means different probabilities) depending upon which test taker is assigned the role of source and which is assigned the role of copier.

Questions concerning the Distribution of Similarity Statistics

The above discussion briefly summarizes some of the research in answer-copying statistics which has led to the development of similarity statistics as a means of detecting potential test fraud. In this regard, the primary decisions that have faced researchers have been:

1. How should the statistical distribution of the selected statistic be modeled?
   Some researchers have compared the observed statistic against empirical data (for example see Hanson, Harris, & Brennan, 1987). Others have assumed that item response selection may be modeled using some parametric equation and estimated from the population of test takers (for example see Wesolowsky, 2000; van der Linden & Sotaridona, 2006; Wollack, 1997). Finally, an argument was made by van der Linden & Sotaridona (2004) that "use of population-based statistical tests may be unfair" because if the population changes the statistical probabilities may change.

2. What statistical assumptions should be made?
3. Which statistic should be used? Should only incorrect identical responses be used? A wide variety of statistics have been studied. For example, Angoff (1974) studied the number of identical incorrect answers, the maximum number of identical incorrect answers within a run of identical answers, and the number of same omitted answers. Statistics studied by Hanson, et. al. (1987) considered bivariate combinations of the number of identical answers, the number of identical incorrect answers, the longest sequence of identical answers, the number of incorrect answers in the longest sequence of identical answers, and the maximum number of incorrect answers within a run of identical answers.
4. How should probabilities be reported?

This paper does not address all of the above questions. However, this paper does address some of the fundamental and more relevant questions. The probability models derived in this paper assume:

1. Response selections between test takers are stochastically and statistically independent.
2. The probabilities of response selections depend upon test taker performance.
3. The statistical test will provide the same result whether one test taker is presumed to be a copier or not.
4. Both identical correct and identical incorrect responses can be evaluated statistically.

Region of Permissible Values for Similarity Statistics

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. 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 is shown in Table 1.

Table 1: Example Agreement between Test Takers T1 and T2 with Scores of 42 and 45

T1 / T2	Correct	Incorrect	Total
Correct	42 to 27	0 to 15	**42**
Incorrect	3 to 18	12 to 0	**18**
Total	**45**	**15**	**60**

The marginal totals in Table 1 are provided in bold font because they are fixed. Given the data in Table 1, the greatest number of questions where both Test Takers T_1 and T_2 answer correctly is 42 and the least number is 27. In other words, the test takers MUST answer at least 27 of the same questions correctly and CANNOT 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 3 remaining questions.

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 2.

Table 2: Agreement between Test Takers T1 and T2 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

The quantities in Table 2 have been shown as

Equation 1-A

$$N_{11} = R,$$

Equation 1-B

$$N_{12} = Y_1-R,$$

Equation 1-C

$$N_{21} = Y_2 - R, \text{ and}$$

Equation 1-D

$$N_{22} = N + R - (Y_1 + Y_2).$$

In Table 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* to the remainder of this paper. After comparing the region of permissible values with the values allowed by proposed statistical methods, it is apparent that most researchers have neglected to account for the region of permissible values.

The remainder of this paper is devoted to the development and validation of probability models that comply with the region of permissible values.

Probability Model 1: Equally difficult items

The simplest model assumes that the probabilities of identical correct and incorrect responses are constant. This is accomplished by assuming that the items are equally difficult for each test taker.

We make the following assumptions:

1. Item responses are conditionally independent and only depend upon test taker performance,
2. Each item has one and only one keyed correct answer,
3. The probability of an identical correct answer depends only upon the performance of Test Takers T1 and T2, and is constant for all items, and
4. The probability of an identical incorrect answer depends only upon the performance of Test Takers T1 and T2, and is constant for all items.

Under the above conditions, the probabilities of identical responses and non-identical responses are:

Probability Model (1)

1. $P(\text{Identical Correct}) = (Y_1/N) \times (Y_2/N)$, where Y_1 is the score for Test Taker T1, Y_2 is the score for Test Taker T2, and N is the number of questions on the test.

2. P(Identical Incorrect) = (1-Y_1/N) x (1-Y_2/N) x $\overline{D}$, where $\overline{D}$ is the mean value of D_i, with D_i being the sum of squared selection frequencies for the incorrect responses for item *i*, divided by the squared total number of incorrect responses for item *i*.
3. P(Different Response) = (1-Y_1/N) x (Y_2/N) + (Y_1/N) x (1-Y_2/N) + (1-Y_1/N) x (1-Y_2/N) x (1- $\overline{D}$).

In order to evaluate Probability Model 1 a simple simulation was performed. The simulation generated test response vectors for a 20-item 4-choice multiple choice test where Y_1 was set to 14 and Y_2 was set to 15. Probabilities of correct answers were assumed constant for all questions and equal to Y_1/N or Y_2/N. Probabilities of incorrect responses conditioned upon selection of an incorrect response were assumed constant and equal to 6/11, 3/11, and 2/11 (Relative ratios of 1/1,1/2, and 1/3). While these values would not be expected to occur in actual testing, they provide the advantage of being able to analyze the probability distribution with a minimum number of extraneous factors (e.g., varying the difficulty of the items and the response selection probabilities). The simulation consisted of 100,000 pairs of random response vectors that were constrained so the scores were equal to the selected values of Y_1 and Y_2. The resulting bivariate counts of R and W are shown in Table 3.

Table 3: Counts of Simulated R-W values from 100,000 Test Pairs

Identical Incorrect Responses, W	Identical Correct Responses, R					
	9	**10**	**11**	**12**	**13**	**14**
0	13003	22907	12536	2496	163	2
1		15717	16992	5105	467	9
3			5615	3444	474	8
4				782	241	7
5					31	1
6						0

In Table 3, above, the value of 13003 indicates the number of times the pair of zero (0) identical incorrect answers and nine (9) identical correct answers was observed. Using the simulation parameters, the probability of an identical correct answer was 0.525 (0.70 x 0.75), the probability of an identical incorrect answer was 0.0304 (0.30 x 0.25 x (0.5455 x 0.5455 + 0.2727 x 0.2727 + 0.1818 x 0.1818)), and the probability of a identical answer was the sum of both, 0.5554.

It has been assumed in the literature that the total number of identical responses would follow a binomial or a generalized binomial distribution (see van der Linden & Sotaridona (2006) or Wesolowsky (2000)). The problem with using a binomial or generalized binomial distribution is that it does not conform to the region of permissible values. This is illustrated in Table 4 where the count data of the number of identical responses are shown, along with expected values from the binomial distribution.

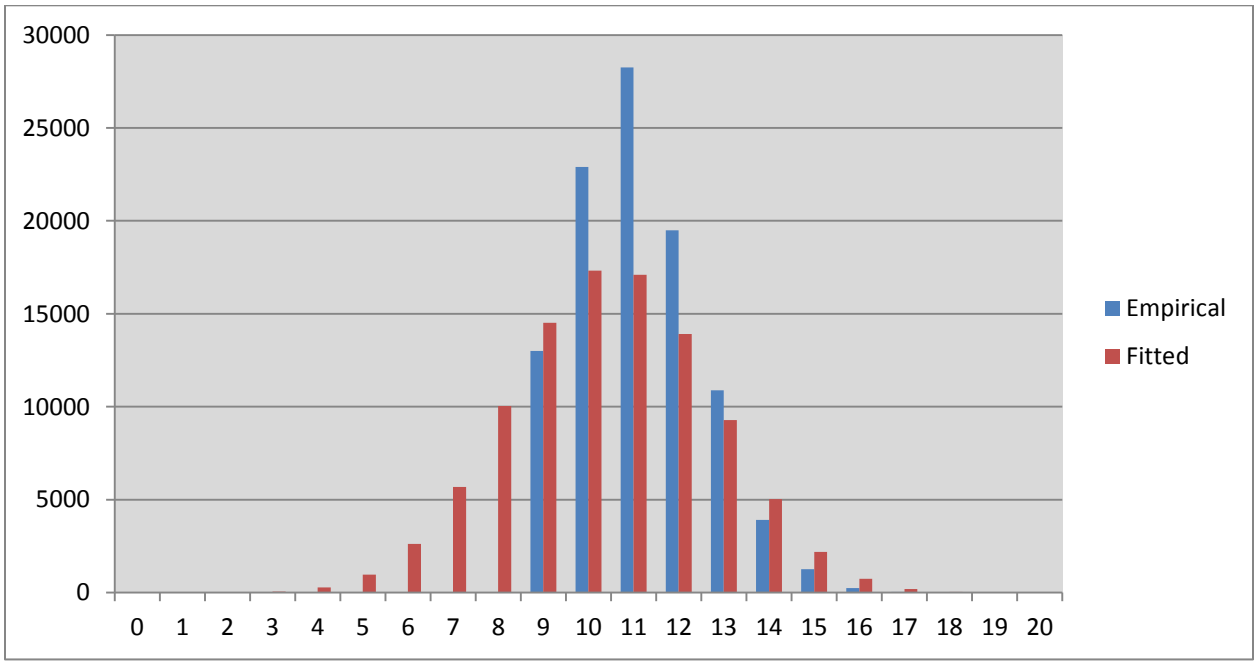
Table 4: Counts of Simulated Identical Responses from 100,000 Test Pairs

	0	1	2	3	4	5	6	7	8	9	10
Count										13003	22907
Expected	0	1	9	60	279	967	2624	5693	10038	14520	17329

	11	12	13	14	15	16	17	18	19	20
Count	28253	19488	10883	3913	1265	249	38	1		
Expected	17092	13908	9286	5037	2186	741	189	34	4	0

Table 4 shows that simply using the binomial distribution as an approximating distribution for the count of the number of identical responses models results in probabilities for impermissible values (i.e., those below 9 and above 18). If a normal distribution were postulated, instead of the binomial distribution, the fit would be worse because the normal distribution is continuous and has infinite tails. This is one of the key points of this paper. *Ignoring the region of permissible values may result in inappropriate probability computations*. Despite this criticism that most researchers have used distributions with tails that extend beyond the region of permissible values, these distributions may be useful as long as the fit of the approximating distribution is sufficiently close to the actual distribution in the upper tail. The data from Table 4 are shown in Figure 1.

Figure 1: Binomial Fit to Counts of Simulated Identical Responses



In Figure 1, above, the binomial distribution appears to provide a reasonable approximation in the upper tail. The error in the approximation appears to be conservative, because the upper tail for the actual distribution is bounded and decreases quicker than the tail for the binomial distribution. As an example, consider a value of 16 identical responses or greater. The binomial distribution would provide an upper tail probability value of 0.0097 where the actual probability value would be estimated from the count data at 0.0029.

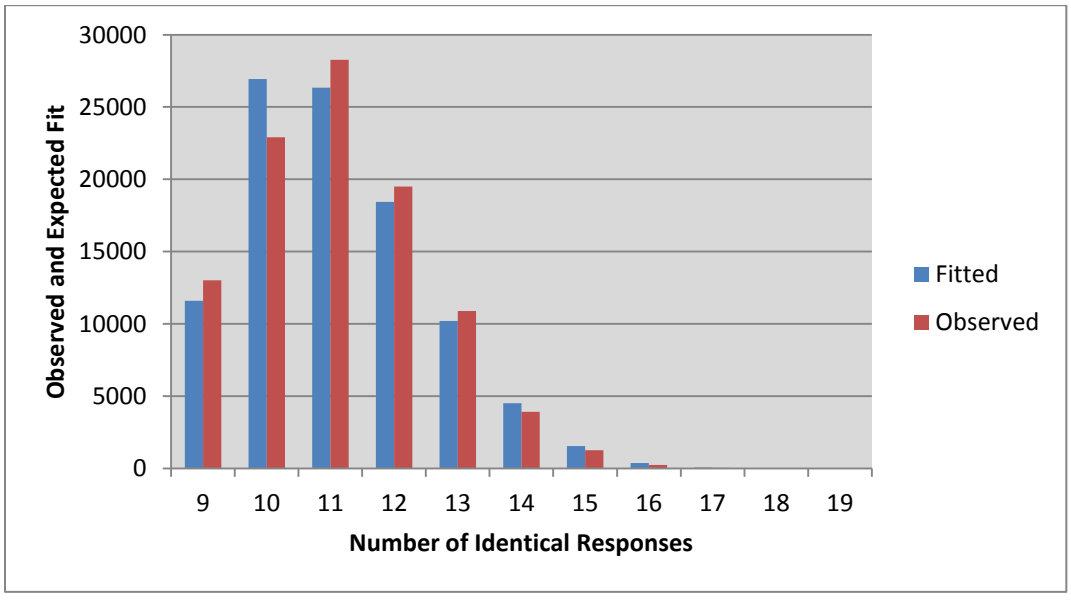
Having established that the unrestricted binomial distribution is a poor fit for the actual distribution of the number of identical responses, it seems reasonable to use another distribution to approximate the theoretical distribution. A probability density function that is able to model a bounded and skewed distribution (like the one seen in Figure 1) is the beta probability density function (pdf). The beta pdf is written as

Equation 2

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{\beta(a,b)}; \quad 0 \le x \le 1; \;\; 0 < a, b; \;\; \text{and} \;\; \beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Using the method of moments (i.e. equating theoretical moments to computed moments), the parameters for the beta pdf are a=2.36 and b=7.60, with boundaries of 8.5 and 19.5. The fitted distribution is shown in Figure 2.

Figure 2: Beta PDF Fit to Counts of Simulated Identical Responses



The Chi-Square statistic is quite large for this fit (approximately 1200 with 10 df), but the approximation appears to closely approximate the theoretical distribution in the upper tail. The advantage of this approximation is the ability to estimate the parameters of the approximating distribution from the computed moments which require less data than estimation techniques, such as maximum likelihood and the method of scoring.

Probability Model 1 assumes that all items were equally difficult for each test taker, but the probability of a correct response depended upon test taker performance. In practice, we would not expect items to be equally difficult. Even so, studying this model has two advantages: (1) the statistical distributions are tractable allowing study of the structure of the probability mechanism, and (2) constant probabilities may allow for reasonable approximations to the actual probabilities. Given the relationships in Equations 1) between the number of correct and incorrect identical responses, a compound hypergeometric-binomial probability model is reasonable. Under this model, the distribution of the number of identical correct responses is assumed to follow the hypergeometric model and the conditional distribution of the number of identical incorrect responses is assumed to follow the binomial model. The hypergeometric probability mass function is written as

Equation 3

$p(R|N,Y_1,Y_2) = \frac{\binom{Y_2}{R}\binom{N-Y_2}{Y_1-R}}{\binom{N}{Y_1}}$, and the conditional binomial probability mass function is written as

Equation 4

$$p(W|N,Y_1,Y_2,R,p_w) = \binom{N+R-Y_1-Y_2}{W}(p_w)^W(1-p_w)^{N+R-Y_1-Y_2-W},$$

where p_w is the probability of an identical incorrect response conditioned upon both test takers answering the item incorrectly, and the other quantities are as defined in Equations 1 and in the accompanying text below those equations. The joint probability distribution of *R* and *W* is found by multiplying the expression of Equation 3 with the expression of Equation 4. The probability mass function for the number of identical responses, *M=R+W*, is given by summing the appropriate probabilities across the joint probability distribution. The joint probabilities have been computed for the illustration using Equation 3 and Equation 4. They are shown in Table 5.

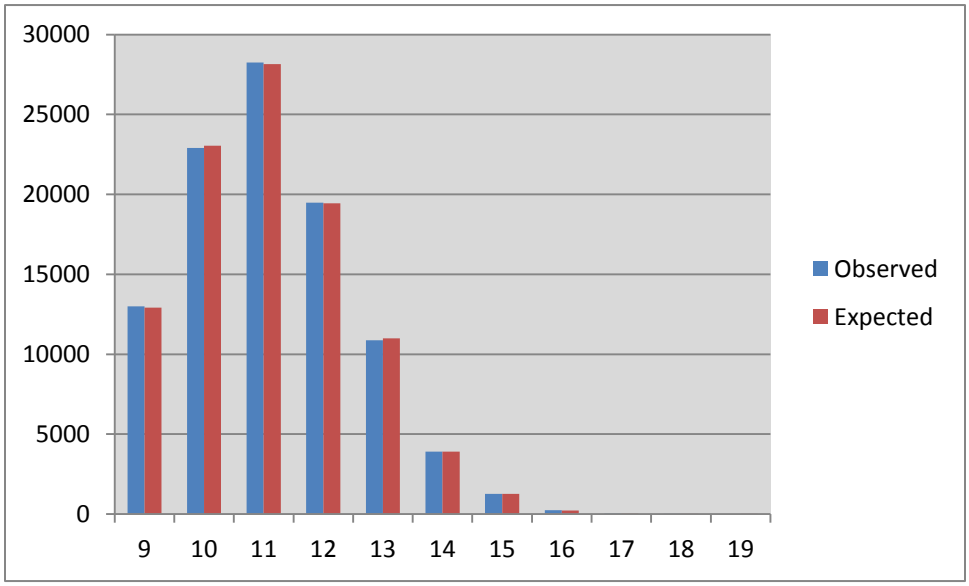
Table 5: Joint Probability Mass Function for Illustrated Data

W/R	9	10	11	12	13	14
0	0.12912797	0.23050943	0.12469330	0.02473256	0.00169810	0.00002887
1		0.15687447	0.16972144	0.05049564	0.00462261	0.00009824
2			0.05775244	0.03436509	0.00471892	0.00013371
3				0.00779578	0.00214099	0.00009100
4					0.00036427	0.00003096
5						0.00000421

The Chi-Square goodness of fit statistic for the probabilities in Table 5 using the simulated counts from Table 3 is equal to 16.6 with 20 degrees of freedom. The goodness-of-fit test is not rejected. The proposed model is a very good fit to the data, even though 100,000 test pairs were sampled. When the probabilities are summed in order to obtain the assumed probability mass function of the number of

identical responses (per Equation 3 and Equation 4), the Chi-Square goodness of fit statistic is equal to 8.3 with 10 degrees of freedom. The goodness-of-fit test is not rejected. The observed and expected values have been plotted in Figure 3.

Figure 3: Hypergeometric-Binomial Fit to Counts of Simulated Identical Responses



The fit to the simulated data shown in Figure 3 is impressive and lends credibility to the assumption that the hypergeometric-binomial compound distribution is a proper model for the number of identical responses shared between two test takers. However, the model was postulated when the items are equally difficult. When the probabilities of answering the test questions correctly vary or are unequal, the hypergeometric-binomial model would provide approximate probabilities.

In this section of the paper, three different probability models have been considered using simulated data only. Two of these models are lacking when the items are equally difficult: the binomial distribution because it does not conform to the region of permissible values, and the beta probability density because it appears to not properly fit the data. At best, both of these models would provide approximate probabilities for assessing the observed similarity between two test response vectors. The final model was a compound distribution of the hypergeometric and binomial distributions. Because this model provides a very good fit to the simulated data, it is investigated more carefully in the next two sections.

Suitability of Hypergeometric-Binomial density for Probability Model 1

The above **Hypergeometric-Binomial** model assumed that the probabilities of correctly answering a question were constant across questions but depended only upon test taker performance (i.e., the score for the test taker was fixed). *Constraining the distribution in this way models the test taker's proficiency appropriately and the test taker's actual probability of answering the test questions correctly is not needed, as long as the probabilities of correctly answering the questions are equal.* For example, the

simulation of 100,000 test pairs was repeated, but instead of using Y_1/N or Y_2/N as the probabilities of responding correctly, the probabilities were set to 0.5 (or one-half). The simulation took more time because more test response vectors were rejected. However, all test response vectors were subject to the constraints that the scores would be exactly equal to Y_1 and Y_2. The comparison between the two simulation runs is shown in Table 6.

Table 6: Effect of Using Constant Probabilities on the Distribution

	Number of Identical Responses										
	9	10	11	12	13	14	15	16	17	18	19
P's = Y1/N & Y2/N	13003	22907	28253	19488	10883	3913	1265	249	38	1	0
P's = 1/2	12825	23181	28110	19476	10965	3909	1262	227	42	3	0
Expected Value	12914	23044	28181.5	19482	10924	3911	1263.5	238	40	2	0
Chi Square	1.23	1.63	0.36	0.00	0.31	0.00	0.00	1.02	0.20	1.00	0.00

The values in Table 6 that correspond to the rows "P's = Y1/N & Y2/N " and "P's =1/2" are the numbers of test pairs that were observed for each count of identical responses. The Chi-Square statistic (5.75 with 9 d.f.) is not significant. As far as can be verified, the simulated distributions of the number of identical responses, after conditioning upon the scores of Y_1 and Y_2 are identical. The test takers' actual probability of answering the question correctly has been conditioned out and is no longer needed. It is for this reason that the probabilities of correct responses are not present in Equation 3.

Consequences of assuming that items are equally difficult

The assumption of that all items are equally difficult, while appearing to be trivial is not supportable in practice. To demonstrate the difficulty with this assumption, assume that both test takers have identical scores. Under the assumption that the items are equally difficult, the probability of an identical correct answer is a constant value and is equal to $c^2 = \left(\frac{Y_1}{N}\right)^2$. Then the expected number of identical correct responses is equal to Nc^2. However, if the individual matching probabilities are not constant, the expected number of identical correct responses will be greater than Nc^2, as illustrated by writing

Equation 5

$$V(x) = \frac{\sum(p_i - c)^2}{N-1} = \frac{\sum p_i^2 - Nc^2}{N-1} \geq 0,$$

where p_i is the probability of a correct response and c is the average or expected value of the p_i values. It is well known in statistics that the minimum value of the quantity in Equation 5 is only realized when all the p_i values are the same. Under any other condition, the expected value of the sum of squares will be greater than N times the squared expected value. *Hence, an assumption of equal difficulty always results in underestimating the expected value of the number of identical correct responses in practice*. The underestimation can be quite large and is directly related to the variance of the matching probabilities. The same analysis applies to the conditional probability of observing an identical incorrect response.

The primary finding of the preceding analysis is that the test response data should not be pooled across items to estimate matching probabilities. Doing so will result in approximating distributions that are biased towards the lower tail which will spuriously raise false positive rates. Researchers have made this assumption explicitly (Belleza & Belleza, 1989) and implicitly (Angoff, 1974).

Probability Model 2: Unequally difficult items

Because of the under-estimation bias in the expected value of the number of matching responses that results from assuming the items are equally difficult, the Hypergeometric-Binomial model cannot be recommended for use in practice. Thus, Probability Model 1, while useful for didactic purposes, is not useful for practical application. The model needs to assume that items are unequally difficult.

At this point it is useful to summarize key points made in this paper so far:

1. The distribution of the match statistic depends upon the number of correctly answered questions by each test taker.
2. An assumption of equally difficult items is untenable and should not be used because it leads to a biased approximation of the distribution of the match statistic. The bias is due to underestimating the expected value of the distribution. This underestimation will likely result in computing upper tail probabilities which are too small.
3. Equation 5 can be used to show that an assumption of constant probability for providing incorrect identical answers between two test takers will result in a biased distribution.
4. We should not expect that a continuous distribution with infinite tails, such as the normal, to provide a reasonable approximation for the left tail of the similarity distribution, because the distribution is truncated and right-skewed.
5. We should not expect a continuous bounded distribution, such as the beta, to provide close approximations in the upper tail of the distribution. On the other hand, a probability-based discrete distribution should be able to approximate the upper tail of the distribution properly.

The above points indicate that a probability model should be considered which includes unequally difficult items. It is hoped that such a model will provide sufficiently accurate results for practical implementation. The revised model should

1. Assume local independence,
2. Use item response probabilities that are based on item difficulties *and* test taker performance, and
3. Generate a discrete probability mass function which obeys the constraints of the possible number of matching responses.

To further define the enhancements, we make the following assumptions:

1. Item responses are conditionally independent and depend upon test taker performance and item difficulty,
2. Each item may have more than one correct answer,

3. Each item response is dichotomously scored with a value of 0 or 1,
4. The probability of an identical correct answer depends upon the performance of Test Taker T1 and T2 and the item difficulty, and
5. The probability of an identical incorrect answer depends upon the performance of Test Taker T1 and T2, and the item difficulty.

Under the above conditions, the probabilities of identical responses and non-identical responses may be written as

Probability Model (2)

1. p_i is the item difficulty and is the overall probability of answering an item correctly,
2. C_i is the sum of squared selection frequencies for the correct responses for item *i*, divided by the squared total number of correct responses for item *i*.
3. D_i is the sum of squared selection frequencies for the incorrect responses for item *i*, divided by the squared total number of incorrect responses for item *i*.
4. θ_1 is the value which solves $\frac{Y_1}{N} = \sum_{i=1,N}\left[\frac{p_i}{p_i+(1-p_i)e^{-\theta_1}}\right] = \sum_{i=1,N} p_{i,\theta_1}$.
5. θ_2 is the value which solves $\frac{Y_2}{N} = \sum_{i=1,N}\left[\frac{p_i}{p_i+(1-p_i)e^{-\theta_2}}\right] = \sum_{i=1,N} p_{i,\theta_2}$.
6. P_i(Identical Correct) = $p_{i,\theta_1} p_{i,\theta_2} C_i$.
7. P_i(Different Correct) = $p_{i,\theta_1} p_{i,\theta_2}(1 - C_i)$.
8. P_i(R_1 Correct, R_2 Incorrect) = $p_{i,\theta_1}\left(1 - p_{i,\theta_2}\right)$.
9. P_i(R_1 Incorrect, R_2 Correct) = $\left(1 - p_{i,\theta_1}\right)p_{i,\theta_2}$.
10. P_i (Identical Incorrect) = $\left(1 - p_{i,\theta_1}\right)\left(1 - p_{i,\theta_2}\right)D_i$
11. P_i (Different Incorrect) = $\left(1 - p_{i,\theta_1}\right)\left(1 - p_{i,\theta_2}\right)(1 - D_i)$.

In Probability Model 2, the variables R_1 and R_2 represent the responses of Test Taker 1 and Test Taker 2. The values of θ and the probabilities that Test Takers 1 and 2 answer the item correctly are derived using a logit transformation of the item difficulty, which is expressed as $\tau_i = ln\left(\frac{p_i}{1-p_i}\right)$.

Probability estimates may be obtained from Probability Model 2 by simulation or by using recurrence equations.

In order to validate Probability Model 2, two simulations were performed and compared. The first simulation, based on actual test result data, was designed to model the distribution of independent test taking. The second simulation, produced using Probability Model 2, was performed to assess the model's goodness-of-fit to the distribution of independent test taking produced by simulation using the live data.

Simulation 1: Independent Test Taking

This simulation was set up and performed in the following manner:

1. The live test result data were taken from the FCAT (Florida Comprehensive Assessment Test), Grade 6 Reading test.
2. The Grade 6 Reading FCAT consists of a set of scored questions. The 45 scored questions were the same on all forms. And, 8 field test questions were unique to each form. The answers to the field test questions were discarded in order to ensure that every pair of compared tests was based on exactly the same questions (45).
3. Each test record within a school and subject was assigned a random value between 1 and 1,999 without replacement, which is referred to as a "virtual school." The test records in excess of 1,999 for a school were discarded. Only tests within a virtual school were compared. This procedure guaranteed that every pair of test responses which were compared came from different physical schools. This procedure allows live data to be used which conform to the assumption of independent test taking, which is the null hypothesis for evaluating the amount of similarity observed between two test response vectors.
4. The number of matching answers was tabulated into a histogram for every pair of test responses which was compared. The histograms were subdivided by the number of correct answers for the two tests which were compared. This allowed conditional distributions to be tabulated which would conform to the constrained region of permissible values.

Simulation 2: Probability Model 2

This simulation was set up and performed in the following manner:

1. Two test scores, Y_1 and Y_2, were randomly selected using a uniform distribution between the values of 0 and N.
2. Values of θ_1 and θ_2 were computed using the two test scores.
3. For the N items, binary values of correct/incorrect and matching/not matching were randomly generated as specified by Probability Model 2. These values were summed to compute the number of correct responses, the number identical correct responses, and the number identical incorrect responses for the simulated pair of tests.
4. The results were tabulated into a histogram for every pair of test responses which was generated. As with Simulation 1, the histograms were subdivided by the number of correct answers for the two tests which were compared. This allowed conditional distributions to be tabulated which would conform to the constrained region of permissible values.

For Simulation 1, there were a total of 9,721,344 pairwise comparisons performed. The number of pairs for each score combination varied greatly, with the greatest number of pairs associated with scores in the middle of the distribution.

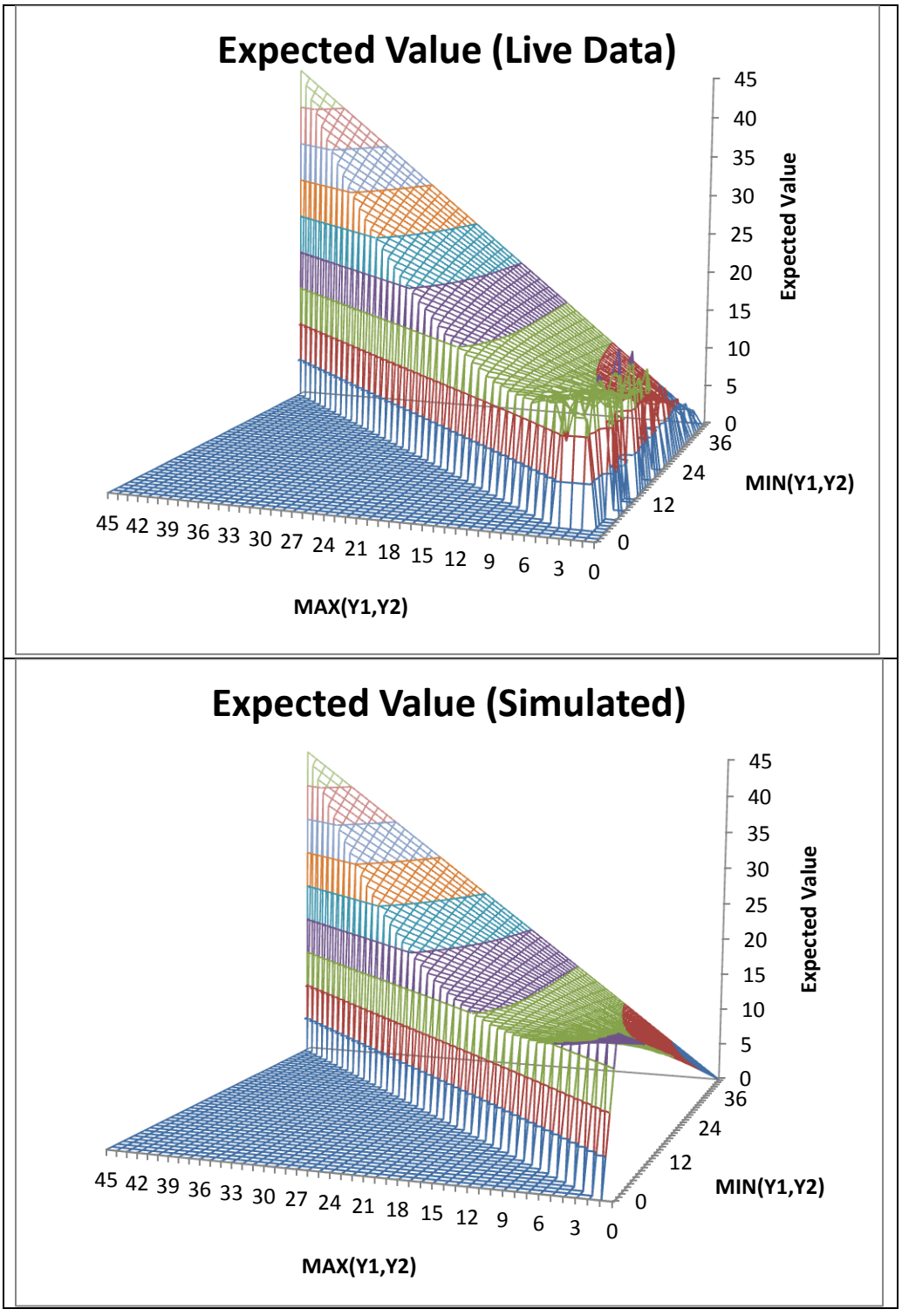
For Simulation 2, there were a total of 10,000,000 pairwise comparisons generated. There were approximately 9,400 generated pairs for score combinations where the two scores differed and approximately 4,700 generated pairs for score combinations where the two scores were the same.

There were a total of 1,081 score combination pairs (46 x 47 / 2). Of these, Chi-Square goodness-of-fit tests were performed for 889 combinations. The remaining 192 combinations were associated with very

low scores which were not present in the live data set. Nearly all of the Chi-Square goodness-of-fit tests were statistically significant due to large sample sizes. However, plots of the data reveal where the probability model is aligned with the live data.

Figure 4 shows the expected value of the number of matches from the live and the simulated data. In order to compare the expected values and show the relationship to the number correct scores, the plots have been created using a three-dimensional wireframe. Because the similarity function (i.e. the count of the number of matching answers) is symmetric in the number of correct responses for Test Taker 1 and Test Taker 2 (i.e. f(Y1,Y2) = f(Y2,Y1)), the distributions in Figures 4, 5, and 6 have been folded. Doing so, allows increased precision in the computations and it also aids in visualizing the shapes of the response curves of the expected value and the standard deviation.

Figure 4: Expected Value of the Number of Matches

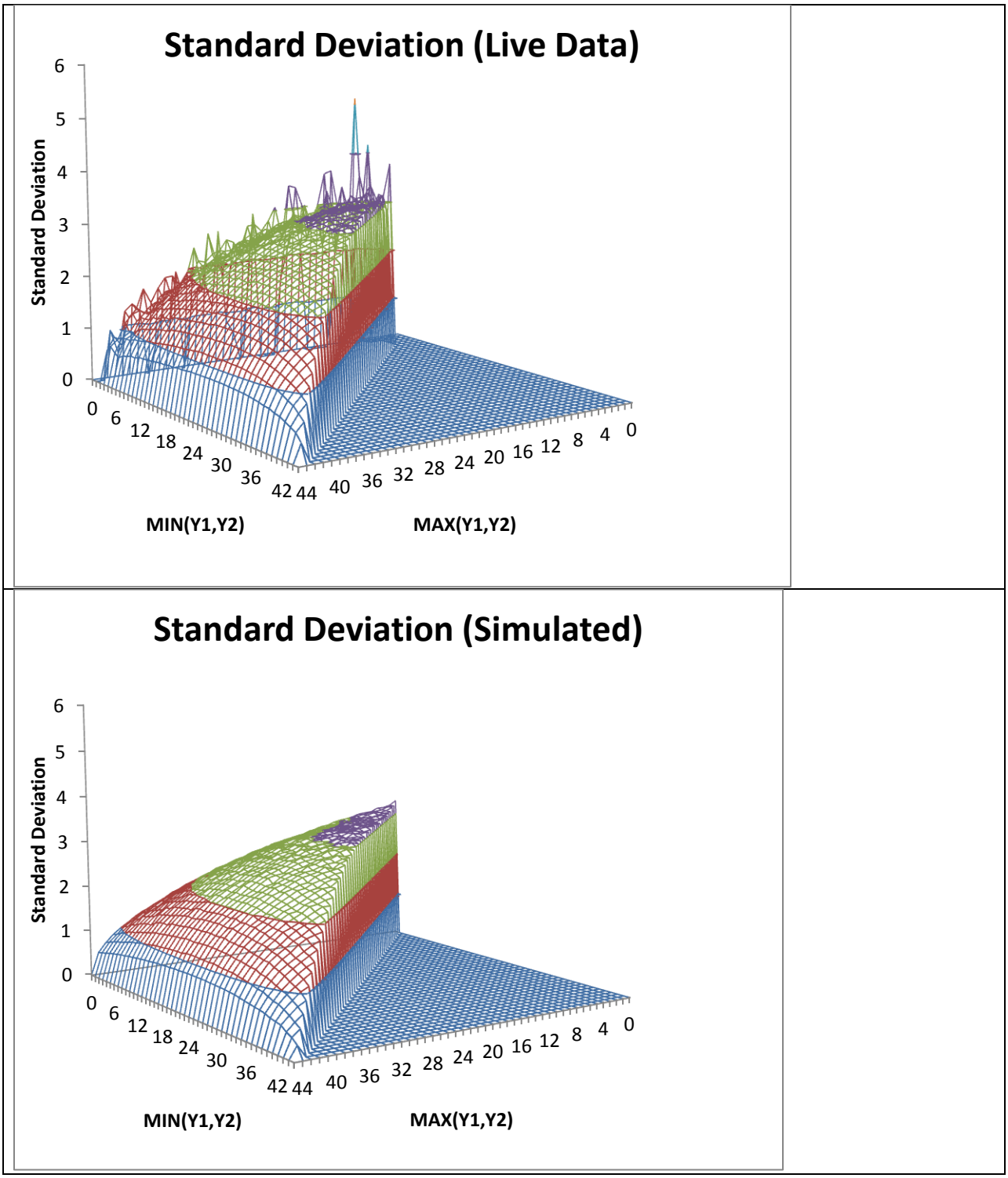


In Figure 4, above, the upper panel shows the expected values computed from the live data and the lower panel shows the expected values computed from the simulated data (i.e. the model). The

response functions for the two data sets appear to be very close, suggesting that the model has captured important characteristics of the expected value of the probability density function for the number of identical answers between two test responses. The jaggedness that is observed in the lower score ranges in the upper panel is due to a scarcity of scores in the live data. It should also be noted that the response surface is not linear. The amount of curvature is not constant, thus if a regression model were used to estimate the expected value, it should minimally include interaction and quadratic terms.

Figure 5 shows the standard deviation of the number of matches from the live and the simulated data. In order to compare the standard deviations and show the relationship to the number correct scores, the plots have been created using a three-dimensional wireframe.

Figure 5: Standard Deviation of the Number of Matches



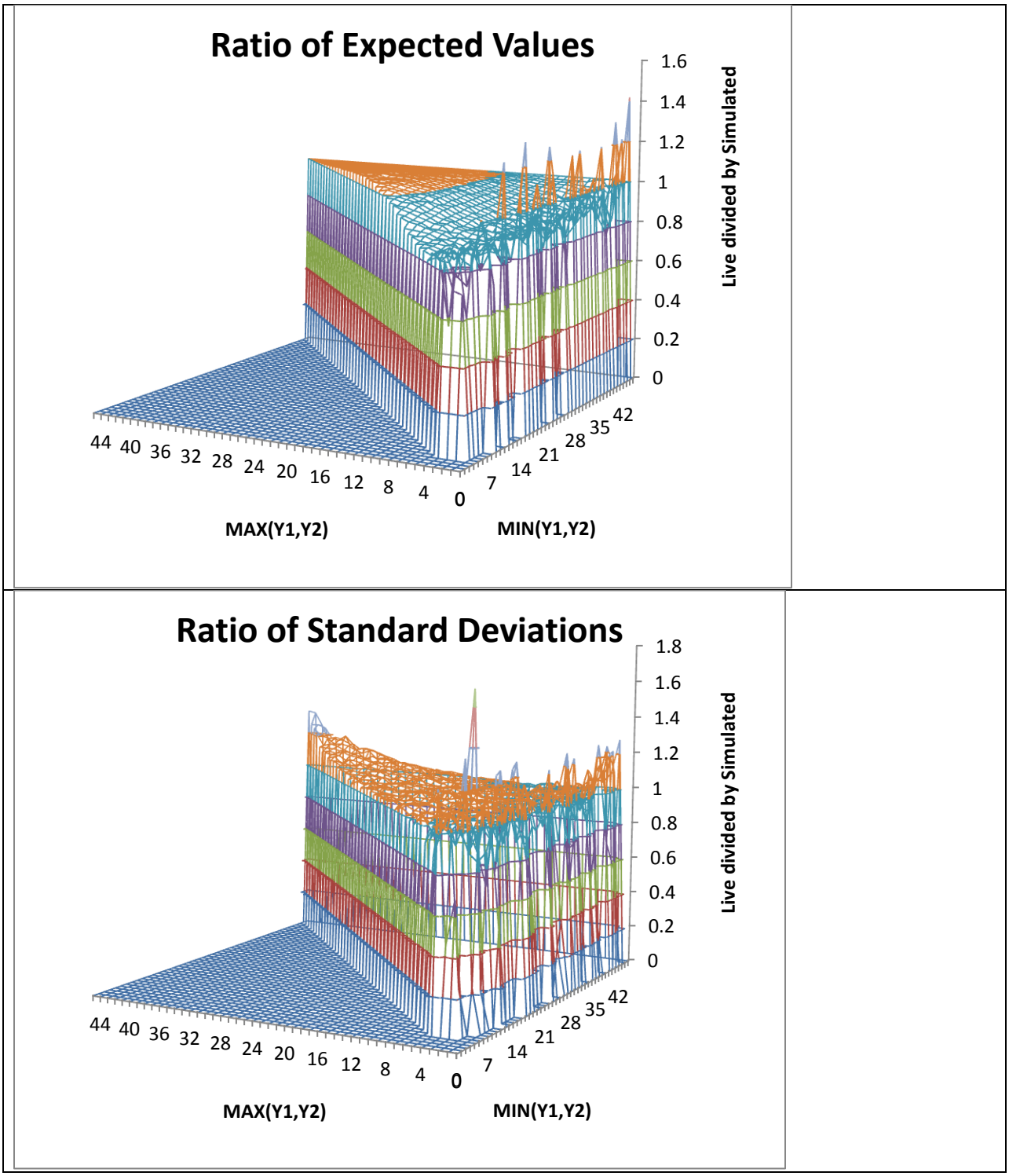
In Figure 5, above, the upper panel shows the standard deviations computed from the live data and the lower panel shows the standard deviations computed from the simulated data (i.e. the model). The

response functions for the two data sets appear to be very close, suggesting that the model has captured important characteristics of the standard deviation of the probability density function for the number of identical answers between two test responses. The jaggedness that is observed in the lower score ranges in the upper panel is due to a scarcity of scores in the live data. It should also be noted that the response surface is not flat and it is not linear. Or, in other words, the variances are not homogeneous. This means that weighted least squares is required if a regression model were used to compute probabilities of similarities.

A problem for future research is the comparison of variances from the model with variances computed using the generalized binomial distribution. One would expect that the latter variances would be larger than the variances estimated from the model because the model conforms to the bounded nature of the region of permissible values and the generalized binomial distribution does not.

In order to compare the expected value and the standard deviation for the live and simulated data, the ratios of the two have been plotted in Figure 6, with the value from the simulated data in the denominator and the value from the live data in the numerator.

Figure 6: Ratio of Expected Values and Standard Deviations



In Figure 6, above, the upper panel provides the ratio of the expected value from the live data to the expected value from the simulated data (i.e. the model). The expected value from the simulated data is

nearly equal to the expected value from the live data for medium to high score ranges. However, the expected value from the simulated data is greater than the expected value from the live data for low score ranges. In general, over-estimation of the expected value by Probability Model 2 will result in a lower value of the Type I error (i.e., detection of false positives) than the live data would indicate. The analysis of Equation 5 suggests that the over-estimation of the expected value is due to more variability in the probabilities of identical incorrect answers at the lower score ranges than is actually present. Thus, the assumption of a constant ratio of identical incorrect responses, after conditioning for the total test score, is called into question.

In Figure 6, above, the lower panel provides the ratio of the standard deviation from the live data to the standard deviation from the simulated data (i.e. the model).The standard deviation from the simulated data is generally smaller than the standard deviation from the live data (as evidenced by a ratio that is greater than 1). This is especially the case for very high score values. In general, under-estimation of the standard deviation by Probability Model 2 will result in a higher value of the Type I error (i.e., detection of false positives) than the live data would indicate.

In order to assess how the over- and under-estimation errors by Probability Model 2 may affect inferences concerning independent test taking, histograms have been produced with score values of 8, 15, 22, 29, 36, and 43. These are shown in Figures 7-A through 7-F, below. These values were selected in order to provide a visualization of the similarity distributions across the entire space. Because of symmetry and folding of the distributions, only the upper triangle needs to be evaluated. The histogram titles list the score of Test Taker 1 separated by the score of Test Taker 2 with an “x” (e.g. 15x22 indicates scores of 15 and 22). In the plots below, the “Counted” data provide the relative frequencies from the live data and the “Simulated” data provide the relative frequencies from the simulated data. In order to compactly represent the data, three and four plots are provided per page. These pages could be extracted and arranged into a tableau in order to compare the histograms as functions of the total test scores.

Figure 7-A: Histograms for 8x8, 8x15, and 15x15

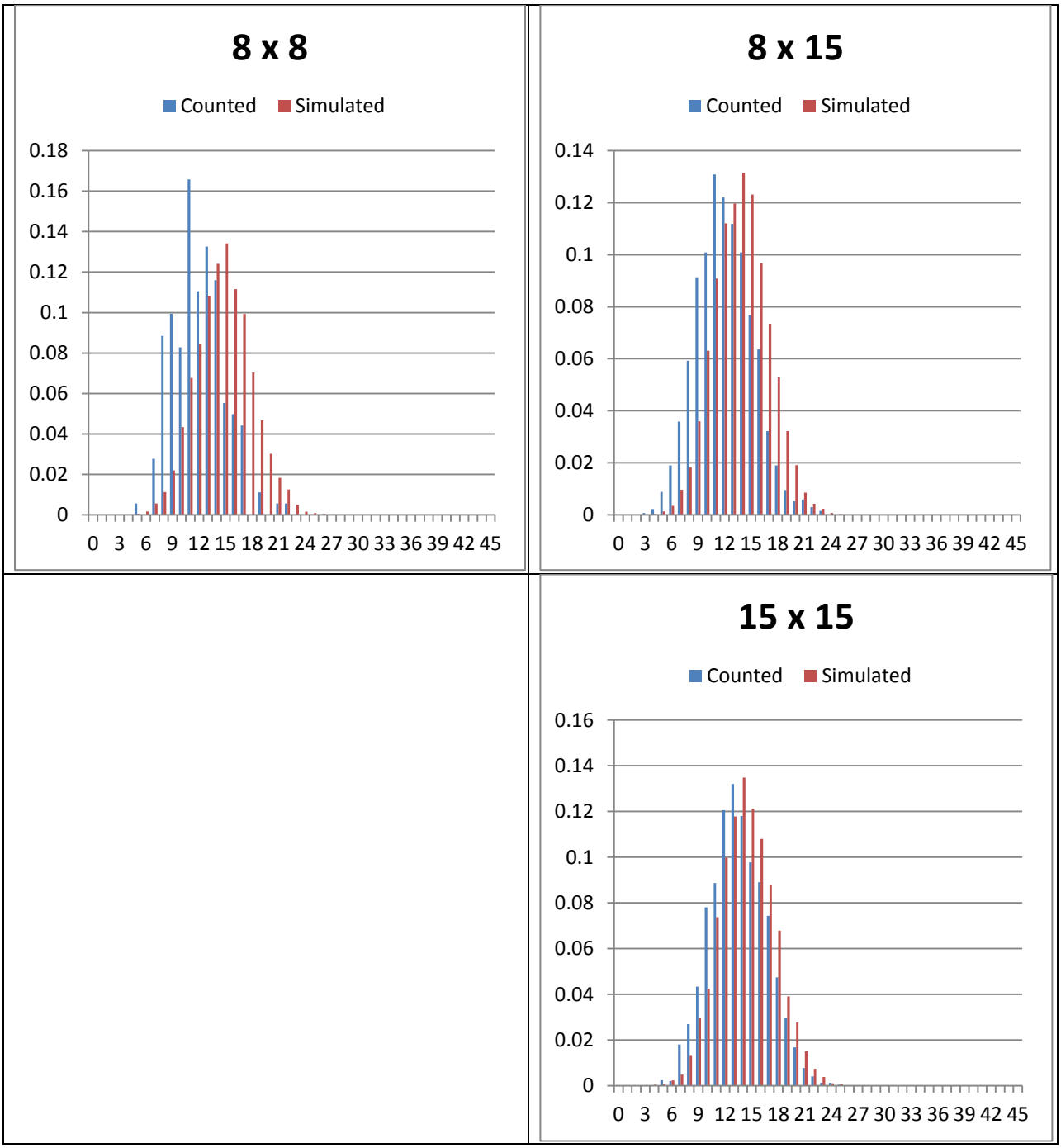


Figure 7-B: Histograms for 8x22, 8x29, 15x22, and 15x29

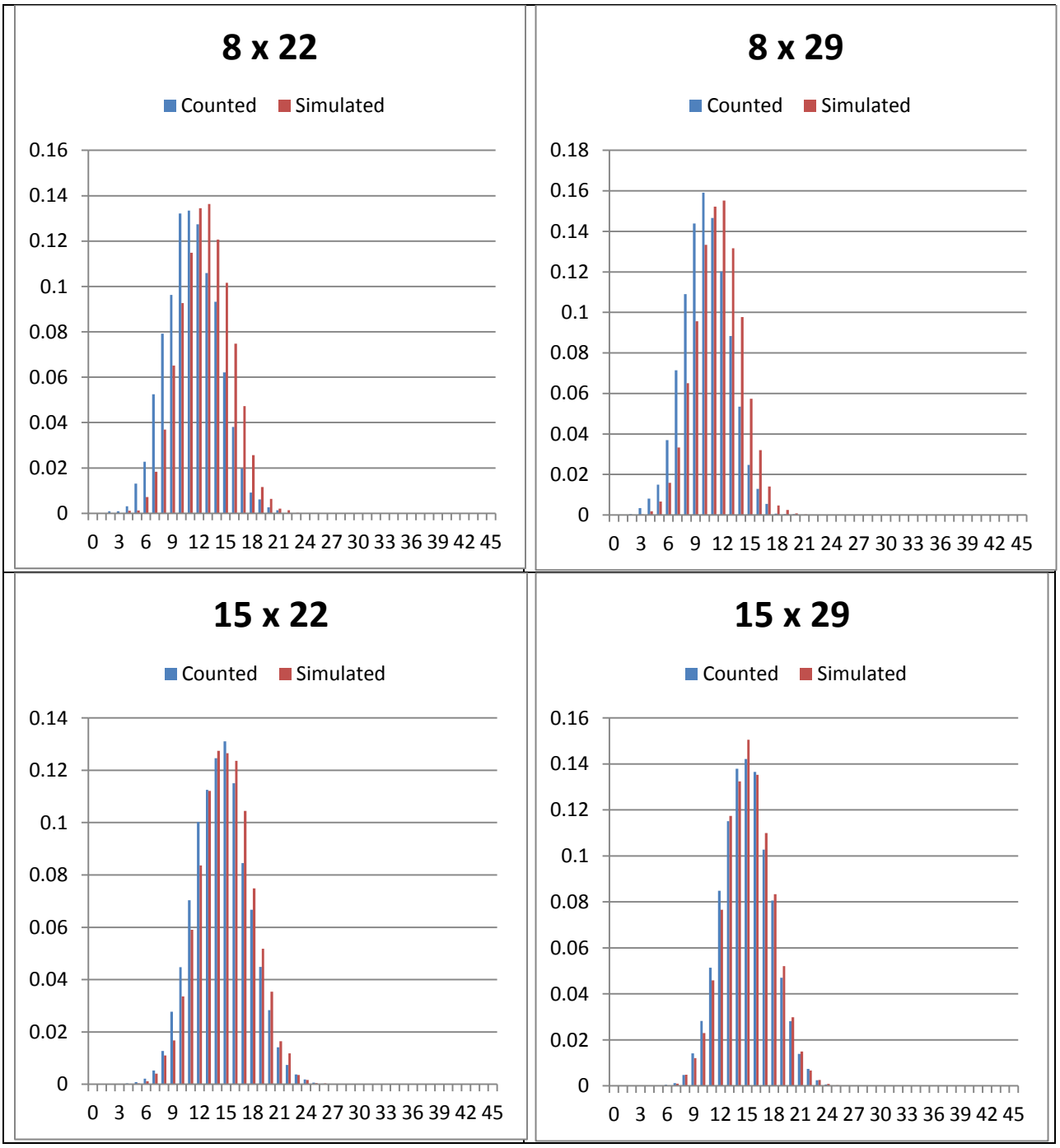


Figure 7-C: Histograms for 8x36, 8x43, 15x36, and 15x43

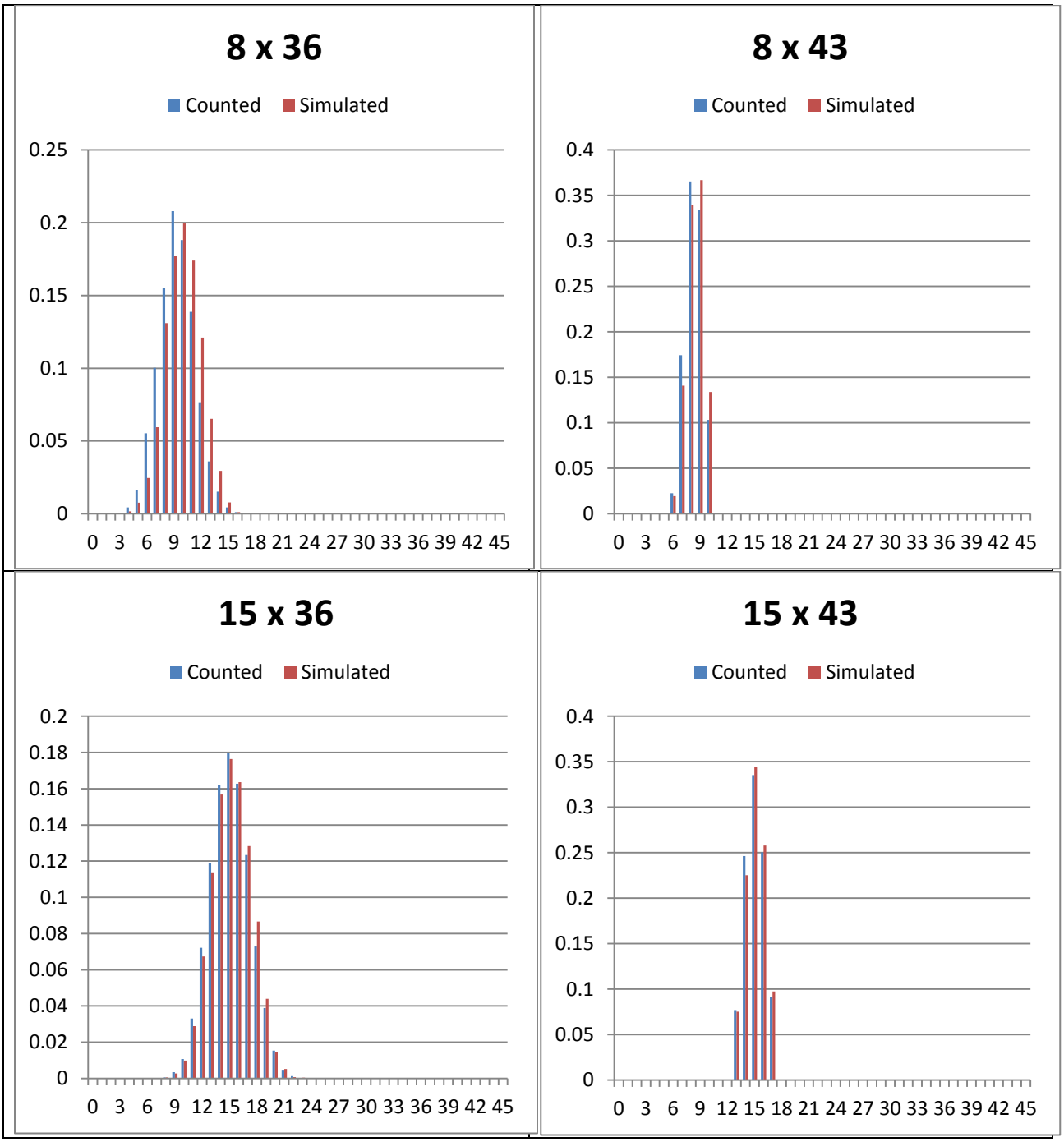


Figure 7-D: Histograms for 22x22, 22x29, and 29x29

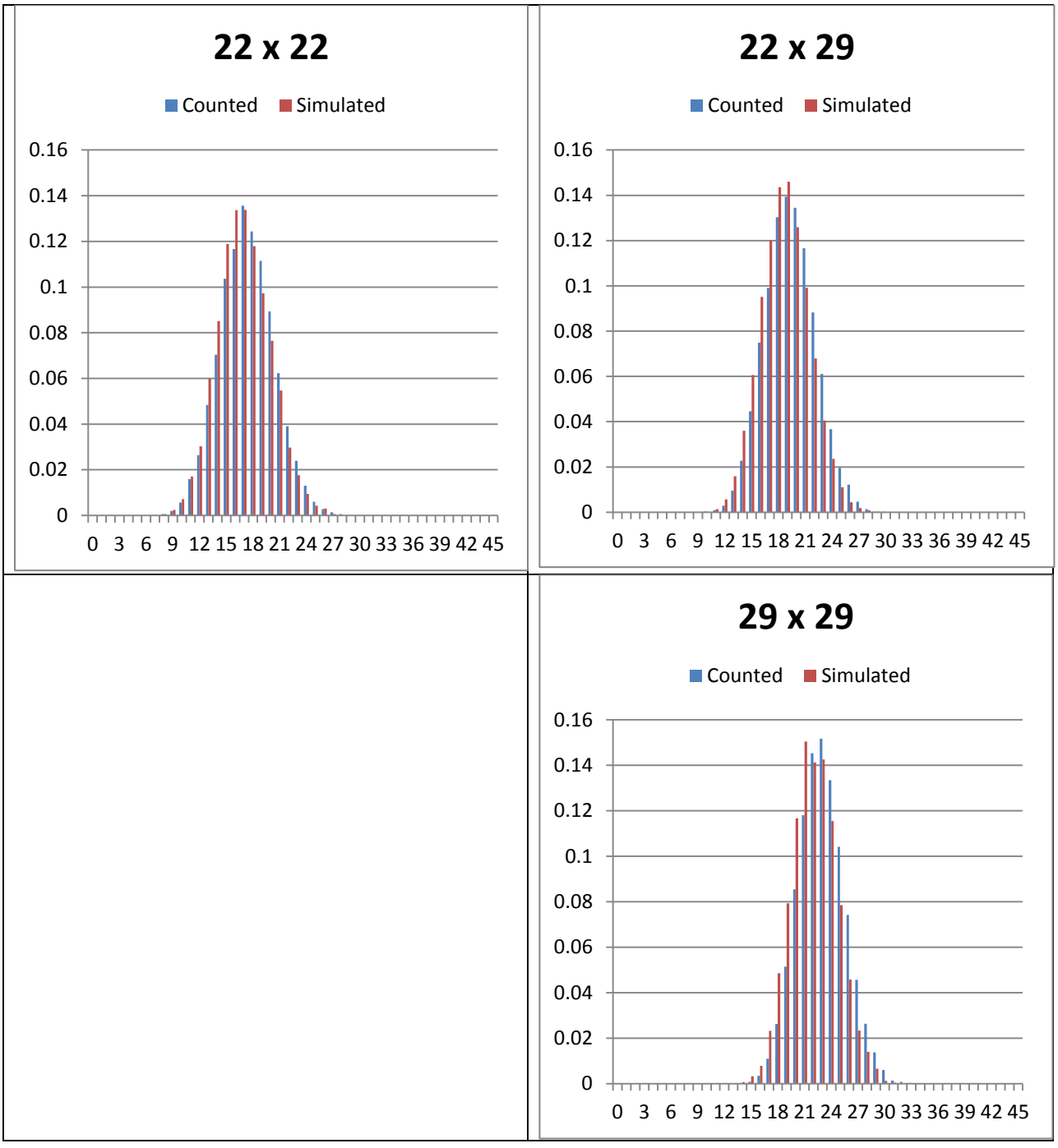


Figure 7-E: Histograms for 22x36, 22x43, 29x36, and 29x43

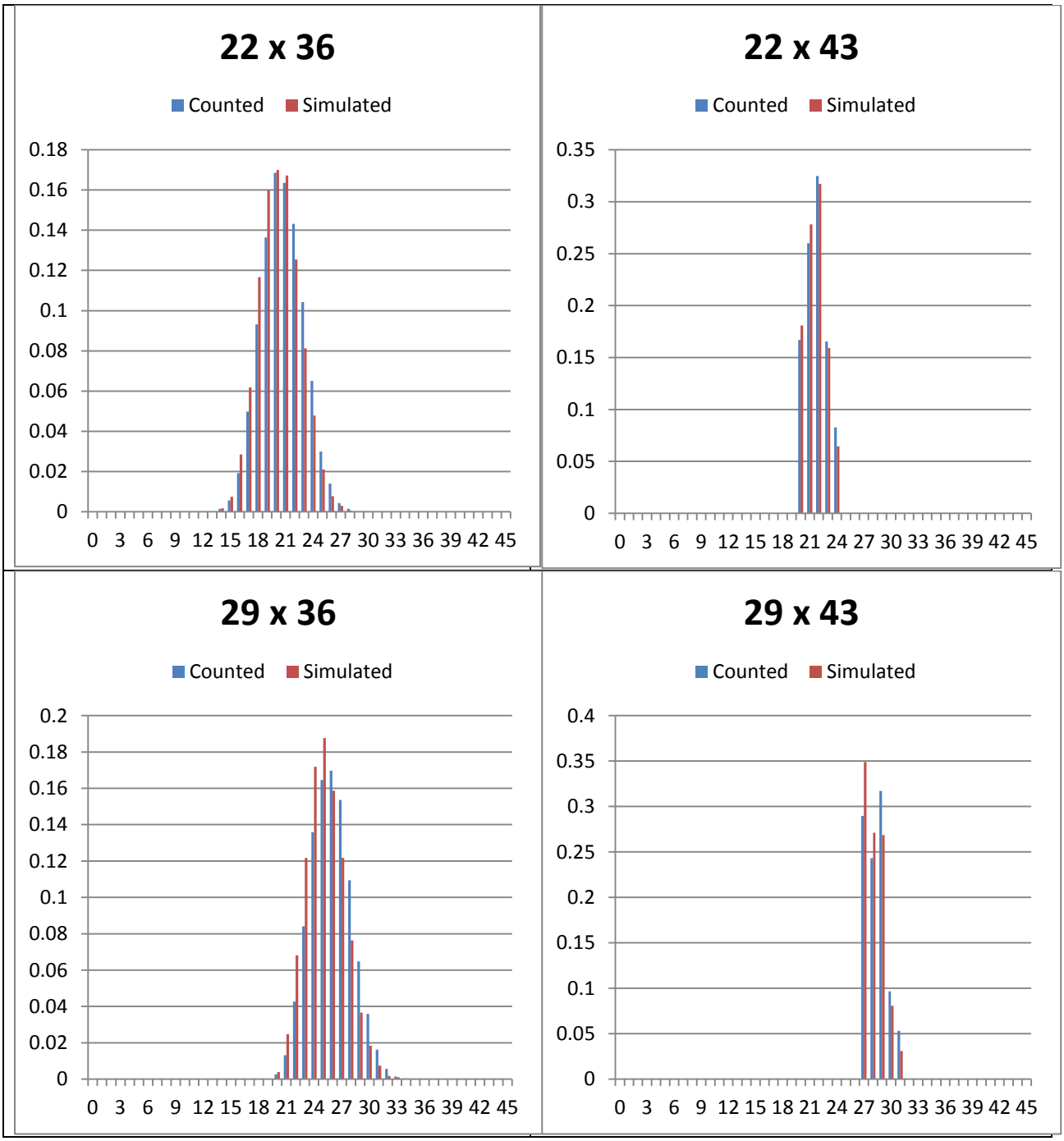
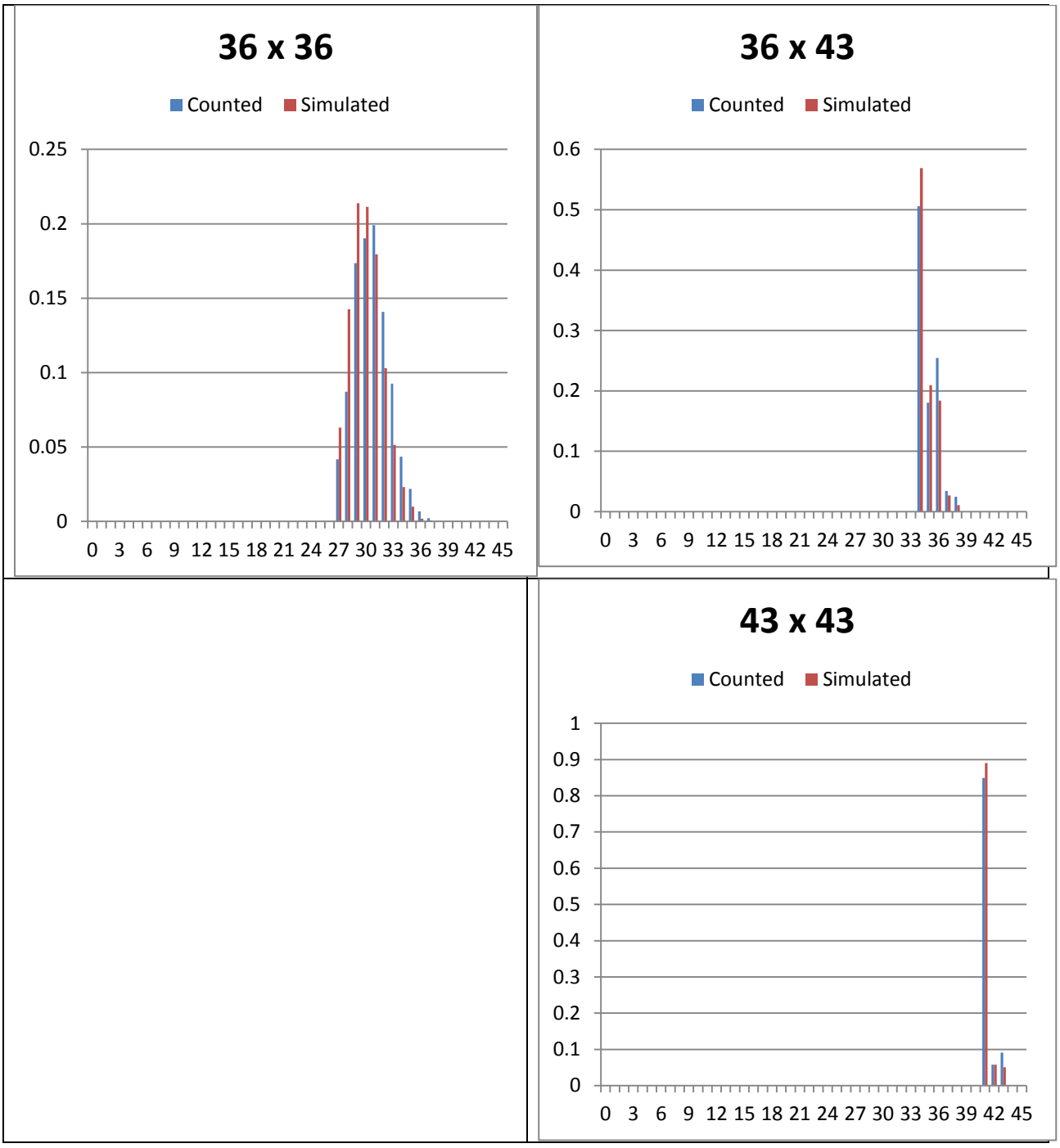


Figure 7-F: Histograms for 36x36, 36x43, and 36x43



Conditional similarity distributions estimated from the live data (labeled "Counted") and the simulated data using the model (labeled "Simulated") are shown in Figures 7-A through 7-F. It should be apparent that assumptions of normality will not hold for all test scores. Except for score values below the guessing

level, Probability Model 2 provides a very close fit to the empirical data. This is evidenced by the high level of concurrency between the blue and red bars.

Summary and Conclusions

There are always nuances that need to be evaluated when actual cases of potential answer copying are considered. Models can provide guidance on where to search for potential misbehaviors and the magnitude of the potential problems. Models cannot identify why two tests were not taken independently or how the misbehavior was accomplished. Those questions must be answered in other ways, such as by interviewing test takers and inspecting seating charts.

This paper shows that taking into account test taker performance and item difficulty are essential for estimating the probability of similar responses. Many of the models that have been suggested in the literature do not incorporate both of these elements. It also shows that use of continuous distributions (e.g. the normal distribution or the beta probability density function) to approximate the actual distribution of the number of identical responses between two test response vectors may not be appropriate across the entire score range.

This paper shows that the region of permissible values constrains the distributions of the numbers of identical responses when the responses from two tests are compared. Most researchers have neglected this constraint on the data. A topic for future research is the impact on approximating distributions when the region of permissible values is ignored. Logic would suggest that variances of those distributions would be larger than the variances of the actual distributions. If so, accuracy of probability estimation would improve by incorporating this information into the model.

In this paper. a model was presented (Probability Model 2) which uses two person parameters: performance for Test Takers 1 and 2, and three item parameters: item difficulty, and conditional proportions of identical correct and identical incorrect answers. Because item discrimination is not used, the probability model for a correct answer is very similar to the Rasch model. Probability Model 2 is similar to the model implemented by Wesolowsky (2000) with the following exceptions:

- Instead of using a logit transformation for computing the probability of a correct response, Wesolowsky's model used a function "suggested by ℓ_p distance from location theory."
- Wesolowsky proposed using a normal distribution for computing the probabilities of similar responses instead of a discrete, constrained distribution as proposed by Probability Model 2.

The model assumes that the conditional proportion of identical incorrect answers is constant after conditioning for item difficulty, and the performance of Test Takers 1 and 2. Removing this assumption would require an extension to the model (e.g. use the nominal response model – Bock, 1972). The analysis in this paper suggests that this would be an appropriate area for future research.

References

Angoff, W. H. (1974). The Development of Statistical Indices for Detecting Cheaters. Journal of the American Statistical Association, 69(345), 44-49.

Belleza, F. S. & Belleza, S. F. (1989). Detection of Cheating on Multiple-Choice Tests by Using Error-Similarity Analysis. Teaching of Psychology, 16(3), 151-155.

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

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.

Frary, R. B. (1993). Statistical Detection of Multiple-Choice Answer Copying: Review and Commentary. Applied Measurement in Education, 6(2), 153-165.

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

Holland, P. W. (1996). Assessing Unusual Agreement Between the Incorrect Answers of Two Examinees Using the K-Index. Educational Testing Service. Technical Report No. 96-4.

Van der Linden, W. J. & Sotaridona, L. (2004). A Statistical Test for Detecting Answer Copying on Multiple-Choice Tests. Journal of Educational Measurement, 41(4), 361-377.

Van der Linden, W. J. & Sotaridona, L. (2006). Detecting Answer Copying When the Regular Response Process Follows a Known Response Model. Journal of Educational and Behavioral Statistics, 31(3), 283-304.

Wesolowsky, G. O. (2000). Detecting 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.