McMaster University

MICHAEL G. DEGROOTE SCHOOL OF BUSINESS

RESEARCH AND WORKING PAPER SERIES

DETECTING EXCESSIVE SIMILARITY IN ANSWERS ON MULTIPLE CHOICE EXAMS

By

George O. Wesolowsky

Michael G. DeGroote School of Business
McMaster University
Hamilton, Ontario

Working Paper # 442

October, 1999

STER UNIVERSITY
in Street West
ı, Ontario, Canada L8S 4M4
ıe (905) 525-9140

15 October 1999

Detecting excessive similarity in answers on multiple choice exams

George O. Wesolowsky, Michael G. De Groote School of Business, McMaster University, Hamilton, Ontario, Canada

ABSTRACT This paper provides a simple and robust method for detecting cheating. Unlike some methods, non-cheating behaviour and not cheating behaviour is modelled because this requires the fewest assumptions. The main concern is the prevention of false accusations. The model is suitable for screening large classes and the results are simple to interpret. Simulation and the Bonferroni inequality are used to prevent false accusation due to 'data dredging'. The model has received considerable application in practice and has been verified through the adjacent seating method.

1 Introduction

Multiple choice examinations are a fact of academic life, as is the suspicion that some students cheat in various ways, particularly by copying from each other. The latter method of cheating is susceptible to detection by statistical means and many such methods have been discovered and re-discovered in the literature. Attitudes towards statistical detection of cheating vary from labelling this one of "the success stories of educational measurement" (Frary and Tideman (1997)) to strong reservations (Dwyer and Hecht (1996)). The existence of these methods of detecting copying appears to be very little known among most instructors using multiple choice tests and examinations. The majority of instructors also take few precautions, such as using multiple versions of examinations, against copying.

Even when statistical detection methods are known, they are often pointedly ignored. The issue of prosecuting cheating creates fear of unpleasant and time-consuming confrontations and legal entanglements, arouses ideological positions, and raises unwelcome publicity for educational institutions. There is also the problem that many who become involved in such cases tend to view statistical methods with suspicion. They fear that these methods will lead to accusations of the innocent 'just by chance'. Further doubt is caused because the common claim of accused students that their multiple choice exam answers are similar because they studied together, seems plausible. In fact, these are concerns which must be addressed when statistical detection is used. However, the reluctance to deal with individual cases often also leads to the rejection of a much more important role of such detection methods, namely the estimation of the extent of the copying problem. This is unfortunate because this is one form of academic dishonesty which can readily be controlled.

Although there were earlier papers, the literature on the statistical detection of copying on multiple choice tests began in earnest in the 1970's. Papers with a broad review of the literature include Hanson, Harris and Brennan (1987), Frary (1993) and Post (1996).

The method described here does not attempt to model any particular cheating behaviour. Other works in the literature, for example Link and Day (1992), Frary et al. (1977), and Kvam (1996) model cheating behavior by making assumptions in the alternative hypothesis; for example, who cheated from whom, and how. It is certainly possible to design a more powerful test this way. Our test simply looks at the number of matching answers and ignores other suspicious patterns such as groupings or sequences in answer matches. One reason for this is to make as few assumptions as possible, especially about copying behavior, which seems to take many forms. Another reason is to keep the method understandable and easily applicable. The emphasis is on preventing false accusations (controlling Type I error) and not on increasing the number of detections. This number has been substantial enough.

Most of the methods in the literature adopt basically similar models. Probabilities of correct and incorrect responses on multiple choice questions are estimated from actual class responses. An important concern is how to approximate the probability that a student will answer correctly on a given question. It seems reasonable to assume that this depends on the student's overall score on the examination, as well as on the difficulty of the particular question. Some early (Saupe (1960)), and even some fairly recent papers, however, did not incorporate the student's ability into the model. One approach to incorporating this consideration is to divide the class into strata so that it can be assumed that the students in each stratum are of approximately equal ability (Harpp and Hogan (1993), (1996)). An alternative approach is found in the seminal paper by Frary et al. (1977) which modelled the probability of a correct answer using the student's overall score as well as the ratio of correct responses on the question. Our paper suggests a modification of the Frary et al. approach. Our model has a more intuitively appealing form, and exhibits more consistency with respect to certain other constraints.

There are two basic ways of using a statistical detection program. One is to provide evidence if there is some prior reason to suspect particular students, for example an invigilator's report on suspicious behaviour during an examination. This is a standard hypothesis test. The other use is scanning, that is comparing the responses of all pairs of students in an attempt to detect copying for which there was no previous evidence. Some authors have missed the point that for the latter case the evidence must be much stronger; not realizing this may lead to false accusations (Chaffin (1979)). One common approach to dealing with this difficulty is the Bonferroni inequality. In addition to incorporating this inequality, we use simulation to demonstrate that the detection cutoffs suggested have a considerable safety margin with respect to Type I error (false accusation).

2 The model

2.1 Definitions

Let:

n = number of students in the class,

q = number of questions on the exam,

m_{jki} = the probability of a match between students j and k on question i,

M_{jk} = the number of matches (same answers) observed between students j and k,

p_{ji} = probability of student j being correct on question i,

r_i = proportion of the class that answered correctly on question i,

c_j = proportion of questions answered correctly by student j,

v_i = the number of wrong choices on question i,

w_{ti} = the probability that, given the answer is wrong, wrong choice t is chosen on question i .

Note that $\sum_{t=1}^{v_i} w_{ti} = 1$, and w_{ti} is assumed, for simplicity, to be the same for all students. It is possible to have w_{ti} depend on student performance. For example, a different w_{ti} could be estimated for each quartile of c_j.

2.2 Calculating m_{jki}

As illustrated in Fig. 1, the probability of a match between students j and k on question i, namely m_{jki}, is equal to the probability that they both have the correct answer plus the probability that they both have the wrong answer and their wrong answers match:

$$m_{jki} = p_{ji}\, p_{ki} + (1 - p_{ji})\,(1 - p_{ki}) \sum_{t=1}^{v_i} w_{ti}^{\;2} \qquad j = 1,\ldots,n\text{-}1;\;\; k = j+1,\ldots,n;\; i = 1,..q\,. \qquad 1)$$

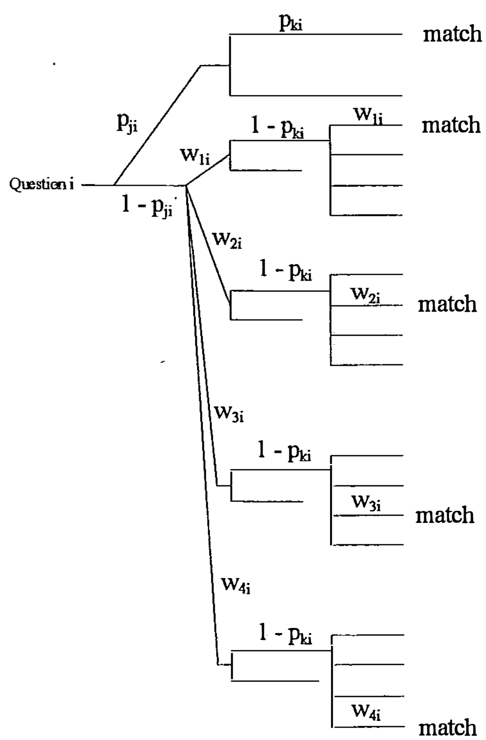

Fig. 1. The probability of a match between students j and k on question i

2.3 Estimating p_{ji}

As discussed in the introduction, it is reasonable to assume that:

1) Students who have a higher score overall are more likely to answer a question correctly than those who have a lower score overall.

2) The probability that a student has an answer right depends on the difficulty of the question; this difficulty is reflected in the proportion of the class that answers the question incorrectly.

These principles can be used to estimate p_{ij} as a function of r_i and c_j. Our approach is similar to that used by Frary et al. (1977), except that their piecewise function is replaced by a smooth one suggested by l_p distance iso-contours from location theory (see Love, Morris, and Wesolowsky (1988), page 258). The Frary et al.model, in diagram form, is given in Fig. 2. It has anomalies at $r_i = 0$ and $r_i = 1$.

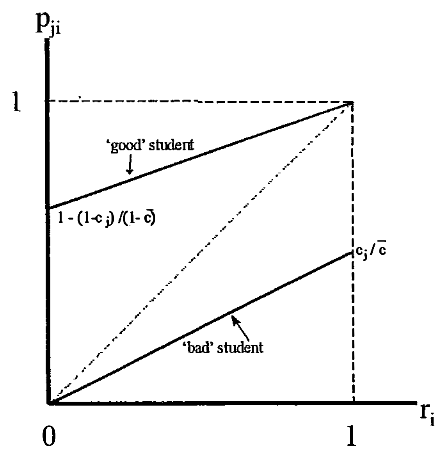

Fig. 2: The Frary et al. model of p_{ji}

In our approach, $\hat{p}_{ji}$ is a function of r_i and c_j in the following way:

$$\hat{p}_{ji} = (1-(1-r_i)^{a_j})^{1/a_j} \qquad j = 1,..,n;\ i = 1,\ldots q, \qquad 2)$$

where:

$$\frac{\sum_{i=1}^{q} \hat{p}_{ji}}{q} = c_j \qquad j = 1,\ldots,n. \qquad 3)$$

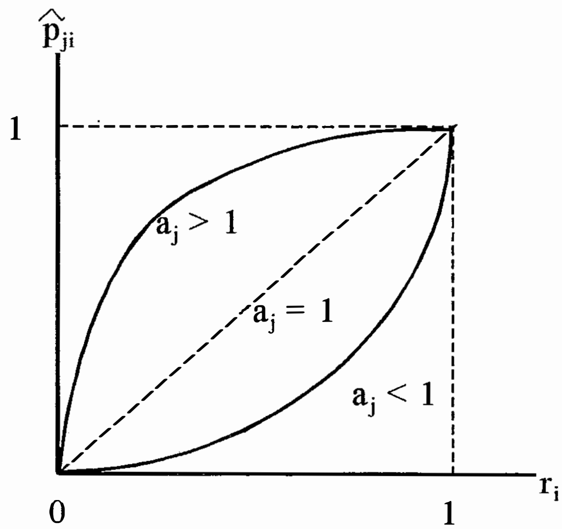

Fig. 3: Expression 2)

This is illustrated in Fig. 3.

The parameter a_j is found by solving $\frac{\sum_{i=1}^{q} \hat{p}_{ji}}{q} = c_j$ by bisection search. The constant a_j is thus found for each student so that the average of the estimated probabilities of correct responses on each question is equal to the proportion of all questions answered correctly by that student. This condition is also met by the Frary et al. model.

Lemma: if $c_j = \bar{r}$ then $\hat{p}_{ji} = r_i$.

Proof:

$$\frac{\sum_{i=1}^{n} \hat{p}_{ji}}{n} = c_j = \bar{r} = \frac{\sum_{i=1}^{n} r_i}{n}$$

This is true if $\hat{p}_{ji} = r_i$ ❑

Therefore, the curve in Fig. 2 is a straight line for $c_j = \bar{r}$. Note $\bar{r} = \bar{c}$.

If $c_j > \bar{r}$, the curve is concave because the student has higher probabilities than average on each question. If $c_j < \bar{r}$ then the curve is convex and the student's probabilities are below average. However, in addition to 3) there is another condition that would be desirable. This condition requires that the estimates for the probabilities of correct responses by students ($\hat{p}_{ji}$) be consistent with the proportion of correct responses on each question by the class:

$$\frac{\sum_{j=1}^{n} \hat{p}_{ji}}{n} \approx r_i \qquad i = 1,..,q. \qquad 4)$$

It turns out that 4) is closely met by $\hat{p}_{ji}$ in actual class data. When the Frary et al. model is used these equations hold approximately true only for r_i near $\bar{r}$. Our model, therefore, has more 'structural consistency'. The example in Fig. 4 gives the difference between the left and right sides of 4) for our model and for the Frary et al. model for a class with 348 students and 50 questions. The class average was .67; note that both models meet 4) closely for r_i near this average.

However, for difficult (low r_i) and easy (high r_i) questions, the Frary et al. model gives implied values of r_i that are too high and too low respectively. This means that the estimated probabilities of correct responses for such questions are respectively too high and too low on the average. While some instructors attempt to make all questions of the same difficulty, others mix easy and hard questions, and even give different weights to different questions when marking. Our model is better suited to cope with this. Also, the fact that real classes comply closely with 4) is an implicit confirmation that response probabilities are being modeled in a reasonable manner.

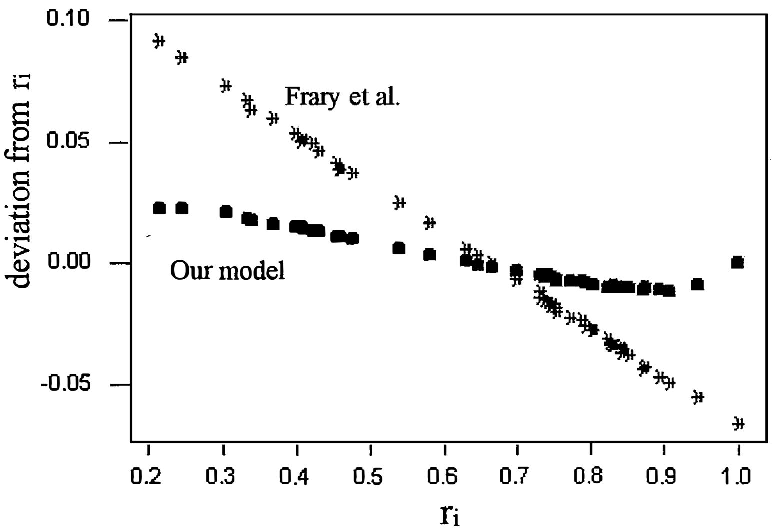

Fig. 4: Comparing compliance with 4)

3 The probability distribution of the number of matches

In order to determine if M_{jk}, the number of matches (identically answered questions) for two students j and k, is suspiciously high, we need to estimate the probability that that number of matches or more could occur. Basically, this is the significance in a one tailed test of the null hypothesis that there is no cheating. For each pair of students j and k the probability of a match on question i is estimated by:

$$\hat{m}_{jki} = \hat{p}_{ji}\,\hat{p}_{ki} + (1-\hat{p}_{ji})(1-\hat{p}_{ki})\sum_{t=1}^{v_i}\hat{w}_{ti}^{\,2} \qquad j = 1,\ldots,n\text{-}1;\; k = j\text{+}1,\ldots,n;\; i = 1,..q \qquad 5)$$

For each pair of students j and k we have q Bernoulli trials with different probabilities. This has been called the repeated independent trials model (page 31 of Sveshnikov (1968)). Let the variable $\tilde{M}_{jk}$ be the number of matches between students j and k. It has the following probability generating function:

$$P(\tilde{M}_{jk} = M_{jk}) = \frac{1}{M_{jk}!}\left(\frac{\partial^{M_{jk}} G(u)}{\partial u^{M_{jk}}}\right)_{u=0} \qquad 6)$$

where:

$$G(u) = \prod_{i=1}^{q}(m_{jki}u + (1 - m_{jki}))$$

This probability distribution is sometimes known as the compound binomial. Equation 6) is not a practical method for computing the estimated probability distribution of $\tilde{M}_{jk}$. This can be better done by a recursive method which is sometimes used in the literature. It is given in the appendix for completeness. However, this recursion, although easy to describe, is computationally slow. It should be recalled that we wish to look at all possible pairs of students and this number is n(n-1)/2. Slowness in computations for each pair is therefore inconvenient in large classes. It is faster to compute these probabilities by the normal approximation, which is the usual approach. Our algorithm uses both methods without sacrificing accuracy.

4 Normal approximation to the distribution of $\widetilde{M}_{jk}$

Since the generation of matches between two students can be viewed as q Bernoulli processes operating in parallel, the mean of the number of matches is simply the sum of the means of these processes, and the variance is the sum of the individual variances.

Therefore, the estimated mean of the number of matches between students j and k is:

$$\hat{\mu}_{jk} = \sum_{i=1}^{q} (\hat{p}_{ji} \quad \hat{p}_{ki} + (1-\hat{p}_{ji})(1-\hat{p}_{ki}) \sum_{t=1}^{v_i} \hat{w}_{ti}^2) \quad j = 1,\ldots,n-1; \; k = j+1,\ldots,n, \qquad 7)$$

and the estimated variance of the number of matches is:

$$\hat{\sigma}_{jk}^2 = \sum_{i=1}^{q} (h_{jki})(1-h_{jki}) \quad j = 1,\ldots,n-1; \; k = j+1,\ldots,n, \qquad 8)$$

where:

$$h_{jki} = \hat{p}_{ji} \quad \hat{p}_{ki} + (1-\hat{p}_{ji})(1-\hat{p}_{ki}) \sum_{t=1}^{v_i} \hat{w}_{ti}^2 .$$

The approximate Z value (standardized normal variate) is:

$$Z_{jk} = \frac{M_{jk} - 1/2 - \hat{\mu}_{jk}}{\hat{\sigma}_{jk}} \qquad j = 1,\ldots,n-1; \; k = j+1,\ldots,n. \qquad 9)$$

5 The significance of the number of matches

Let $\gamma(M_{jk}) = P(\widetilde{M}_{jk} \geq M_{jk})$, which can be obtained from equation 6) by summing terms. This is the significance in testing the hypothesis that there is no copying, against the alternative that there is copying. Let $\lambda(M_{jk})$ be an approximation for $\gamma(M_{jk})$ that is obtained by substituting $\hat{m}_{jki}$ for m_{jki} in G(u). Let $\varphi(Z_{jk}) = P(Z \geq Z_{jk})$. $\varphi(Z_{jk})$ is a less accurate approximation for the significance of the number of matches between students j and k, but is faster to compute than $\lambda(M_{jk})$.

An important difficulty is missed in many articles in the literature. The number of matches that is suspicious depends on whether the pair is question was suspected prior to the running the program or whether the pair was identified entirely on the basis of a high M_{jk} or Z_{jk}. Intuitively, if a student pair is suspected prior to the running of the program (say, for behaviour during the test), then a Z_{jk} of 3 or more is strong evidence because this is an 'a priori' hypothesis test. However, it should be noted that it is almost certain that there will be some Z_{jk} values over 3 in a class of, say, 400 even when there is no cheating whatsoever. This occurs because in a class of 400 there will be many (79,800 to be exact) pairs scrutinized. Essentially this means that pairs that have not been pre-selected in some reasonable way must be judged by a higher standard of evidence. This can be accomplished as follows.

There are $N = \dfrac{n^2-1}{2}$ pairs being examined. Suppose that we choose a critical value Z_c, for Z_{jk}. P(at least one $Z_{jk} \geq Z_c$) $\leq N\,\varphi(Z_{jk}) = \Phi(Z_{jk})$ by the well known Bonferroni inequality. Similarly, we can define $\Lambda(M_{jk}) = N\,\lambda(M_{jk})$.

Suppose that we have the rule that students will be flagged if $\Lambda(M_{jk}) \leq \Omega$ or, using the normal approximation, if $\Phi(Z_{jk}) \leq \Omega$. Then Ω will be the upper limit (approximately derived) on the probability that there will one or more pairs falsely accused in a class.

Some comments should be made about N. This program was developed in response to problems on centrally administered examinations. Instructors were not routinely given seating plans; these were available, after the exam was written, on request only. The program therefore conservatively makes no use of the seating plans. It is interesting to look into this further.

Seating of classes was done in columns, with students from different classes on both sides of the column. For simplicity assume that all students are seated in a single continuous column. To be charged with copying, students flagged by the program must be found to be seated adjacently (one behind the other). If the program randomly selects a pair, then the probability that they will be found seated together is 2/n. The bound on the significance of this joint event is therefore $(2/n)(n(n-1)/2)\lambda(M_{jk}) = (n-1)\,\lambda(M_{jk})$.

Suppose now that we obtain the seating arrangement for the column prior to running the program and only compare the (n-1) pairs that are seated adjacently. Note that the program still requires response information for the entire class to calculate its estimates. In this case, however, $N = n-1$, and the bound on the significance of the number of matches for each pair is $(n-1)\,\lambda(M_{jk})$, which is exactly the same as before. However, examining the $(n^2 - n)/2 - (n-1) = (n-1)(n-2)/2$ pairs, who presumably could not have cheated, is a very powerful check on the validity of the model, as will be discussed subsequently.

For more complex seating arrangements, the analysis can be difficult. Consider, however, a rectangular array with f rows and g columns. If adjacent seating means horizontal, vertical, or diagonal adjacency, then there are $(4fg - 3(f+g) + 2)$ adjacent pairs. For example, a 20x20 array of

400 students would have 1482 adjacent pairs out of 79,800 possible pairs. A square array will have more adjacent pairs than any rectangular array of the same size.

6 Simulation experiments

There are two important questions regarding the probability of 'false accusations' as estimated in the preceding section. The first question is what will be the effect of the fact that $\lambda(M_{jk})$, $\varphi(Z_{jk})$, $\Lambda(M_{jk})$, and $\Phi(Z_{jk})$ use not actual probabilities but probabilities which are estimated from the class data? The second is that if the Bonferroni bound is conservative, what are the actual probabilities of false accusations that will occur for a pre-set level of Ω? Both of these questions are very difficult to answer analytically but can be answered by simulation.

The simulation procedure is very simple and is based on the question answering model in Figure 1. The estimated probabilities in Figure 1 result in a series of simple Bernoulli simulations to create a set of simulated responses for each student on each question. The simulated class is then subjected to the algorithm as if it were an actual class. Thus an actual class provides the 'true' probabilities and generates its own series of simulated classes.

Simulations were generated using four real classes A, B, C and D. The subject matter was economics, statistics, information systems, and mathematics. Simulations were done using the normal approximation to calculate Ω, instead of the more accurate compound binomial distribution. This was done simply to reduce computation time. The classes had up to 412 students and 50 questions. To reduce computation time only part of the classes was used. Some methods in the literature would consider 20 students and 20 questions to be too small a data set. This small size was included to show that it does not lead to false accusations.

Table 1: Simulation of the safety factor

Class	No. of simulated classes	No. of students	No. of questions	Ω using normal approximation	Proportion of classes with false accusations	Safety factor
A	10000	20	20	1.0	.1197	8.4
A	10000	20	20	.05	.0034	14.7
A	50000	20	20	.01	.00044	22.7
A	10000	100	30	1.0	.1691	5.9
A	10000	100	30	.05	.0063	7.9
A	100000	100	30	.01	.00152	6.6
B	10000	20	20	1.0	.0661	15.1
B	10000	20	20	.05	.0012	41.7
B	10000	100	30	1.0	.073	13.7
B	10000	100	30	.05	.0017	29.4
C	10000	20	20	1.0	.0603.	16.6
C	10000	20	20	.05	.0008	62.5
C	10000	100	30	1.0	.0285	35.1
C	10000	100	30	.05	.0006	83.3
D	10000	20	20	1.0	.0407	24.6
D	10000	20	20	.05	.0009	55.6
D	10000	100	30	1.0	.0759	13.2
D	10000	100	30	.05	.0028	17.9

As is seen in all of the simulations, the actual proportion of false positives was well below the permitted level Ω. This means there is a safety factor created by the conservative nature of the Bonferroni inequality which more than compensates for the fact that the model probabilities were estimated from the data. This "safety factor" is given in Table 1 as Ω divided by the proportion of false accusations in the simulation. Each row of the table is a different set of simulated classes. Note that for any data set, the safety factor tends to increase as Ω decreases. Note also that classes B, C, and D have mainly similar results but class A has generally lower safety factors. Class A has a higher average standard deviation of Z's in the simulations than the other classes. One reason may be that the average grade in that exam was just over 50%, while the other exams had average grades over 60%.

One could use simulation on every class to obtain more accurate estimates of Type I error; the only reason for not doing this is that it is time consuming.

7 Is the model appropriate?

The most common plea of students accused of copying (by far) is that "our answers are similar because we studied together". This claim may seem to conveniently explain why the students were found to be in adjacent seating: friends want to be together. The model assumes independence in student responses and hence this apparently plausible claim has to be considered. However, it does not take much study of the diagram in Figure 1 to see that other assumptions could also be questioned, either in general, or in the particular structure of certain tests.

Let us consider the "studied together" claim in particular. First, we must recognise that programs such as this one evaluate a very large number of pairs of students. Note that a class of 100 will have 4950 pairs (not all independent) and recall that a class of 400 will have 79,800 pairs. This particular program has scrutinized some 40 or more classes using a setting of $\Omega = 1$. One case was reported, at a significance close to the cutoff, where the students were found to have written the exam in different rooms. All of the rest of the student pairs flagged were found to be in adjacent seating. The proportion of classes with flagged non-adjacent students seems entirely consistent with what would be predicted for false accusations by the simulations of the preceding section.

As discussed earlier, there are generally very limited opportunities for adjacent seating, and a large number of students work together. The "studied together" effect, or other model imperfections, should have produced large numbers of strongly similar pairs who were not in adjacent seating. It seems quite implausible that model violations are only restricted to adjacently seated pairs. It thus seems implausible that the 'working together' effect is strong enough to cause false accusations except perhaps in tests of very unusual design.

This experience is augmented by the fact that it has often been noted in the literature that such programs are very consistent in what they consider to be strong similarity; differences are usually in marginal cases. Our computational experience with the Harpp and Hogan (1996) method bears this out. We can therefore add their experience (David Harpp, personal communication, March 25, 1998), which includes hundreds of classes in many disciplines, to ours. This experience is that strong similarity is always accompanied by adjacent seating, and that when precautions such as multiple versions and randomized seating are taken, such strong similarities disappear.

This indicates that the model is very robust with respect to violations of its assumptions. It also indicates that the "studied together" defence is plausible only on the surface. One should not, of course, use the program with no thought as to whether the examination or test was constructed in a manner compatible with the model.

8 Operation of the algorithm

The algorithm examines all N possible pairings of students and selects those pairs that have a similarity above the given threshold: that is $\Lambda(M_{jk}) \leq \Omega$, or alternatively $\Phi(Z_{jk}) \leq \Omega$. Because the normal approximation is faster to compute, it is used for the bulk of the sorting. It is only for pairs close to the $\Phi(Z_{jk}) \leq \Omega$ criterion that the switch is made to the $\Lambda(M_{jk}) \leq \Omega$ criterion. Generally, the normal approximation is accurate and on the conservative side. It is important to note that the accused pairs are always judged on the more accurate criterion.

Table 1 gives the output of a 'run' on a class of 348 students who were given 50 multiple choice questions. The first 20 questions were true-false, and the remainder 5 choice. The dots mean that answers were correct, the numbers indicate which incorrect answers were chosen. For example '2' would mean that choice 'b' was made and that this choice was incorrect.

The default level of Ω was set at 1.0. Because the Bonferroni inequality is conservative this actually means that this level would result in a 'false accusation' very roughly once in 10 or 20 classes. This is borne out by experience and simulation experiments. This threshold results in a very conservative estimate of the level of copying in the class but would not, of course, be normally used for charging students unless there is additional evidence.

A few comments on the interpretation of the output in Table 2 are appropriate. BVP_p = $\lambda(M_{jk}) = 6.97\ 10^{-12}$ for students number 53 and 195. The 'BVP' stands for Bernoulli with variable probabilities. Zb = 6.76 is not from the normal approximation: it was derived by solving $P(Z \geq Zb) = 6.97\ 10^{-12}$ in a standardized normal distribution. This equivalent Z is provided to give a well-known index for intuitive comparison. BBVP_p= 4.21E-07 = $\Lambda(M_{jk})$ and this means that the number of false accusations would be fewer than 4.21 in 10 million classes on the average. As mentioned previously, the actual number would be much less. This degree of similarity is not at all unusual. The output H-Hstat = 6.00 gives the Harpp-Hogan similarity statistic (Harpp and Hogan (1996)). It should be noted that when the seating plan was checked after this computer run, all six pairs were found to be in adjacent seating.

Table 2: Example of output

```
'course' in course.txt  ? xxx  ||  #stud. =  348 #ques= 50
No. of students ?          No. of questions ?
Generate .nam file?
Input y to include pre-specified suspect pairs?
Start checking at question  ?
The symbol for a correct answer is (default is .)    ?
Maximum z value (default= 8)  ?
zincrement (.1->.5,  default= .1)  ?
Bonferroni significance (default =1 )  ?
Z Bonferroni= 4.150725
____________________________________________________________________

** pair: 14   80 ************  H-Hstat =   1.44 *************
Zb = 4.26  BVP_p= 1.03E-05  BBVP_p= 6.21E-01 m= 41 | 50 (mu,s)=( 27.11  3.27)
.22.1..... ..21.2.1.2 .42......5 .3..2....1 .3..45.1.1
.22.1.21.. ......21.2 .........5 .3..2....1 .2..45.1.1
** pair: 48  188 ************  H-Hstat =   2.13 *************
Zb = 5.31  BVP_p= 5.51E-08  BBVP_p= 3.32E-03 m= 42 | 50 (mu,s)=( 24.42  3.33)
.2.2...11. 1.21..211. ..32..2.3. .3.5.114.5 .....2.1.2
.2.2...11. 1.21..211. ..322.2.3. .3...4..35 .....2...1
** pair: 49  126 ************  H-Hstat =   1.86 *************
Zb = 4.75  BVP_p= 1.04E-06  BBVP_p= 6.25E-02 m= 43 | 50 (mu,s)=( 27.91  3.24)
.2....211. ...122.... ......1... ...3.12.15 .5...2.1..
.2...12.1. .2.122.... ....1.2... ..33.12..5 .5...2.1..
** pair: 53  195 ************  H-Hstat =   6.00 *************
Zb = 6.76  BVP_p= 6.97E-12  BBVP_p= 4.21E-07 m= 47 | 50 (mu,s)=( 25.96  3.30)
.2....211. 1.2....1.. ...21.2... ...3412.15 .354.2.4..
.2....211. 1.2....... ....1.2... ...3412.15 .3.4.2.4..
** pair: 82  241 ************  H-Hstat =   1.88 *************
Zb = 5.04  BVP_p= 2.58E-07  BBVP_p= 1.56E-02 m= 42 | 50 (mu,s)=( 25.54  3.31)
.2...1..1. 1.2...2... ..1.5.24.. .3.3.21544 .2..31.54.
.2........ 1.2...2.... ..1...1... .3.3.21544 .2..31.44.
** pair: 173  205 ************  H-Hstat =   2.00 *************
Zb = 4.65  BVP_p= 1.67E-06  BBVP_p= 1.01E-01 m= 44 | 50 (mu,s)=( 29.66  3.17)
........1. .2...2.1.. .41......5 .332..5... .....1.2.1
...2...111 .2..2..... .41......5 .332..5... .....1.2.1


 347     mean= -0.0809 stdev= 0.8528
```

The program has several features that are useful. It is possible, for example, to run this program in two modes: one which identifies the selected pairs and one which does not. It is possible, therefore, to do anonymous studies. The program can also force the inclusion of pairs that are not included by the choice of Ω. This is done, for example, when examination invigilators report suspicious behaviour by a pair. The program also permits the exclusion of questions from the beginning and the end. This permits identification of exam versions and exclusion of some questions. The program also provides output and graphics for checking the internal assumptions of the model.

9 Some comments on application

There are two basic uses for the program: to support charges of academic dishonesty, and to investigate the security of examination writing conditions. Obviously, the first use is more controversial because it raises the fear that students may be unjustly accused, either by chance, or because the model assumptions are faulty. These are valid concerns but should be viewed objectively.

The claim that students could be accused just by chance (even though the probability may be somewhat infinitesimal) is true. There is a tendency by some to divide evidence into the categories 'real' and 'statistical'. The testimony of witnesses, for example, would be in the real category. One could counter by saying that all forms of evidence can lead to unfair accusations and hence this distinction is an illusion; there is ample evidence for this in our legal systems. Such arguments, however, enter into the murky domain of legal and philosophical issues, and are not within the scope of this paper.

Our model assumptions are violated in real applications (classes) to varying degrees. This would be true of any model and is true of virtually every statistical technique. However, there is a very large body of indirect evidence that unless (possibly) these violations are extreme, they do not lead to false accusations.

Nevertheless, there is one rather compelling, if indirect, reason why such programs should not be used to support charges against students except in unusual circumstances. The reason is that this type of cheating is almost entirely preventable by measures such as randomized seating and multiple versions of exams. When academic dishonesty charges are made, this is an admission that effective, if unfortunately sometimes expensive, security measures were not taken..

It is difficult to think of a noble reason not to use a program such as this to scan classes to determine if there is a cheating problem. While one might not wish to make individual accusations for various reasons, it is difficult to argue that the method (or others like it) does not provide a conservative estimate of the extent of cheating. Scanning can even be done anonymously by a third party without revealing the identities of the students 'flagged'.

It is important to note that the degree of copying revealed by the program will be an underestimate. Although the method does use the information from correct answers and not just the wrong answers (as some do), detection is more effective when there is a larger proportion of wrong answers. In the extreme, when a student copies another's 100% paper completely, this is not detectable by any purely statistical. means.

The Bonferroni cutoffs used are very conservative, and although simulation could obtain more accurate values, this is very time consuming. Finally, this method does not hypothesize any particular patterns of cheating, as some methods do, and hence does not have an advantage in detecting them.

REFERENCES

Chaffin, W.W. (1979) Dangers in using the Z Index for detection of cheating on tests, *Psychological Reports*, 45, pp. 776-778.

Dwyer, D.J. & Hecht, J.B. (1996) Using statistics to catch cheaters: methodological and legal issues for Student Personnel Administrators, *NASPA Journal,* 33, 2, pp. 125-35.

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

Frary, R.B., (1993) Statistical detection of multiple-choice answer copying: review and commentary, *Applied Measurement in Education,* 6, 2, pp.153-65.

Frary R.B., & Tideman, T.N. (1997) Comparison of two indices of answer copying and development of a spliced index, *Educational and Psychological Measurement,* 57, 1, pp. 20-32.

Hansen B.A , Harris D.J. , & Brennan, R.L. (1987) A comparison of several statistical methods for examining allegations of copying, ACT Research Research Report Series 87-15). Iowa City, IA: American College Testing Program.

Harpp D.N. & Hogan, J.J. (1993) Crime in the classroom – detection and prevention of cheating on multiple-choice exams", *Journal of Chemical Education* 70, 4, pp.306-311.

Harpp D.N. & Hogan, J.J. (1996) Crime in the classroom – Part II, an update *Journal of Chemical Education*, 73, 4, pp.349-351.

Hynes, K., Givner, N.& Patil K. (1978) Detection of Test Cheating Behavior, *Psychological Reports,* 42, pp.1070.

Kvam, P.H. (1996) Using exam scores to estimate the prevalence of classroom cheating, *The American Statistician*, 50,3, pp. 238-242.

Link, S.W. & Day, R.B. (1992) A Theory of Cheating, *Behavior Research Methods*, 24, 2, pp. 311-316.

Love, R.F., Morris, J. & Wesolowsky, G., (1988) *Facilities Location: Models and Methods*, Elsevier Publishing, New York, New York .

Post, G.V. (1996) A quantal choice model for the detection of copying on multiple choice exams, *Decision Sciences*, 25, 1, pp.123-142.

Saupe J.L. (1960) An empirical model for the corroboration of suspected cheating on multiple-choice tests, *Educational and Psychological Measurement,* XX, 3, pp.475-489.

Sveshnikov A.A.(1968) *Problems in Probability Theory, Mathematical Statistics and Theory of Random Functions*, Dover

Appendix

Consider n independent Bernoulli processes. The probability of success in experiment i is m_i for i = 1,...n. What is the probability of x successes?

Let f(q,j) be the probability of j successes in the first q trials. We wish to find f(n,x).

$f(q, j) = f(q-1, j-1)\, m_q + f(q-1, j)\, (1- m_q)$.

The probability of j successes in the first q trials is equal to the probability of j-1 successes in the first q-1 trials and success in trial q, plus the probability of j successes in the first q-1 trials, and no success in trial q .

To apply this recursion one notes that $f(1, 1) = m_1$, $f(1,0) = 1 - m_1$ and that $f(q, -1) = 0$ for all q.

Faculty of Business
McMaster University

WORKING PAPERS - RECENT RELEASES

415. N. Archer, "Characterizing World Wide Web Search Strategies", April, 1996.

416. J. Rose, "Immediacy and Saliency in Remedying Employer Opposition to Union Organizing Campaigns", July, 1996.

417. Roy J. Adams and Parbudyal Singh, "Worker Rights Under NAFTA: Experience With the North American Agreement on Labor Cooperation", September, 1996.

418. George Steiner and Paul Stephenson, "Subset-Restricted Interchange for Dynamic Min-Max Scheduling Problems", September, 1996.

419. Robert F. Love and Halit Uster, "Comparison of the Properties and the Performance of the Criteria Used to Evaluate the Accuracy of Distance Predicting Functions", November, 1996.

420. Harish C. Jain and Simon Taggar, "The Status of Employment Equity in Canada", December, 1996.

421. Harish C. Jain and Parbudyal Singh, "Beyond The Rhetoric: An Assessment of the Political Arguments and Legal Principles on Strike Replacement Laws in North America", January, 1997.

422. Jason Schwandt, "Electronic Data Interchange: An Overview of Its Origins, Status, and Future", March, 1997.

423. Christopher K. Bart with John C. Tabone, "Mission Statement Rationales and Organizational Alignment in the Not-for-profit Healthcare Sector", November, 1997.

424. Harish C. Jain, Michael Piczak, Işik Urla Zeytinoğlu, "Workplace Substance Testing - An Exploratory Study", November, 1997.

425. S. Suarga, Yufei Yuan, Joseph B. Rose, and Norman Archer, "Web-based Collective Bargaining Support System: A Valid Process Support Tool for Remote Negotiation", January, 1998.

426. Pawan S. Budhwar and Harish C. Jain, "Evaluating Levels of Strategic Integration and Development of Human Resource Management in Britain", March, 1998.

427. Halit Üster and Robert F. Love, "Application of Weighted Sums of Order p to Distance Estimation", April, 1998.

428. Halit Üster and Robert F. Love, "On the Directional Bias of the ℓ_{bp}-norm", April, 1998.

429. Milena Head, Norm Archer, and Yufei Yuan, "MEMOS: A World Wide Web Navigation Aid", October, 1998.

430. Harish C. Jain and Parbudyal Singh, "The Effects of the Use of Strike Replacement Workers on Strike Duration in Canada", February, 1999.

431. Parbudyal Singh and Harish C. Jain, "Strike Replacements in the United States, Canada and Mexico: A Review of the Law and Empirical Research", February, 1999.

432. John W. Medcof and Jeremy Boyko, "Reinforcing, Revising and Reconciling Attributions in the Employment Interview", March, 1999.

433. Norm Archer, "World Wide Web Business Catalogs in Business-to-Business Procurement", March, 1999.

434. Diwakar Gupta and Saifallah Benjaafar, "Make-to-order, Make-to-stock, or Delay Product Differentiation? - A Common Framework for Modeling and Analysis", April, 1999.

435. Harish C. Jain, Parbudyal Singh and Carol Agocs, "Recruitment, Selection and Promotion of Visible Minorities and Aboriginals in Selected Canadian Police Services", April, 1999.

436. Harish C. Jain and Angus Bowmaker-Falconer, "Employment Equity/Affirmative Action Codes of Practice and Best Practices in USA, Britain, Canada and Other Selected Countries", May, 1999.

437. Diwakar Gupta, Yavuz Günalay, and Mandyam M. Srinivasan, "On the Relationship Between Preventive Maintenance and Manufacturing System Performance", June, 1999.

438. Jinliang Cheng, George Steiner, and Paul Stephenson, "A Fast Algorithm to Minimize Makespan for the Two-Machine Flow-Shop Problem with Release Times", June, 1999.

439. Jinliang Cheng, George Steiner, and Paul Stephenson, "A Fast Algorithm to Minimize Maximum Lateness for the Two-Machine Flow-Shop Problem with Release Times, June, 1999.

440. Norm Archer and Yufei Yuan, "Electronic Commerce and the Business-to-Business Customer Relationship Life Cycle", August, 1999.

441. Halit Üster and Robert F. Love, "Calculation of Confidence Intervals for Estimated Distances", August, 1999.