Statistical Methods for the DETECTION OF ANSWER COPYING on Achievement Tests

Leonardo S. Sotaridona

Samenstelling promotiecommissie

Voorzitter/secretaris	Prof.dr. J.M. Pieters
Promotor	Prof.dr. W.J. van der Linden
Assistant Promotor	Dr. R.R. Meijer
Referent	Dr. P.W. Holland (Univ. of California at Berkeley/ Educational Testing Service)
Leden	Prof.dr. C.A.W. Glas (Univ. Twente GW)
	Dr. C.W.A.M. Aarts (Univ. Twente BBT)
	Prof.dr. K. Sijtsma (Univ. van Tilburg)
	Prof.dr. P.F. Sanders (Cito)
	Dr. H.J.M. van Berkel (Univ. Maastricht)

Publisher: Twente University Press, P.O. Box 217, 7500 AE Enschede, the Netherlands, www.tup.utwente.nl

Cover design: Geronimo S. Sotaridona

Print: Océ Facility Services, Enschede

ISBN 903651942x

STATISTICAL METHODS FOR THE DETECTION OF ANSWER COPYING ON ACHIEVEMENT TESTS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof.dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 5 september 2003 te 15:00 uur

door

Leonardo Sitchirita Sotaridona

geboren op 11 april 1968
te Sta. Barbara, Filippijnen

This thesis has been approved by

the promotor prof.dr. W.J. van der Linden

the assistant promotor dr. R.R. Meijer

Acknowledgment

I have very much appreciated the opportunity to conduct my Ph.D. research at the Department of Research Methodology, Measurement, and Data Analysis at the University of Twente. The atmosphere there has been very friendly, supportive, and conducive to my research. I thank all of my colleagues for their support; it has been a pleasure working with them.

I am grateful to Professor Wim J. van der Linden, my thesis supervisor, for his very reliable and highly competent guidance. Working with him during these years has been a great pleasure. His fatherly guidance and recommendations for my further career have been equally important to me.

Special thanks are due to Dr. Rob R. Meijer, assistant supervisor. Our cooperation over the past three-four years "has not been characterized by any misfit". He gave me much space to pursue my research ideas and provided me with necessary and timely assistance when the tasks were difficult. Rob's enthusiasm in the various research projects has resulted in a high output. At many times and in a variety of cases, he demonstrated his willingness to lend me his helping hand. It has been a privilege to have him as my daily supervisor.

Furthermore, I would like to thank Professor Cees A. W. Glas, Department Head, for his comments and suggestions on my research and for his quick response to all questions related to my research. His fatherly advice on the direction of my future career is also appreciated.

I consider Bernard and Jean Paul as "big brothers", always ready to spend time with me on queries of any sorts, ranging from research-related topics to personal and family-related concerns. It was great to have them around especially during my initial adjustment in Enschede.

I thank IOPS for organizing its regular (inter)national conferences and for the financial support received to attend the 2001 International Meeting of the Psychometric Society in Osaka, Japan. Likewise, I thank Educational Testing Service (ETS) for the Harold Gulliksen Psychometric Fellowship awarded to me as well as for the opportunity to visit its campus as a summer intern in June-July, 2002. These experiences have been very helpful in getting a good perspective on my future career.

Thanks also due to my fellow Ph.D. students—Irene, Anna, Jonald, and Adelaide—for the casual exchange of ideas with them that were often nec-

essary to get closer to the solution of a problem. Special thanks are due to Irene, my office mate; our discussions and cooperation have proven to be mutually beneficial.

I have very much appreciated the support of Dr. Renato V. Alba and friends at the Western Visayas College of Science and Technology in Iloilo City, Philippines and also the support of the Filipino-Dutch family friends in the Netherlands.

I would like to thank my parents, brothers and sisters for their endless spiritual support and words of encouragement.

I humbly dedicate this work to GingGing, KZ, and Sean.

Finally, I would like to thank the Almighty Father, the source of all wisdom.

Leonardo S. Sotaridona
Enschede, September 2003

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Testing, or "assessment," plays a vital role in education today. Test results are often a major force in shaping public perception about the quality of our schools. As a primary tool of educators and policy makers, assessment is used for a multitude of purposes. Educators use assessment results to help improve teaching and learning and to evaluate programs and schools. Assessment is also used to generate the data on which policy decisions are made. Because of its important role, educational assessment is a foundation activity in every school, every school district and every state–a vital component in innovation, higher standards and educational excellence. (A Guide to Effective Assessment, Online Resources: CTB/McGraw-Hill).

It is important that the test scores are reliable, accurate, and valid to be able to make sound inferences about a person's knowledge, skills, or ability. There are several factors that may invalidate a test score. One factor is cheating on the test. Cheating violates academic integrity and is a threat to educational quality.

Because cheating on tests will invalidate the accuracy of inferences made about the knowledge or competence measured (Cizek, 1999; Meijer, 1998), it is a serious problem that need serious attention from a number of key players, such as students, parents, educators, testing agencies, and policy makers.

There are a variety of ways the problem of cheating can be addressed, for example, by developing the right attitude among the students, by prevention, or by detection. Although prevention and developing the right attitude among the students may reduce cheating considerably, detection of cheating

is important to warrant fair test taking.

Among the many cheating methods employed are acquisition without permission of tests or other academic materials and/or distribution of these materials and other academic work. In this thesis, the focus is mainly on answer copying, that is the situation where one examinee copies the answers from another examinee.

To detect answer copying on multiple-choice tests, both observational and statistical methods can be used. Observational methods use a human observer to detect answer copying such as talking to another examinee during the test or collecting physical evidence such as cheat sheets exchanged between two examinees. Statistical methods address cheating by modeling the response probabilities of examinees under the assumption of no cheating and looking for patterns of similar answers between examinees that are unlikely under the model.

It is important to note that statistical evidence should not be used as the only evidence for accusing an examinee of cheating but that it is supported by other types of evidence. Interesting in this respect is a remark in a case in the United States

> "Statistics are not irrefutable; they come in infinitive variety and, like any other kind of evidence, they may be rebutted. In short, their usefulness depends on all of the surrounding facts and circumstances (*Teamster v. U.S.*: Good, 2001, p. 13)."

Surrounding facts and circumstances in a testing situation may be observations of cheating or very large differences between an examinee's test score in two administration of a similar test. Thus, often some kind of trigger should precede the use of statistical methods to minimize false accusation of cheating behavior.

1.1 An overview of the Thesis

This thesis contains a collection of studies where statistical methods for the detection of answer copying on achievement tests in multiple-choice format are proposed and investigated. Although all methods are suited to detect answer copying, each method is designed to address specific characteristics of a testing situation.

Chapter 2 presents a simulation study to investigate the statistical properties of the K-index (Holland, 1996; Lewis & Thayer, 1998). The K-index is used as a copying index at the Educational Testing Service, the largest

testing company in the United States. In this chapter the performance of this index is compared to an index that is based on item response theory. Furthermore, small sample properties of the K-index are investigated and some modifications to the K-index are proposed.

Building on the insights gained from the study in Chapter 2, two new modifications of the K-index are proposed in Chapter 3. The first modification uses a Poisson distribution to model the number of similar wrong responses of the source and the copier instead of the binomial distribution that is used in the original version of the K-index. The second modification uses the number of similar incorrect and weighted correct responses of the source and the copier. The main idea behind these indices is to incorporate the additional information about copying that is contained in the similar correct responses. Instead of using a uniform weight of one for all similar correct responses, a weight is used that depends on the probability of a correct response. The Type I error and detection rates of these statistics are compared to that of other statistics.

In both Chapters 4 and 5 new statistical tests are presented under a null distribution that is independent of the behavior of any other examinees than the examinee suspected of copying and the examinee believed to have served as a source. The basic difference between the two approaches is the way the response process is modelled and the type of conditioning employed.

The statistical test of answer copying presented in Chapter 4 is based on the response process model described as "knowledge-copying-or-random-guessing model". This model assumes that the answers of examinees to test items may be the result of three possible processes: (1) knowing, (2) guessing, and (3) copying, but that examinees who do not have access to the answers of other examinees can arrive at their answers only through the first two processes. This assumption leads to a distribution for the number of matched incorrect alternatives between the examinee suspected of copying and the examinee believed to be the source that belongs to a family of "shifted binomials". It is shown that the test is uniformly most powerful (UMP) at the level of significance chosen. The test is compared to the K-index and several power functions are presented for several sets of parameter values.

The statistical test of answer copying presented in Chapter 5 is based on Cohen's Kappa (Cohen, 1960). The research was motivated by the question how much could be inferred about copying behavior of examinees without assuming any specific response model. The only assumption in this chapter is that the response behavior of the copier and the source is probabilistic, that is, can be characterized by a (possibly different) probability distribution

over the alternatives of each item. Simulation studies are conducted to investigate (1) the impact of different response probabilities across the items on the Type I error of the test, (2) the effects of recoding the alternatives of the items before pooling their information in a table, and (3) the power of test.

A conceptually new approach to screening possible cheaters on a high-stakes test is presented in Chapter 6. This chapter gives an overview of the basic principles of neural networks and discusses how this technique can be applied to identify cheaters. The results of a small simulation study shows that this technique has high power when the configuration of the item scores of a cheater has similar characteristics as the configuration used to train the network.

Chapter 2

Statistical Properties of the K-Index

Abstract. We investigated the statistical properties of the K-index (Holland, 1996) that can be used to detect copying behavior on a test. A simulation study was conducted to investigate the applicability of the K-index for small, medium, and large datasets. Furthermore, the Type I error rate and the detection rate of this index were compared with the copying index, ω (Wollack, 1997). Several approximations were used to calculate the K-index. Results showed that all approximations were able to hold the Type I error rates below the nominal level. Results further showed that using ω resulted in higher detection rates than the K-indices for small and medium sample sizes (100 and 500 simulees).

2.1 Introduction

The variety of methods to cheat on educational tests seems to be only restricted to one's imagination. In his book on cheating on tests, Cizek (1999, chap. 3) gives an overview of several cheating methods. Among the methods discussed are using forbidden materials, circumventing the testing process, or even using microrecorders.

This chapter has been published as: Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39, 115-132.

In the present study, we will be concerned with a form of cheating that has received some attention in the recent literature, namely, answer copying. In this type of cheating, one examinee copies the answers from another examinee. This copying may take place from an examinee who is sitting in the neighborhood of the copier, although answer copying may also take place using all kinds of code for transmitting answers; a code for doing so, for example, may be the clicking of pens, tapping of the foot, and the like. Thus the examinees do not have to be physically near each other. Because answer copying may invalidate an examinee's test score (Meijer, 1998), it is necessary to prevent those practices by employing well-instructed proctors and construct the seating arrangements so that there is ample room between the examinees. However, if a proctor observes some irregularities, statistical methods may be used to obtain additional evidence of answer copying.

Two types of method have been proposed to detect answer copying: person-fit statistics and answer-copying statistics. Person-fit statistics compare the likelihood of an observed item score pattern with the likelihood under a test model (Meijer & Sijtsma, 2001; Reise, 2000), whereas answer-copying statistics determine the probability that the observed score patterns of two examinees under suspicion are similar. We will focus on answer-copying statistics because they have higher power for detecting answer copying than person-fit statistics (Cizek, 1999, pp. 217-218). Answer-copying statistics can be classified into two types (Cizek, 1999, pp. 138-139). One type of method compares an observed pattern of responses to a known theoretical distribution (e.g., Bay, 1994; Frary, Tideman, & Watts, 1977; Wollack, 1997). In the second type of method, the probability of an observed pattern is compared with a distribution of values derived from independent pairs of examinees who took the same test. An example of such a statistic is the K-index (Holland, 1996).

In this article we investigated the statistical properties of the K-index, in terms of Type I error and detection rates, which thus far is only described in a paper by Holland (1996) and applied on a few empirical datasets from Educational Testing Service (ETS). A modification of the K-index was described in Lewis and Thayer (1998). As Cizek (1999) noted, no comparative studies of the performance of this index are known, so it is unknown whether it performs better, worse, or the same as the other available indices. In this article we investigated the small sample properties of the K-index, in particular. Furthermore, we compared the detection rate of this index with the index, ω, proposed by Wollack (1997). The major difference between the indices is that the K-index does not assume any test model, whereas ω is based on item response theory modeling (e.g., van der Linden & Hambleton,

1997).

This study is organized as follows. First, we discussed the rationale behind the K-index and discussed several methods proposed by Holland (1996) to calculate this index. Second, we discussed some existing practical problems when this index is applied in practice and proposed two new methods to calculate this index. Third, we conducted a simulation study to investigate the statistical properties of this index and finally, we conducted a simulation study in which we compared the Type I error rate and detection rate of the K-index with the ω statistic.

2.2 The K-Index

The K-index is a statistic that can be used to assess the degree of unusual agreement between the incorrect answers on a multiple-choice test of two examinees; one referred to as the *source* (s) and the other as the *copier* (c). The copier is suspected of copying answers from the source. Note that the K-index only takes the incorrect answers of the examinees into account. For a rationale behind this strategy, see Holland (1996).

2.2.1 Notation

The following notations will be used throughout this chapter. Let j ($j = 1, \ldots, J$) denote examinees; i ($i = 1, \ldots, I$) denote items; v ($v = 1, \ldots, V$) denote the item response categories; s denote an examinee identified as the source; c denote an examinee suspected of copying answers from s; w_j denote the number of "wrong" answers of examinee j, and M with realization m denote the number of matching wrong answers between examinee j and s. Furthermore, let $r = 1, \ldots, c', \ldots, R$ denote subgroups of examinees, where each group has a distinct number of wrong answers and c' is the group where examinee c belongs; $j' = 1, \ldots, n_r$ denote an examinee in subgroup r, where each subgroup has at least one examinee and $\sum_{r=1}^{R} n_r = J - 1$; $\mathbf{M}_r = (M_{r1}, \ldots, M_{rj'}, \ldots, M_{rn_r})$ denote a vector of the number of matching wrong answers in a particular subgroup r; $\mathbf{M}_{c'} = (M_{c'1}, \ldots, M_{c'n_{c'}})$ denote a vector of the number of matching wrong answers of $n_{c'}$ examinees in subgroup c' where subgroup c' consists of examinees having the same number-incorrect score as the copier; and let $Q_r = \frac{w_r}{I}$ denote the proportion of wrong answers of subgroup r where I is the total number of items in the test.

2.2.2 K-Index Based on the Empirical Distribution

The K-index can be determined using empirical data of J examinees answering I items. To calculate the K-index based on the empirical data, we first determine the group of examinees with the same number-incorrect score as c (subgroup c') and then for each of these examinees in subgroup c' we determine the number of items that match the incorrect answers of the source. This is the vector $\mathbf{M}_{c'}$ and the distribution of $\mathbf{M}_{c'}$ comprises the empirical agreement distribution. For examinee c, we specifically denote $m_{c'c}$ as the number of matching wrong answers between c and s. The random variable M_{rj} will simply be denoted as M if it is not necessary to identify the group membership of j. The K-index is defined as the proportion of examinees having the *same* number-incorrect score as c and whose number of matching incorrect item scores with s is at least as large as $m_{c'c}$.

For $j' = 1, \ldots, n_{c'}$, let $I_{c'j'}$ denote an indicator variable, coded as 1 for $m_{c'j'} \geq m_{c'c}$, and 0 otherwise, then K is defined as

$$K = \frac{\sum_{j'=1}^{n_{c'}} I_{c'j'}}{n_{c'}}. \tag{2.1}$$

The idea is that when K is very small there is statistical evidence that examinee c copied from examinee s.

Note that, in general, the number of matching incorrect scores depends on the ability level of s and c. The number of matching incorrect answers is necessarily small when either s or c or both have many correct scores (high ability), whereas it is large when both examinees have many wrong answers (low ability). In order to minimize the dependency of M on the ability level of the population of examinees, the K-index is computed conditional on the number of incorrect scores of the suspected copier. As a consequence, the number of examinees involved in the actual computation of the K-index (subgroup c') becomes very small. We emphasize this because the number of examinees in subgroup c' influences the accuracy of the value of the K-index. When the sample size is small ($J = 100$), one alternative is to use a theoretical approximation to the empirical agreement distribution.

2.2.3 K-Index Based on Theoretical Approximations

To use the K-index, one has to specify first the Type I error (α) which is defined as the probability of misclassifying an examinee as a copier. Ideally, we would like to have a statistic for which the nominal and empirical Type I

error rates are similar. Note that in this type of statistical application, the main concern is to have a statistic that is not liberal—a statistic for which the empirical Type I error rate is at most as large as the nominal Type I error rate—because the consequence of misclassifying an honest examinee as a copier can be very serious at the individual level.

Seaman, Levin, and Serlin (1991; see also Wollack, 1997), stressed that copying indices that fail to hold the nominal Type I error rate should be considered unacceptable. On the other hand, a copying index should not be overly conservative; otherwise, its power to detect true examinee copiers will be very low.

In general, a disadvantage of using the discrete empirical distribution in small samples is that the random variable M can only take a small number of values. As a result, it is often not possible to obtain a prespecified Type I error of say .01 (Agresti, 1996, p. 43).

Holland (1996) noted that the distribution of M can be approximated by the binomial distribution, that is: $M \overset{approx.}{\sim} B(w_s, p)$ where w_s, the number of wrong answers of s is known, but p is unknown. Holland (1996) suggested two ways of approximating p. In the first approach, p is computed such that the binomial distribution and the empirical distribution of M have the same means. Let $\overline{m}_{c'}$ denote the mean of the empirical agreement distribution which equals

$$\overline{m}_{c'} = \frac{\sum_{j'=1}^{n_{c'}} m_{c'j'}}{n_{c'}}. \tag{2.2}$$

Then, an estimate of p denoted as $p^*_{c'}$ is defined as

$$p^*_{c'} = \frac{\overline{m}_{c'}}{w_s}. \tag{2.3}$$

Let K^* denote the K-index based on $p^*_{c'}$, then K^* is given by

$$K^* = P(M \geq m_{c'c}) = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} (p^*_{c'})^g (1 - p^*_{c'})^{w_s - g}. \tag{2.4}$$

Holland (1996) showed using large empirical datasets that the binomial distribution using $p^*_{c'}$ yielded a "conservative" estimate of the empirical agreement distribution. That is, the K-index based on the binomial approximation is often stochastically higher than the K-index based on the empirical distribution (Agresti, 1990, p. 9).

To calculate $p^*_{c'}$, the response pattern of examinees in subgroup c' must be available. Note that the value of $p^*_{c'}$ is affected by the sample size– the

smaller the sample size, the less reliable is the estimate of $p^*_{c'}$. Holland suggested to approximate $p^*_{c'}$ using linear regression by utilizing the proportion of wrong answers (Q_r) of each examinee in each number incorrect score subgroup r as regressors. Using large datasets from ETS, Holland (1996) showed empirically that p^*_r, where p^*_r is defined analogously as in (2.3), is linearly related to Q_r. Let $\widehat{p}_r$ be the estimate of the binomial probability p^*_r using Q_r. The expression for $\widehat{p}_r$ is given as a piece-wise linear function with a and b as the intercept and slope parameters, respectively:

$$\widehat{p}_r = \begin{cases} a + bQ_r & \text{if} \quad 0 < Q_r \leq 0.3 \\ [a + .3b] + .4b[Q_r - .3] & \text{if} \quad 0.3 < Q_r \leq 1 \end{cases} . \tag{2.5}$$

Note that a and b have to be specified in order to estimate $\widehat{p}_r$ in (2.5). Holland (1996) used $a = 0.085$ and different values for b depending on the particular test that was used. However, from his study it is unclear how these values were obtained. Besides, they may vary across different tests.

In the present study, we proposed $\widehat{p}^*_1$ and $\widehat{p}^*_2$ as estimates of p^*_r based on linear and quadratic regression approach. Based on these estimates of p^*, two versions of K-index, $\overline{K}_1$ and $\overline{K}_2$ are defined as

$$\overline{K}_1 = P(M \geq m_{c'c}) = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} (\widehat{p}^*_1)^g (1 - \widehat{p}^*_1)^{w_s - g} \tag{2.6}$$

and

$$\overline{K}_2 = P(M \geq m_{c'c}) = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} (\widehat{p}^*_2)^g (1 - \widehat{p}^*_2)^{w_s - g}. \tag{2.7}$$

Only those examinees belonging to subgroup c' are used to estimate p by $p^*_{c'}$. On the other hand, $\widehat{p}^*_1$ and $\widehat{p}^*_2$ use relevant information from R subgroups. Therefore, $\widehat{p}^*_1$ and $\widehat{p}^*_2$ are expected to provide better estimates of p than $p^*_{c'}$.

The main aim of this study is to explore the usefulness of the K-index and its approximations given in (2.4), (2.6), and (2.7) under varying testing conditions. First, we investigated if the linear relationship between p^*_r and Q_r found by Holland (1996) also applies for relatively small datasets. Second, we investigated the fit of the binomial distribution using $p^*_{c'}$, $\widehat{p}^*_1$, and $\widehat{p}^*_2$ as an approximation to the distribution of M. Finally, we determined the empirical Type I error rates and detection rates of the K-index and the ω statistic. Because we used ω to evaluate the performance of the K-index, we introduced this statistic first.

2.3 The ω statistic

Wollack (1997) proposed the ω copying index that is formulated in the context of the nominal response model (NRM) (Bock, 1972). To determine ω, the NRM is used to estimate the probability of an examinee´s response to one of the item response categories $v[= 1, \ldots, h, \ldots, V]$. Under the NRM, the probability of examinee j with ability level θ_j responding to option h of item i with intercept and slope parameters ζ_{ih} and λ_{ih} is given as

$$P_{ih}(\theta_j) = \frac{\exp(\zeta_{ih} + \lambda_{ih}\theta_j)}{\sum_{v=1}^{V} \exp(\zeta_{iv} + \lambda_{iv}\theta_j)}. \tag{2.8}$$

Let h_{cs} be the number of identically answered items of s and c, and let $E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi})$ be the expected value of h_{cs} conditional on the ability level of the copier (θ_c), the item response vector of the source ($\mathbf{U}_s$), and the item parameters ($\boldsymbol{\xi}$). Furthermore, let $\sigma_{h_{cs}}$ be the standard deviation of h_{cs}. Then ω is given by

$$\omega = \frac{h_{cs} - E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi})}{\sigma_{h_{cs}}}, \tag{2.9}$$

where

$$E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi}) = \sum_{i=1}^{I} P(u_{ic} = u_{is}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi}). \tag{2.10}$$

Using the NRM, the probabilities of c selecting the responses of s can be determined. For any pair of examinees s and c, the distribution of ω approaches the standard normal (Wollack, 1997) as the number of test items becomes infinitely large. Thus, the ω values can be evaluated for statistical significance using the standard normal distribution.

The ω statistic is very similar to the g_2 index proposed by Frary, Tideman, and Watts (1977). The main difference is in the way the expected value of h_{cs} is computed; ω uses the nominal response model conditional on $\theta_c, \mathbf{U}_s$, and $\boldsymbol{\xi}$, whereas g_2 uses item distractors and difficulties from classical test theory and the ratio of the copier's number-correct score to the mean number-correct score for all examinees.

Wollack (1997) compared the empirical Type I error rates and the power of ω and g_2. The results showed that ω performed better than g_2 under the conditions simulated. In particular, g_2 failed to maintain the nominal Type I error rate which he found was too liberal in all circumstances. Therefore, in this study, the empirical Type I error and detection rates of the K-index were compared with ω.

Although both the K-index and the ω make use of item response similarities, ω compares the responses of the copier to the entire responses of the source, whereas in the K-index, the incorrect responses of the copier are compared with the incorrect responses of the source. Wollack (1997) pointed out that the power of a statistic that does not take into account the information from correctly answered items is likely to decrease due to a reduction in the number of operational items used. Besides, examinees that are most likely to be caught are those who miss several items. He added that "it is often not worthwhile to pursue a cheating claim if the alleged copier received a low score" (p. 13); an argument against a copying index that disregards correctly answered items such as the K-index.

The ω statistic is based on IRT modeling, in particular the nominal response model. First, it is reasonable to assume that the fit of the model to the data is important for the ω statistic to perform well. Second, if the suspected examinee copied a considerable number of items from the source, the ability level of the copier will be overestimated which consequently affects the value of ω. Finally, the estimation of the item parameters used in the NRM requires large number of examinees (Wollack, 1997); a requirement which may restrict the usefulness of this index in cases where large datasets are not available. For the latter case, Wollack and Cohen (1998) showed that estimating the item parameters on sample size as small as 100 for 40 and 80 items tests did not result in an increase in Type I error or a significant loss in power.

The K-index on the other hand, does not assume any IRT model and is therefore easier to apply in practice. However, a drawback of this index is that the number of examinees in each score group based on the number-incorrect scores should be large enough to obtain a reliable estimate of the binomial parameter p. For example, when simulating a standard test consisting of 40 items 10 times and drawing θ from the standard normal distribution for 30 simulees, the number of score groups ranges from 19 through 22 with score groups with only 1 simulee ranging from 12 through 15 (60%-74%) and other score groups consisting of only 2 or 3 simulees. Thus, p is very unreliably estimated for these samples.

2.4 Method

2.4.1 Data Generation

The NRM was used to generate item scores on multiple-choice tests with five options. Test lengths were 40 and 80 items and the number of simulees in

the sample were 100, 500, and 2000. These numbers were chosen to reflect small, medium, and large sample sizes. To be able to compare the results in this study with the results obtained by Wollack (1997), the same item parameters were chosen as in his study which were based on empirical data of a mathematics college placement test. Similarly, the ability parameter, θ_j, was drawn from $N(0, 1)$. Given the item and ability parameters, $P_{ih}(\theta_j)$ was computed for all i, h and j, using (2.8).

The observed response of examinee j to item i was obtained by drawing a sample from the set $v = \{1, \ldots, 5\}$, where each element of v has a probability of being drawn equal to $P_{i1}(\theta_j)$, $P_{i2}(\theta_j), \ldots, P_{i5}(\theta_j)$ respectively. In the NRM, the category with the largest algebraic value for λ has a monotonically increasing response function. As in other studies (e.g., Thissen & Steinberg, 1997), this category was chosen as the keyed alternative.

2.4.2 Simulation of Copying

To simulate copying, s and c were identified based on their ability percentile rank. Because in practice we are mostly interested in obtaining additional statistical evidence of answer copying for examinees that raise their scores by copying answers from an examinee with higher ability, we choose c such that the ability percentile rank of c is lower than that of s. This was also done to reflect the fact that the source is often a person with higher ability level than the copier (Holland, 1996). Simulees were first ordered according to θ. Then, in each dataset, s was selected as the simulee at the 90th or 60th percentile rank. In each dataset, 5% copiers were selected randomly from the simulees with θ level below the θ level of s.

Similar to Wollack (1997), copying was simulated by first randomly selecting an item and then altering the response of c to match the responses of s. This was done as follows. First n% of the items were randomly selected and then the item scores of c on these items were changed to match the item scores of s. For both 40-item and the 80-item tests, 10%, 20%, 30%, and 40% of the item scores were changed corresponding to 4, 8, 12, and 16 items in the 40-items test and 8, 16, 24, and 32 items in the 80-items test. The four factors – sample size (3 levels), number of items (2 levels), ability level of the source (2 levels), and percentage of items copied (4 levels) – were completely crossed to simulate 48 testing conditions. A program in S-plus (MathSoft, 2000) was written by the authors that performed the required simulation and necessary routine calculations.

The data used in this study share the following similar features with the data used by Wollack (1997): (a) the copier copied from a more able source,

(2) the number of copiers in each dataset and the percentage copied were the same, and (3) the same item parameters and distributional assumption were made for the θ parameters.

A difference with Wollack (1997) is that we did not use a seating chart to identify the $s-c$ pair. We assumed that there is a suspicion that c copied the answers from s. The K-index and the ω statistic were then used to check the probability that copying has occurred for a particular $s-c$ pair of examinees. So we did not use the statistics as a screening device. Wollack (1997) pointed out that in situations where there is only one source, ω has the highest power.

2.4.3 Data Analysis

Relationship Between p_r^* and Q_r

Recall that Q_r $(r = 1, \ldots, R)$ denote the proportion of wrong answers in each number-incorrect score group. For each score group r, we computed the binomial probability p_r^* using (2.3) with $\overline{m}_{c'}$ replaced by $\overline{m}_r$. To explore the relationship between Q_r and p_r^*, we first created scatterplots for p_r^* and Q_r. The information derived from visual inspection of these scatterplots suggested the kind of regression models to be fitted. On the basis of the results discussed below and on the empirical results obtained by Holland (1996), two standard linear regression models were proposed: *(a)* $\widehat{p}_1^* = \beta_0 + \beta_1 Q_r + \varepsilon_r$ and *(b)* $\widehat{p}_2^* = \beta_0 + \beta_1 Q_r + \beta_2 Q_r^2 + \varepsilon_r$, where β_0 and β_1 are the intercept and slope parameters respectively, β_2 is a regression parameter that indicates direction and amount of curvature, and ε_r is an error term which is assumed to have a normal distribution with mean 0 and constant variance σ^2. The fit of the two models was determined using the coefficient of multiple determination (R^2) and the magnitude of the residual standard error (see Neter et al., 1996). R^2 measures the proportionate reduction of total variation in p_r^* associated with the use of Q_r. The model with the largest R^2 and the smallest RSE was preferred.

Type I Error and Detection Rates

For a given α, a simulee was identified as a copier when the value of the K-index was less than or equal to α. For the ω statistic, a simulee was identified as a copier when the value of ω was above the one-tailed critical value corresponding to the upper α of the standard normal curve. In this study, assuming suspicion of a specific simulee copying from a specific source, the ω statistic was tested for significance without adjustment for α level. α

=.0001, .0005, .001, .0025, .005, .01 were used. These values were also used in Wollack (1997).

To investigate the empirical Type I error rate, we simulated tests of 40 and 80 items for 100, 500, and 2000 persons and we computed the proportion of noncopiers that were flagged as copiers. We used 100 replications. Similarly, the detection rate was investigated by taking the proportion of the true copiers that were detected.

2.5 Results

2.5.1 Relationship Between p_r^* and Q_r

Scatter plots of p_r^* and Q_r were investigated for different sample sizes and number of items. Results are shown in Figure 2.1. For sample size $J = 100$ (Figure 2.1a-b), the relationship seems to be linear but for sample size $J = 500$ (Figure 2.1c-d) p_r^* initially increases as Q_r increases then levels off at approximately $Q_r = 0.6$, and tends to decrease. For 2000 examinees (Figure 2.1e-f) it is clear that the relationship is curvilinear.

Quantitative assessment of the fit of the *linear* and *quadratic* regression models using R^2 and RSE revealed that the model which included the quadratic term had a better fit, that is, a larger R^2 and a smaller RSE. For example, for $J = 500$ and $I = 40$ (Figure 2.1c), the value of R^2 for the linear fit is 0.6 ($RSE = 0.03$), whereas including Q_r^2, the value of R^2 increases to 0.66 ($RSE = 0.03$). Testing H_o: $\beta_2 = 0$, with the alpha level of .001, β_2 is statistically unequal to zero, $p < .001$. Similar results were obtained for $J = 2000$ and $J = 100$. In general, p_r^* is estimated more accurately when the quadratic term is included.

2.5.2 Empirical and Binomial Agreement Distributions

For a particular choice of the source and the subject, several agreement distributions were constructed for the empirical K-index and for K^*, $\overline{K}_1$, and $\overline{K}_2$ based on the three versions of the binomial distributions ($p_{c'}^*$, $\widehat{p}_1^*$, and $\widehat{p}_2^*$). Results for different sample sizes were similar so we present in Figure 2.2 a typical example of these distributions for a sample size of 500. In Figure 2.2, but also in Figures 2.3-2.6, $\overline{K}_1$ and $\overline{K}_2$ were denoted as K1 and K2 respectively.

In general, the empirical distribution (Figure 2.2a) tends to have larger upper tail (negatively skewed) whereas the distribution based on $\widehat{p}_2^*$ (Figure 2.2d) consistently have smaller upper tails. Note that the size of the upper

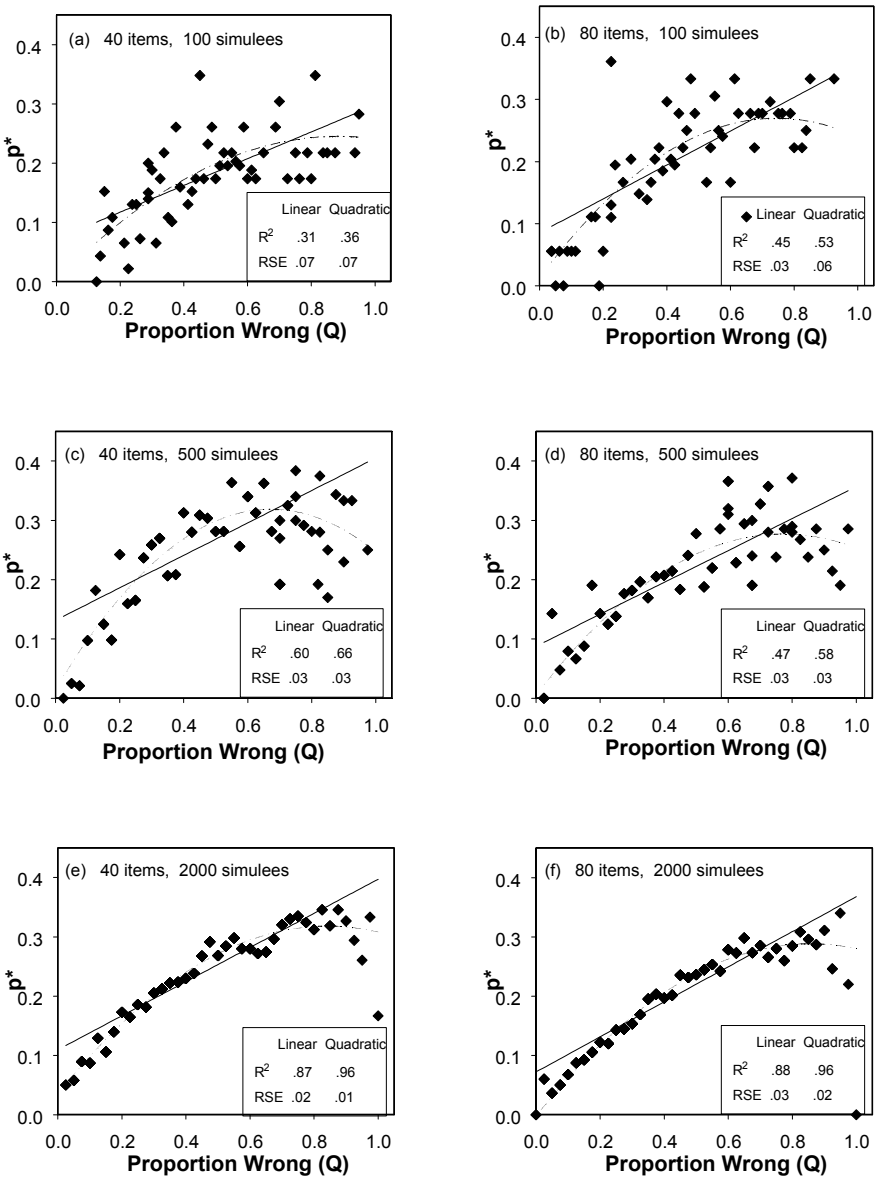


Figure 2.1: Scatter Plots of p^* and Proportion Wrong (Q).

tail of the distribution greatly influences the value of the K-index. As can be seen from (2.1), (2.4), (2.6), and (2.7), the K-index is computed as the sum of the upper tail probability densities. This implies that a distribution with the smallest upper tail yields smallest numerical values of the K-index and thus provides the strongest evidence of answer copying. Since the empirical agreement distribution has a larger upper tail, it is expected that the K-index computed based on this distribution will be large and thus implies low detection rates.

Further, we found that the empirical distribution had the largest upper tail when the number of simulees was smallest, that is for $J = 100$ (graph not presented here). Thus, for $J = 100$, the K-index based on (2.1) is expected to be too conservative.

2.5.3 Type I Error Rate

Figure 2.3 shows the graphical comparison of the empirical Type I error rates of the K-index and ω, across combinations of examinee sizes and number of items. Type I error rates that are on the identity (boundary) line represent perfect Type I error control; Type I errors above the boundary line are larger than the nominal values and those below it are smaller than the nominal values. The Type I error rate of the K-index based on (2.1) was found to be much below the nominal α level and is not presented here.

The K-indices were able to control the Type I error rates below the nominal α level in all situations considered. In most cases, ω was also able to control its Type I error below the nominal level, with the exceptions for the 80-item test with 500 and 2000 simulees wherein the Type I error of ω exceeded its nominal level by approximately .005 for $\alpha = .01$ (see Figures 2.3d and 2.3f).

We also investigated the variance of the K-index and ω across replications. The variance of the K-index decreased with increasing percentage of copying, sample size, and number of items. The variance of ω decreased with increasing percentage of copying but unlike the K-index, ω seems not sensitive to changes in sample size and number of items. The variance of the K-index was almost equal to ω for longer tests, large number of examinees, and a large percentage of copying. For example, for an 80-item test and 100 examinees, the variance of $\overline{K}_1$, $\overline{K}_2$ and ω for 10% copying are .0955, .0939 and .0704 respectively. As the percentage of copying increases to 40%, the three variances decrease to .003, .003, and .002, respectively.

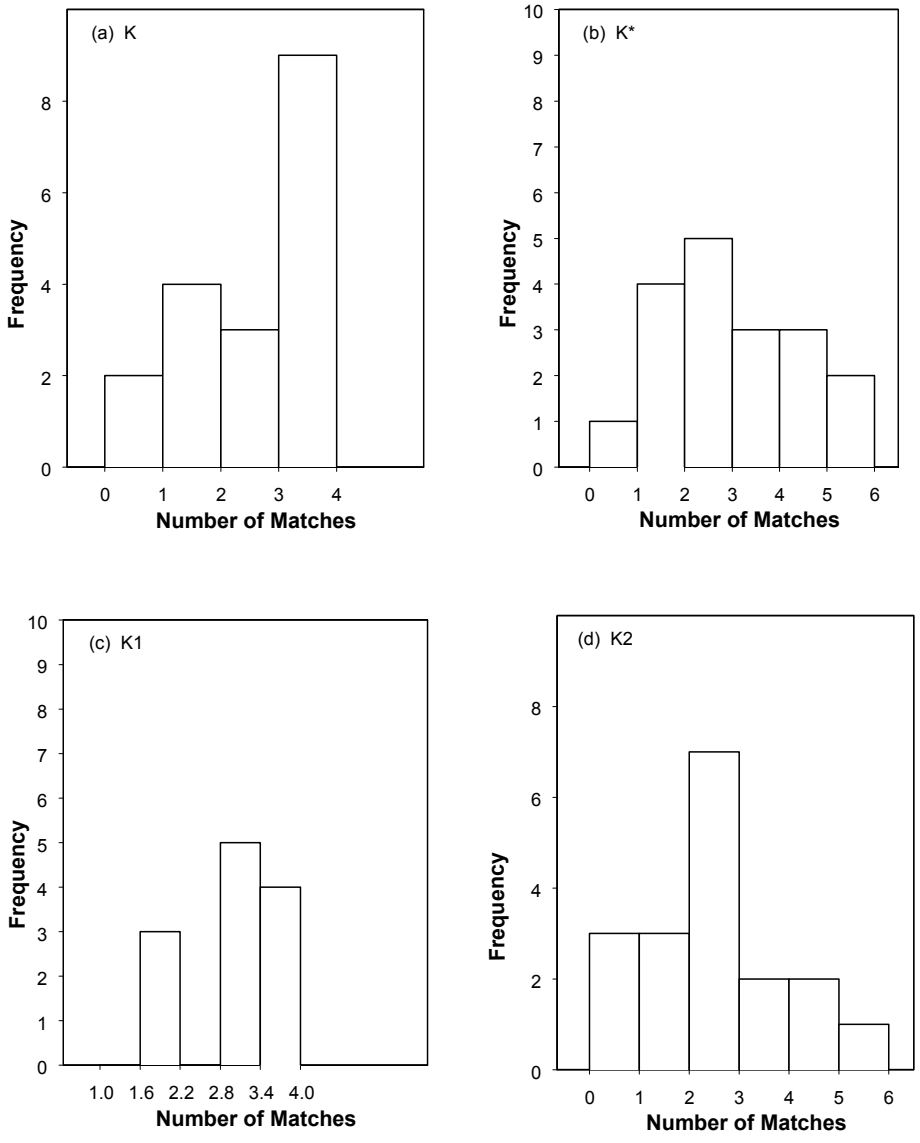


Figure 2.2: Empirical and Binomial Agreement Distribution.

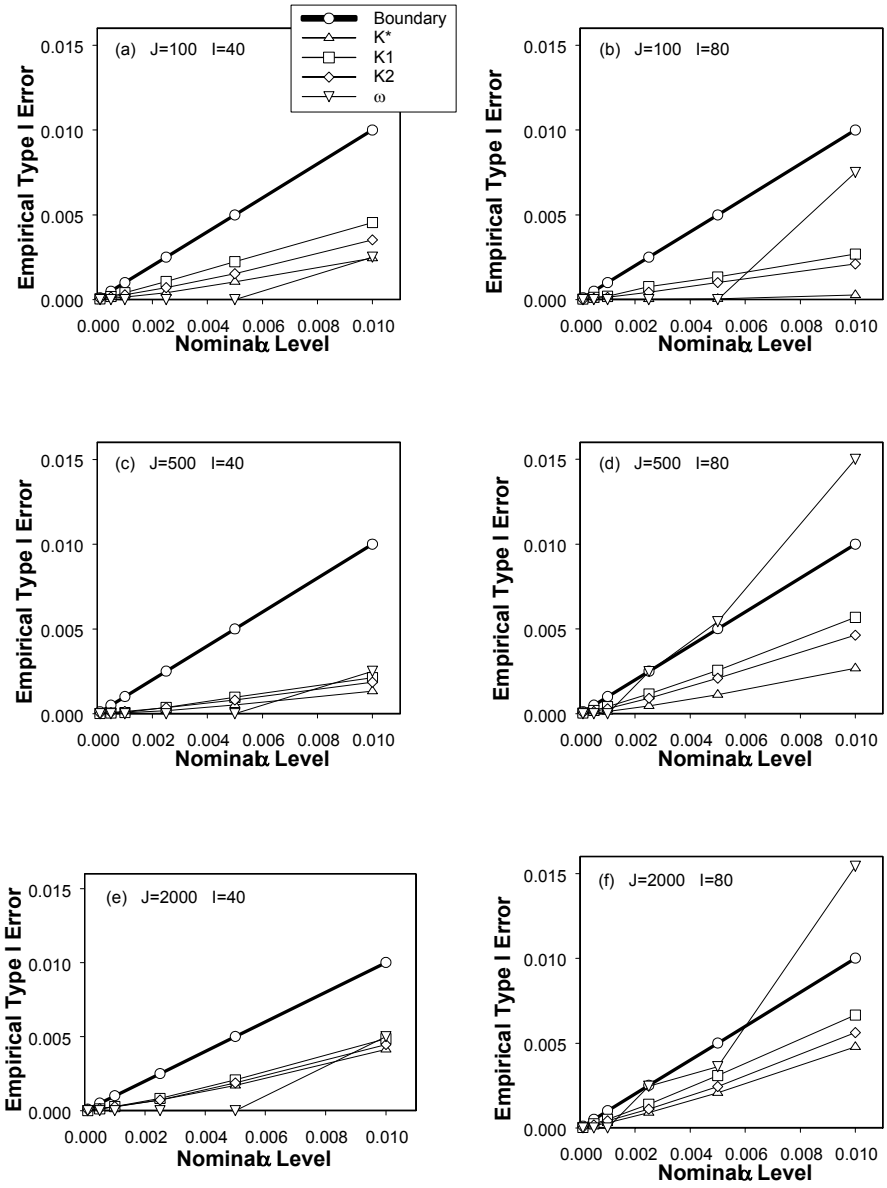


Figure 2.3: Nominal and Empirical Type I Error Rates as a Function of Simulee Size and Test Length.

2.5.4 Detection Rate

The detection rates of K^*, $\overline{K}_1$, $\overline{K}_2$, and ω as a function of α-level for different percentages of copying, sample sizes and test lengths were first investigated for the source fixed at the 90th percentile. The K-index based on (2.1) was not included in the current analysis because its detection rate was very low. Figure 2.4 shows the detection rates for 100 simulees on the 40-item test and Figure 2.5 for 500 simulees on the 80-item test. The detection rates for the other simulated configurations were similar and are not presented here.

In almost all simulated datasets, ω had the highest detection rate. The difference between the detection rates of ω and the K-indices is relatively large for small sample size and test length but tends to diminish as the sample size and test length increased. For example, for $\alpha = .01$, the difference in detection rate between ω and $\overline{K}_2$ is 0.15 for $J = 100$, $I = 40$, and 40% copying (see Figure 2.4a) and it reduce to 0.02 for $J = 500$, $I = 80$, and 40% copying (see Figure 2.5a). The K-index based on the binomial distribution where p was estimated using linear regression with quadratic term included ($\overline{K}_2$), appeared to be slightly better than $\overline{K}_1$. As expected, K^* had the lowest detection rate.

Further note that the detection rates of the ω and the K-indices increased with the percentage of copied answers. Thus, examinees who copied many items are more likely to be detected than examinees who copied few items.

The probability of detecting a copier who copied 10% of the items is very low–at most .08 for ω and less than .05 for the K-indices (see Figures 2.4d and 2.5d)

Increasing the number of simulees had no substantial effect on the detection rates of ω. This is expected since the computation of ω depends only on the response pattern of the source and the copier and not on other examinees. On the other hand, the detection rates of the K-indices increased with the sample size and number of items. For example, for for $\alpha = .01$ and 40% copying the detection rate of $\overline{K}_2$ is 0.69 for $J = 100$ and $I = 40$ (see Figure 2.4a) and it increased to 0.92 for $J = 500$ and $I = 80$ (see Figure 2.5a).

To investigate the influence of the proficiency level of the source, we also investigated the detection rates of the indices when the source was at the 60th percentile rank. Results are shown in Figure 2.6 for 100 simulees and a 40-item test. Comparing Figure 2.6 with Figure 2.4 revealed a slight increase in the detection rate of ω, $\overline{K}_1$ and $\overline{K}_2$ for 40% and 30% copying, but for 20% and 10% copying, the detection rates were almost the same; the detection rate of K^* substantially increased for 40% copying but not for

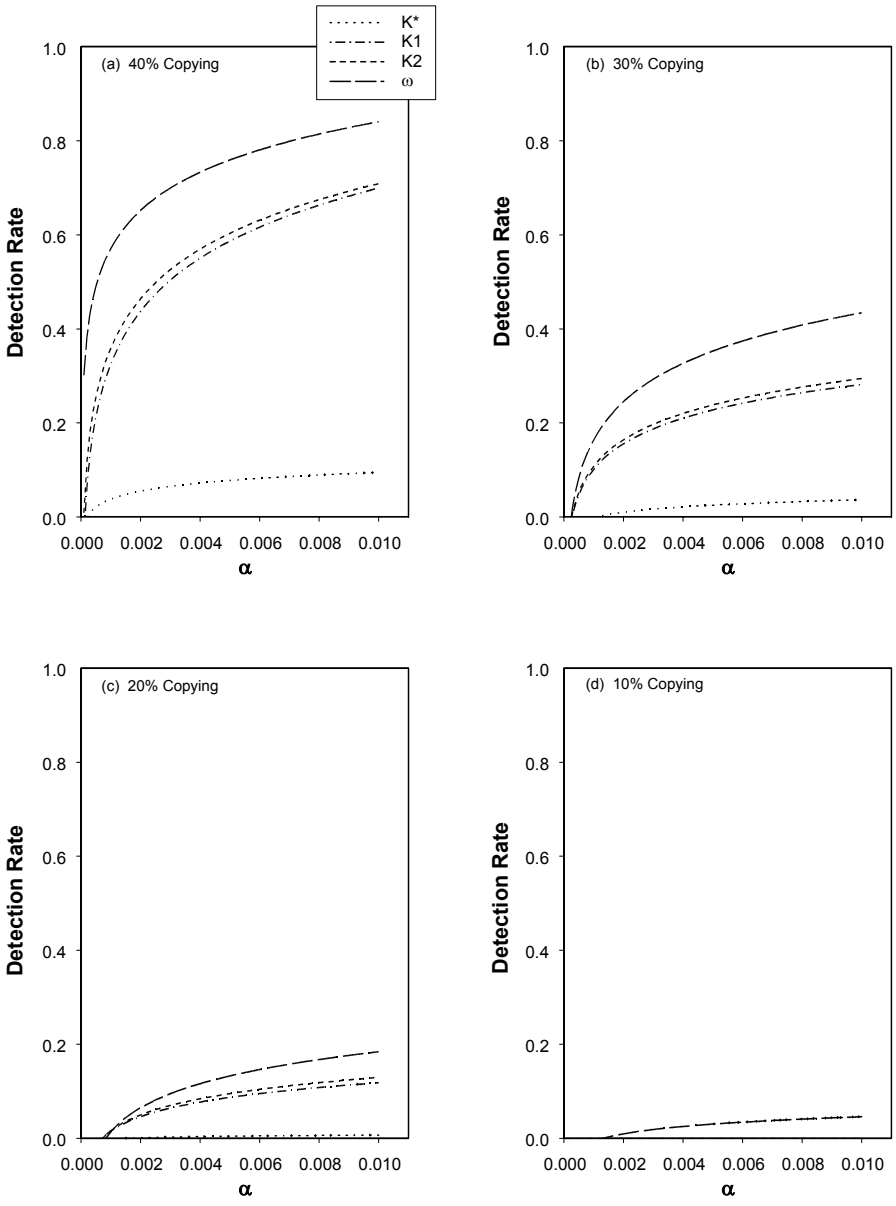


Figure 2.4: Detection Rate of the K-index and ω, as a Function of Copying Percentage, on 40-item Test, 100 Simulees, and the Source at the 90th Percentile Rank.

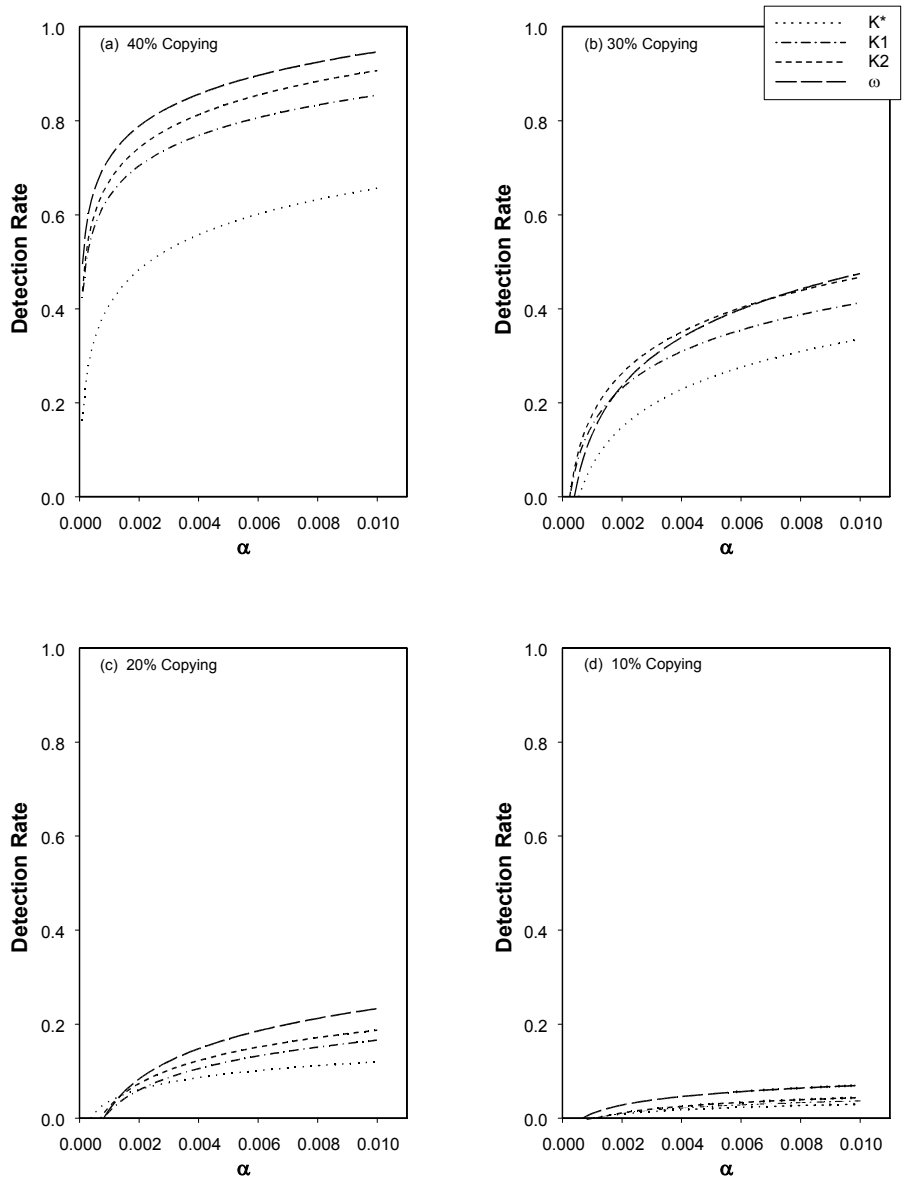


Figure 2.5: Detection Rate of the K-index and ω, as a Function of Copying Percentage, on 80-item Test, 500 Simulees, and the Source at the 90th Percentile Rank.

the other percentages of copying. Comparing the indices within Figure 2.6 revealed that ω still remains the highest detection rate followed by $\overline{K}_2$, $\overline{K}_1$, and K^*.

2.6 Discussion

In this study we investigated the statistical properties of the K-index and compared its detection rate with the detection rate of the ω statistic. The practical usefulness of these statistics will depend on the application at hand. As was shown in this study, the use of these indices need not be restricted to large-scale testing but can also be applied for small samples consisting of 100 examinees. As others have discussed, these indices can be used to obtain additional evidence for answer copying when a proctor has observed irregular behavior during the test. An alternative is to use these indices for routine monitoring of test responses to evaluate efforts to prevent copying or for triggering the need to employ such measures. For example, a teacher can inform privately an examinee pair with a very high index value and suggest that they not sit together on subsequent tests.

Results showed that in general, the binomial success probability, p, is better estimated by a quadratic function than by a linear function of the proportion wrong answers, Q. However, when the dataset is large ($J = 2000$), the relationship between p and Q was nearly linear at the lower end of Q (e.g., $Q < 0.6$). This finding supported the findings by Holland (1996) when he used the linear function to estimate p by Q. In his study, he used ETS data for which the source and the copier generally belonged to the upper end of the ability continuum (e.g., few wrong answers or low value of Q).

When using the K-index for small datasets ($J = 100$), it is not advisable to use the empirical agreement distribution nor its binomial approximation in (2.4). In terms of distributional shape, the empirical agreement distributions was negatively skewed whereas the binomial distributions—especially the one based on $\widehat{p}_2^*$—were positively skewed. This resulted in a larger numerical value of the K-index despite the higher percentage of copying.

Results further showed that all approximations of the K-index were able to hold the Type I error rates below the nominal level in all situations simulated. Thus, the K-index has more favorable statistical properties than the g_2 index (Frary, Tideman, & Watts, 1977) which failed to control the nominal Type I error rates (Wollack, 1996).

Although ω had higher detection rates than $\overline{K}_1$ and $\overline{K}_2$ for simulee sizes

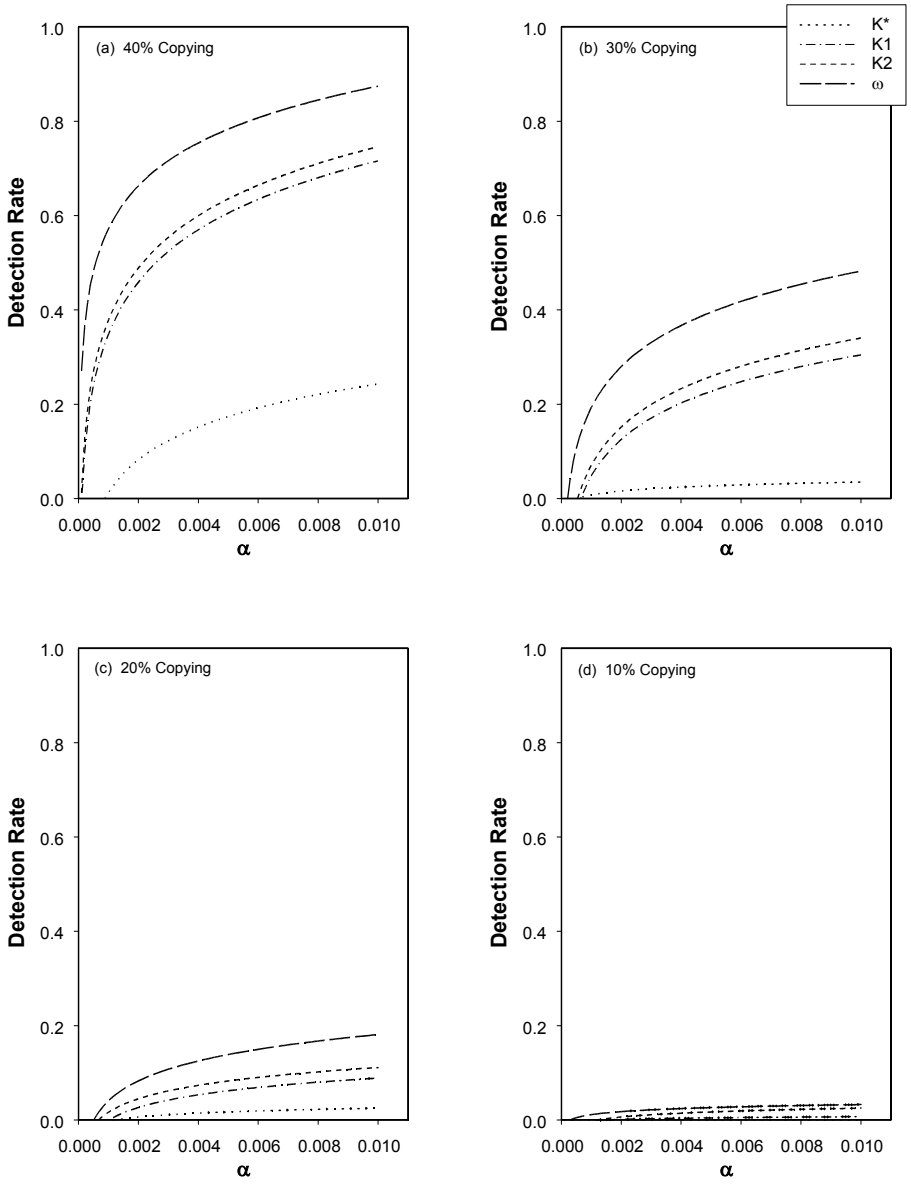


Figure 2.6: Detection Rate of the K-index and ω, as a Function of Copying Percentage, on 40-item Test, 100 Simulees, and the Source at the 60th Percentile Rank.

100 and 500, the differences in detection rates are small using 2000 simulees. It is expected that using more than 2000 simulees will further improve the detection rates of $\overline{K}_1$ and $\overline{K}_2$. We do not recommend the use of K^* in practice while $\overline{K}_2$ might be a good alternative if it is not possible to use ω.

Finally, the random variable M is a non-negative count of matching incorrect answers. For future study, it may be important to investigate the fit of a Poisson distribution as an alternative distribution for the random variable M. Furthermore, a weighted matching correct answers between the source and the copier can be included in the computation of the copying index. The weight may be taken as some function of the probability of correct response. Incorporating the weighted matching correct answers in addition to matching incorrect answers differentiates the K-index from this new index. Also, several measures can be investigated to minimize the impact of discreteness due to small sample size.

Chapter 3

Variants of the K-Index

Abstract. Two new indices to detect answer copying on a multiple-choice test—S_1 and S_2—were proposed. The S_1 index is similar to the K-index (Holland, 1996) and the $\overline{K}_2$-index (Sotaridona & Meijer, 2002) but the distribution of the number of matching incorrect answers of the source and the copier is modeled by the Poisson distribution instead of the binomial distribution to improve the detection rate of K and $\overline{K}_2$. The S_2 index was proposed to overcome a limitation of the K and $\overline{K}_2$ index, namely, their insensitiveness to correct answer copying. The S_2 index incorporates the matching correct answers in addition to the matching incorrect answers. A simulation study was conducted to investigate the usefulness of S_1 and S_2 for 40- and 80-item tests, sample sizes of 100 and 500 simulees, and 10%, 20%, 30%, and 40% answer copying. The Type I errors and detection rates of S_1 and S_2 were compared with those of the $\overline{K}_2$ and the ω copying index (Wollack, 1997). Results showed that all four indices were able to maintain their Type I errors, with S_1 and $\overline{K}_2$ being a little more conservative compared to S_2 and ω. Furthermore, S_1 had higher detection rates than $\overline{K}_2$. The S_2 index showed a significant improvement in detection rate compared to K and $\overline{K}_2$.

This chapter has been published as: Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40, 53-69.

3.1 Introduction

Cheating on tests has a long and rich tradition (Cizek, 1999, Chap. 5). Example of cheating methods are using forbidden materials, circumventing the testing process, or even using microrecorders. In the present study, we will be concerned with answer copying. In this type of cheating, one examinee copies the answers from another examinee, which may take place using all kinds of codes for transmitting answers and a code for doing so, for example, clicking of pens, tapping of the foot, and the like. Thus the examinees do not have to be in the physical neighborhood of each other. Because answer copying may invalidate an examinee's test score, it is necessary to prevent those practices by using well-instructed proctors and construct the seating arrangements so that there is ample room between the examinees. However, if a proctor observes some irregularities, statistical methods may be used to obtain additional evidence of answer copying.

Several methods have been proposed that all are based on determining the probability that the observed score patterns of two examinees under suspicion are similar. These chance methods can be classified into two types (Cizek, 1999, pp. 138-139). One type of method compares an observed pattern of responses to a known theoretical distribution (e.g., Frary, Tideman, & Watts, 1977; Wollack, 1997). In the second type of method, the probability of an observed pattern is compared with a distribution of values derived from independent pairs of examinees who took the same test. An example of such a statistic is the K-index (Holland, 1996).

Sotaridona and Meijer (2002) investigated the statistical properties of different forms of the K-index and compared the detection rate of these indices with the detection rate of the ω index (Wollack, 1997). The major difference between the indices is that the K-index does not assume any test model, whereas ω is based on item response theory modeling (e.g., van der Linden & Hambleton, 1997). Sotaridona and Meijer (2002) discussed that the K-index is less sensitive to answer copying when both the source and the copier have many matching correct answers. Lewis and Thayer (1998) and Sotaridona and Meijer (2002) found that the K-index that used the binomial distribution to model the matching incorrect answers had low power to detect a substantial amount of copying.

In this paper, we will propose an index S_2 that both takes the matching correct and the matching incorrect answers into account. Furthermore, we discuss an index S_1 which mathematical form is similar to the K-index (Holland, 1996) and the $\overline{K}_2$-index (Sotaridona & Meijer, 2002) but the distribution of the number of matching incorrect answers of the source and the

copier is modeled by the Poisson distribution.

This study is organized as follows. First, we introduce the K-index and the ω index. Second, we discuss two new indices S_1 and S_2 that may be used to obtain additional evidence of answer copying. Third, we conducted a simulation study to investigate the Type I error rate and the detection rates of S_1 and S_2.

3.2 The ω and $\overline{K}_2$

In this study, the copying indices ω (Wollack, 1997) and $\overline{K}_2$ (Sotaridona & Meijer, 2002) are compared to the newly proposed copying indices, S_1 and S_2, with respect to the Type I errors and detection rates. A brief description of ω and $\overline{K}_2$ is given below followed by a more elaborate discussion of S_1 and S_2. The reader is referred to Sotaridona and Meijer (2002) and Wollack (1997) for a more detailed treatment of $\overline{K}_2$ and ω respectively.

3.2.1 The ω Index

Let examinee c, the copier, be suspected of copying answers from examinee s, the source. In a multiple-choice test with options $v = 1, 2, \ldots, k, \ldots, V$, let h_{cs} be the number of items $i = 1, 2, \ldots, I$ where the response of c matches the response of s. Given that the response of s on i is k, let $P_{ik}(\theta_c)$ denote the probability of c selecting the same option k on item i. Wollack (1997) used the nominal response model (Bock, 1972) to obtain this probability which is given by

$$P_{ik}(\theta_c)=\frac{\exp(\zeta_{ik}+\lambda_{ik}\theta_c)}{\sum_{v=1}^{V}\exp(\zeta_{iv}+\lambda_{iv}\theta_c)}, \tag{3.1}$$

where ζ_{ik} and λ_{ik} are the intercept and the slope parameters. The expected value of h_{cs} is computed, conditional on the ability level of the copier (θ_c), the item response vector of the source $\mathbf{U}_s = (U_{1s}, \ldots, U_{Is})$ where U_{is} is the response to item i, and the item parameters $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_I)$ with $\xi_i = (\zeta_{i1}, \ldots, \zeta_{iV}, \lambda_{i1}, \ldots, \lambda_{iV})$, as

$$E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi}) = \sum_{i=1}^{I} P_{ik_i}(\theta_c), \tag{3.2}$$

and the standard deviation of h_{cs} is

$$\sigma_{h_{cs}} = \sqrt{\sum_{i=1}^{I} [P_{ik_i}(\theta_c)]\,[1 - P_{ik_i}(\theta_c)]}. \tag{3.3}$$

The ω index is based on the residual between the observed and the expected value of h_{cs}. A standardized residual defines ω, which is asymptotically standard normally distributed (Wollack, 1997). The larger the value of ω, the stronger the evidence that c copied from s. The ω-statistic is given by

$$\omega = \frac{h_{cs} - E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi})}{\sigma_{h_{cs}}}. \tag{3.4}$$

3.2.2 The $\overline{K}_2$ Index

Define the number-incorrect group $r = 1, 2, \ldots, c', \ldots, R$ such that examinees $j = 1, 2, \ldots, J_r$ have the same number of wrong answers, and c' indicate the group membership of c. The number of examinees in number-incorrect group r is denoted by J_r so that $J_{c'}$ is the number of examinees with the same number of wrong answers as examinee c. The index jr will be used to indicate an examinee j in number-incorrect group r. Let U_{ijr} be the response of examinee jr to item i and let W_s be the set of items answered incorrectly by s. The number of items in W_s is denoted by w_s.

For each examinee jr, an indicator variable A_{ijr} equal to 1 if $U_{ijr} = U_{is}$, and 0 otherwise. The item response of s is indexed by is indicating that s does not belong to any number-incorrect group. The number of matching incorrect answers of jr and s, denoted by M_{jr} is then defined as

$$M_{jr} = \sum_{i \in W_s} A_{ijr}. \tag{3.5}$$

For a particular $s - c$ pair, M_{jr} is observed for each examinee jr. For simplicity, M_{jr} will be denoted by M when it is not necessary to identify the examinee.

Let $M_{c'c}$, with realization $m_{c'c}$, be the number of matching wrong answers between c and s, and let p be the success probability parameter in the binomial distribution. The $\overline{K}_2$ index is similar to the K index discussed by Holland (1996). For example, both indices are based on the random variable M, and are computed as the sum of the binomial probabilities

$$\Pr(M_{c'c} \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} p^w (1-p)^{w_s - w}. \tag{3.6}$$

The rationale behind the choice of the binomial distribution for M is discussed in Holland (1996). The main difference between K and $\overline{K}_2$ is the way p is estimated. For the K index, p is estimated by $\widehat{p} = \frac{\overline{M}_{jc'}}{w_s}$, where $\overline{M}_{jc'}$ is the mean number of matching incorrect answers in the number-incorrect group c' (Holland, 1996). For the $\overline{K}_2$ index, p is estimated by $\widehat{p}_2 = E(\beta_0 + \beta_1 Q_r + \beta_2 Q_r^2 + \varepsilon_r)$, where Q_r is the proportion of wrong answers of examinees in number-incorrect group r. The parameters β_0, β_1, and β_2 are regression coefficients, and ε_r is an error term which is assumed to have a normal distribution with mean 0 and variance σ^2.

Note that the value of $\widehat{p}$ is obtained using data in the number incorrect group c' only, whereas the value of $\widehat{p}_2$ is obtained using all relevant information from R number incorrect groups. In this sense, $\widehat{p}_2$ contains more information than $\widehat{p}$, and therefore is expected to give a better estimate of p than $\widehat{p}$.

The $\overline{K}_2$ index is defined as

$$\overline{K}_2 = \Pr(M_{c'c} \geq m_{c'c}) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} \widehat{p}_2^{\,w} (1 - \widehat{p}_2)^{w_s - w}. \tag{3.7}$$

The $\overline{K}_2$ index is an upper-tail probability. This probability can be compared to a chosen nominal level of significance α, such as 0.01. When it is less than or equal to this value, c may be identified as having a pattern of responses unusually similar to that of s.

Sotaridona and Meijer (2003) showed that the detection rates of the $\overline{K}_2$ index were in general higher than those of the K index, whereas ω yielded the highest detection rates. Furthermore, both $\overline{K}_2$ and ω were able to keep their empirical Type I errors below the nominal levels. Note that the negative consequence of falsely identifying a noncopier as copier is severe, so we prefer an index that has a Type I error rate at the nominal level or slightly below the nominal level.

3.3 Two New Indices

3.3.1 The S_1 Index

The S_1 index is similar to the $\overline{K}_2$ index in the sense that it is also based on the random variable M. The S_1 index differs from the $\overline{K}_2$ index in the following ways. First for the $\overline{K}_2$ index, M is assumed to follow a binomial distribution whereas for S_1, M is assumed to follow a Poisson distribution.

Secondly, the Poisson parameter μ (the expected value of M), is estimated using a loglinear model, whereas in the $\overline{K}_2$ index the binomial parameter p was estimated using a linear regression model. The motivation for proposing the Poisson distribution for M, the loglinear model for estimating μ, and a statistic for checking the adequacy of the loglinear model are discussed below. Once the estimate of μ for number incorrect group c', $\widehat{\mu}_{c'}$, is obtained, the S_1 index is computed as

$$S_1 = \sum_{w=m_{c'c}}^{w_s} \frac{e^{-\widehat{\mu}_{c'}} \widehat{\mu}_{c'}^{w}}{w!}. \tag{3.8}$$

Note that S_1 is not an upper-tail probability since the Poisson distribution puts no upper limit on the number of matching wrong answers, whereas there is an upper limit on M which is w_s. Equation (3.8) is interpreted as the probability of w_s being greater than $m_{c'c}$. The smaller the value of S_1, the stronger the evidence of answer copying.

The Choice for the Distribution of M

Several distributions have been assumed for the random variable M by previous researchers dealing with copying indices. Bay (1995) used the compound (or generalized) binomial distribution in developing the B_m copying index. In B_m, all items in the item score pattern are considered. The ESA copying index (Belleza & Belleza, 1989), the K index (Holland, 1996), and the $\overline{K}_2$ index (Sotaridona & Meijer, 2002) only considered the items incorrectly answered and used the binomial distribution for M. Wollack (1997, p. 309) criticized the B_m and ESA indices because they cannot adjust the probabilities associated with an examinee's responses as a function of the test score. Wollack (1997) found that B_m and ESA had lower detection rates compared to other indices based on classical test theory like the g_2 index (Frary, Tideman, & Watts, 1977). We did not include the g_2 index in this study since Wollack (1997) found that the Type I errors of g_2 are grossly inflated. Unlike g_2, the $\overline{K}_2$ index is able to control its Type I error below its nominal level.

Remember that the responses of the source to a set of test items are considered fixed and that given these responses we count the number of wrong responses of the copier that matches that of the source and call it M. Since the binomial distribution resulted in low detection rates for K and $\overline{K}_2$ (Lewis & Thayer, 1998; Sotaridona & Meijer, 2002), we propose the S_1 index that assumes a Poisson distribution as a reasonable approximation to

the distribution of M. One may conceptualize S_1 as monitoring the rate or number of answer matches per incorrect answer by the source. If this rate is sufficiently high, then this provides evidence of answer copying. The extent to which the Poisson distribution approximates the distribution of M was investigated empirically.

Model for Estimating μ

To compute S_1 in (3.8), we should determine w_s, $m_{c'c}$, and μ. The value of w_s and $m_{c'c}$ are known whereas the expected value of M, μ, must be estimated. The mean of M differs across different ability levels. The values of M are small if most of the examinees have high ability level because the number of items incorrectly answered by the copier to match the items answered incorrectly by the source is small. On the other hand, if most of the examinees have low ability level, the number of items answered incorrectly is large and the number of matching items is likely to be large. This information is taken into account when estimating μ by stratifying the examinees according to the number-incorrect score they obtained.

Since the Poisson distribution is assumed for M, it is standard practice to use the loglinear model to model the log of the mean of M (Agresti, 1996, p. 73). Using this model, it allows μ to be nonlinearly related to the predictor variable which in this case is the number of wrong answers (see also Hanson, 1994 for using the loglinear model in the context of the compound binomial distribution).

The relevant data for estimating μ are the number of incorrect scores and the mean number of matching incorrect scores for each number-incorrect group. Let μ_r denote the expected value of the Poisson variate M_{jr}. The loglinear model has the form

$$\log(\mu_r) = \beta_0 + \beta_1 w_r, \forall\ r \tag{3.9}$$

where β_0 is the intercept term signifying the logarithm of the population mean across R number incorrect groups, and β_1 is the slope parameter. Estimation of β_0 and β_1 in (3.9) is discussed in Agresti (1996, p. 93).

To obtain S_1, we need to determine the fitted mean for the number incorrect group to which the copier belongs. This fitted mean equals

$$\widehat{\mu}_{c'} = \exp(\widehat{\beta}_0 + \widehat{\beta}_1 w_{c'}).$$

Model Checking

The fit of the loglinear model in (3.9) was investigated using the likelihood-ratio goodness-of-fit statistic, G^2, (Agresti, 1996, p. 89). The G^2 statistic can be used to test the null hypothesis that the model fits the data against the alternative that the model does not fit the data. Let $\widehat{\mu}_r$ be the fitted mean number of matching incorrect answers of number-incorrect group r. The G^2 statistic is given by

$$G^2 = 2\sum_{r=1}^{R} \mu_r \log\left(\frac{\mu_r}{\widehat{\mu}_r}\right). \tag{3.10}$$

If the model perfectly fits the data, $\widehat{\mu}_r = \mu_r$. In such a case, $\log\left(\frac{\mu_r}{\widehat{\mu}_r}\right) = 0$ and consequently $G^2 = 0$. The distribution of G^2 is approximately chi-squared with degrees of freedom equal to R minus the number of model parameters. For the loglinear model in (3.9), the number of model parameters is 2. The p-value to test the null hypothesis is the right-tail probability. Large values of G^2 or small p-values, for example, less than .01, would suggest a poor model fit (Agresti 1996, p. 89). If the fit of the model to the data is poor then it would not be appropriate to use (3.8) as a statistical test of answer copying.

3.3.2 The S_2 Index

Copying indices that are based solely on the matching incorrect answers, such as the K and $\overline{K}_2$ indices, discard the additional information about copying that is available in the matching correct answers. By excluding the number of matching correct answers in the analysis of answer copying, we explicitly assume that c completely knows the answer to item i whenever c and s give a correct response to item i. However, this is not always the case. An examinee may obtain the correct answer to an item by copying or by guessing.

Note that the K and $\overline{K}_2$ indices are not sensitive to a copier who is copying only the correct answers of the source. This may be the case when s and c are friends and s shares his or her answers to c on items where he or she is almost sure of the correct answers. Another example is a high-stakes examination where c may bribe s for sharing his correctly answered items to c.

The new copying index S_2 is proposed to overcome this limitation. We propose S_2 to incorporate information about copying that is contained in

the matching correct answers in addition to the information in the matching incorrect answers. Note that in K and $\overline{K}_2$, the evidence of answer copying is given a weight of 1 if s and c choose the same wrong option to an item, and 0 otherwise. For S_2, the weight is 1 if s and c choose the same wrong option to an item, δ (to be defined below) if s and c both choose the same correct option, and 0 otherwise. The variable δ quantifies the amount of correct-answer copying information to an item for a particular source-copier pair.

Let i^* be an item that was answered correctly by s, and let U_{i^*jr} be the response of examinee jr to item i^*. Then, δ_{i^*jr} gives the estimate of copying information on item i^* by examinee jr. The value of δ_{i^*jr} satisfies the inequality

$$1 \geq \delta_{i^*jr} \geq 0,$$

that is, $\delta_{i^*jr} = 0$ if jr knows the correct answer to item i^* and $\delta_{i^*jr} = 1$ if jr is completely ignorant about the correct answer to item i^* (see conditions 1-2 below). The problem is to quantify the amount of knowledge that jr has on i^*. To do this we have to obtain the probability of jr answering item i^* correctly. This probability can be estimated as the proportion of examinees in number-incorrect group r getting the correct answer to item i^*. A drawback of this approach is that the estimate is highly dependent on the population of examinees taking the test. For example, the estimate will be low for a high ability population of examinees, whereas the estimate will be high for a low ability population of examinees. A solution is to condition on the ability level of the suspected copier.

Let P_{i^*jr} denote the probability that jr answers item i^* correctly, and let A_{i^*jr} be an indicator variable equal to 1 if $U_{i^*jr} = u_{i^*s}$, and 0 otherwise. Given U_{i^*s}, this probability is given by

$$P_{i^*jr} = \Pr\left(U_{i^*jr} = u_{i^*s} \mid U_{i^*s}\right), \tag{3.11}$$

and the maximum likelihood estimate of P_{i^*jr} equals

$$\widehat{P}_{i^*jr} = \frac{\sum_{j=1}^{J_r} A_{i^*jr}}{J_r}. \tag{3.12}$$

Note that P_{i^*jr} is determined conditional on the observed response pattern of the source. Given the estimate of P_{i^*jr}, what remains is to transform this estimate into δ_{i^*jr}. A suitable transformation function, $f(P_{i^*jr})$, satisfies the following conditions:

1. $f(P_{i^*jr})$ approaches 0 as P_{i^*jr} approaches 1; that is, the evidence of answer copying diminishes as P_{i^*jr} approaches 1.

2. $f(P_{i^*jr})$ approaches 1 as P_{i^*jr} approaches 0; that is, the evidence of answer copying approaches 1 if the suspected copier answers an item correctly despite low probability of giving the correct answer to such an item.

3. Tests with different number of options must have different weight functions. Let f and f' be two different weight functions and let i^* and $i^{*\prime}$ be items taken from two tests with number of options V and V' such that $V < V'$. Then it should hold that $f(P_{i^*jr}) > f'(P_{i^{*\prime}jr})$ whenever $P_{i^*jr} = P_{i^{*\prime}jr}$.

The basis for conditions 1-2 should be clear from the above discussions. Condition 3 arises from the idea that multiple-choice tests with different number of options should have different transformation functions that differ by a factor that is a function of the number of options. The proposed function should account for the probability of guessing to an item.

For notational convenience, let g denote the probability of answering item i correctly by guessing. For example, for 5-option test, $g = 0.20$ for a 4-option test, $g = 0.25$. A function satisfying conditions 1-3 is shown in (3.13) is

$$\delta_{i^*jr} = f(P_{i^*jr}) = d_1 e^{d_2 P_{i^*jr}}, \tag{3.13}$$

where

$$d_2 = -\left(\frac{1+g}{g}\right) \text{ and } d_1 = \left(\frac{1+g}{1-g}\right)^{d_2 P_{i^*jr}}.$$

Equation (3.13) is a monotone decreasing function of P_{i^*jr} where g a scaling constant. Figure 3.1 shows the graph of (3.13) with $g = .2$ (denoted as $F1-$ 5 options) and $g = .25$ (denoted as $F2-$ 4 options). As shown in the graph, the value of δ_{i^*jr} for both $F1$ and $F2$ approaches zero as P_{i^*jr} approaches 1 and δ_{i^*jr} approaches 1 as P_{i^*jr} approaches zero (conditions 1-2). Furthermore, $F1(P_{i^*jr}) < F2(P_{i^*jr})$ for $P_{i^*jr} \in (0, 1]$ (condition 3).

Let M^*_{jr} denote the sum of the number of matching incorrect answers and weighted matching correct answers by examinee jr and examinee s. The expression for M^*_{jr} is given by

$$M^*_{jr} = M_{jr} + \sum_{i^*} \delta_{i^*jr}. \tag{3.14}$$

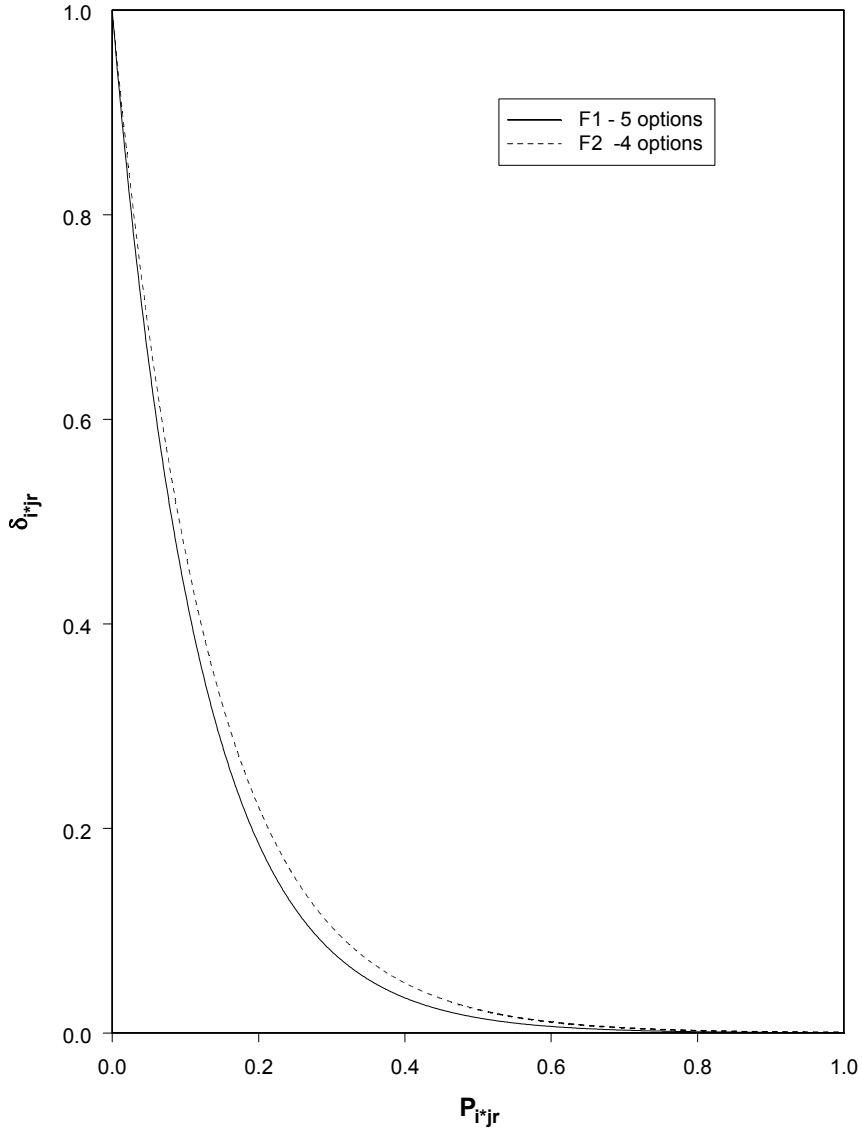


Figure 3.1: Graph of δ_{i^*jr} as a Function of P_{i^*jr} with $g = .25$ and $g = .20$.

In (3.14), the contribution of each item to the value of M^*_{jr} is 0 if the response of jr did not match that of s, 1 if the wrong response of jr matches that of s, and δ_{i^*jr} if the correct response of jr matches that of s. The value of M^*_{jr} is large if most of the incorrect responses of jr matches the wrong responses of s or if P_{i^*jr} is small and most of the correct responses of jr matches the correct responses of s. The larger the value of M^*_{jr} relative to the number of items, the stronger the evidence of answer copying.

Note that if there are no matching correct answers between s and jr, the second term in (3.14) sums up to zero and $M^*_{jr} = M_{jr}$. Hence, M_{jr} becomes a special case of M^*_{jr}. On the other hand, if there are no matching incorrect items but only matching correct answers, then $M_{jr} = 0$ and $M^*_{jr} = \sum_{i^*} \delta_{i^*jr}$. Thus, while M_{jr} is only sensitive to incorrect answer copying, M^*_{jr} is sensitive to both correct and incorrect answer copying.

The random variable M^*_{jr} is a nonnegative real-valued random variable. We treat M^*_{jr} as an integer by rounding it off to the nearest integer. Although some error is introduced by doing this, we expect that this will only have a minor influence on the effectiveness of the statistic. Like M_{jr}, we use the Poisson distribution for and the loglinear model to estimate its mean. We explored empirically the usefulness of the Poisson distribution to model M^*_{jr} using the G^2 statistics.

The S_2 index is then defined as

$$S_2 = \sum_{w=m^*_{c'c}}^{I} \frac{e^{-\widehat{\mu}_{c'}} \widehat{\mu}^w_{c'}}{w!}, \tag{3.15}$$

where $M^*_{c'c}$, with realization $m^*_{c'c}$, is the sum of the number of matching incorrect and weighted matching correct answers between c and s. The smaller the value of S_2, the more likely that answer copying occurred.

3.4 Method

3.4.1 Data Generation and Simulation of Copying

The data were simulated in the same way as in Sotaridona and Meijer (2002). Multiple choice test items with five options were considered. Test length was chosen to be 40 and 80 items and samples of 100 and 500 simulees were generated. Item parameters were chosen in accordance with the study by Wollack (1997). As described in Wollack (1997), the item parameters were estimated under the nominal response model using MULTILOG (Thissen, 1991) for an 80-item, 5-option English college placement test and a 40-item, 5-option

mathematics college placement test used at a large Midwestern research university. We draw the ability parameter, θ, from $N(0,1)$. Given the item and ability parameters, $P_{iv}(\theta)$ was computed based on the nominal response model. The item response was drawn randomly from $v = [1, 2, \ldots, V]$, each having probability of being drawn equal to $P_{i1}(\theta), P_{i2}(\theta), \ldots, P_{iV}(\theta)$, respectively. The source was drawn at random from a sample of simulees having ability percentile rank ranging from 40 to 90. In both 40- and 80-item tests, five percent copiers were selected randomly from the simulees with θ level below the θ level of the source. The percentage of items copied were 10, 20, 30, and 40.

Similar to Wollack (1997), copying was simulated by first randomly selecting a specified percentage of items from the copier and then altering the responses of c to match the responses of s on those items.

We crossed the three factors – sample size (2 levels), number of items (2 levels), and percentage of items copied (4 levels)–resulting in $2 \times 2 \times 4 = 16$ testing conditions. The dataset in each condition was replicated 100 times.

3.4.2 Type I Error and Detection Rates

A simulee was identified as a copier by $\overline{K}_2$, S_1, or S_2 index if the values were less than or equal to the level of significance α. The α levels were set to .0001, .0005, .001, .005, and .01; similar α levels used in Wollack (1997) with the exclusion of .05, .10 and .0025.

For the ω statistic, a simulee was identified as a copier when the value of ω was above the one-tailed critical value corresponding to the right tail of the standard normal distribution. ω was computed using the item and ability parameters that were used in the simulation. This was done because Wollack and Cohen (1998) showed that the Type I error rate of ω is not affected by using estimated item and ability parameters. As in Sotaridona and Meijer (2002), the copying indices were computed based on prior suspicion of a particular simulee copying from a specific source. The statistics were therefore used without adjustment for the α level.

To determine the empirical Type I error rate, we computed the proportion of noncopiers who were identified by the copying index as copiers. For datasets with 100 examinees, this computation was based on 9400 noncopiers (94 non-copiers per replication$\times$100 replications) and for datasets with 500 examinees, 47400 non-copiers (474 non-copiers per replication$\times$100 replications).

Likewise, the detection rate was computed as the proportion of true copiers classified as copier by $\overline{K}_2$, S_1, S_2, and ω. For datasets with 100

examinees, this computation was based on 500 true copiers (5 true copiers per replication×100 replications) and for datasets with 500 examinee on 2500 true copiers (25 true copiers per replication×100 replications).

3.5 Results

3.5.1 Adequacy of the Loglinear Model

The fit of the loglinear model in (3.9) was assessed using the G^2 statistic. The results were similar for M and M^* and also for the 40− and 80−item test so only the results for M^* and the 40−item test are presented and discussed here.

Figure 3.2 shows the scatter plots of 100 p-values ranked in increasing order for the 40-item test with 100 and 500 simulees and for different percentages of items copied. Remember that the null hypothesis being tested is that the loglinear model fits the data; large p-values therefore supports the null hypothesis.

The loglinear model fits the data in every situation simulated as reflected by the high p-values both for $J = 100$ and $J = 500$. For example, for $J = 100$, the minimum p-value for 10% copying is 0.332, for 20% copying it is 0.418, for 30% copying it is 0.182, and for 40% copying it is 0.481 (Figure 3.2a-d). For $J = 500$, all the p-values are almost 1 across four percentages of copying (Figure 3.2e-h).

3.5.2 Type I Error Rate

Figure 3.3 shows the empirical Type I error of ω, $\overline{K}_2$ (denoted as $K2$), S_1, and S_2 (denoted as $S1$ and $S2$) for different α-levels and across different combinations of sample sizes and test lengths. The solid line in the graph is a boundary line indicating perfect agreement between the nominal and empirical Type I errors. A copying index having Type I errors above the boundary line is liberal and below the boundary line is conservative in classifying a simulee as copier. An ideal copying index maintains its Type I error on or slightly below the nominal α level, but not too far below, otherwise, its detection rate will be reduced.

The S_2 index holds its Type I error for $J = 100$ (Figures 3.3 a-b) and tend to be slightly liberal for $J = 500$ (Figure 3.3c-d). The ω index on the other hand is slightly liberal in most cases for $J = 100$ and slightly conservative for $J = 500$. Both the S_1 and $\overline{K}_2$ were able to hold their Type I errors below the nominal levels. The empirical Type I error for S_1 and $\overline{K}_2$

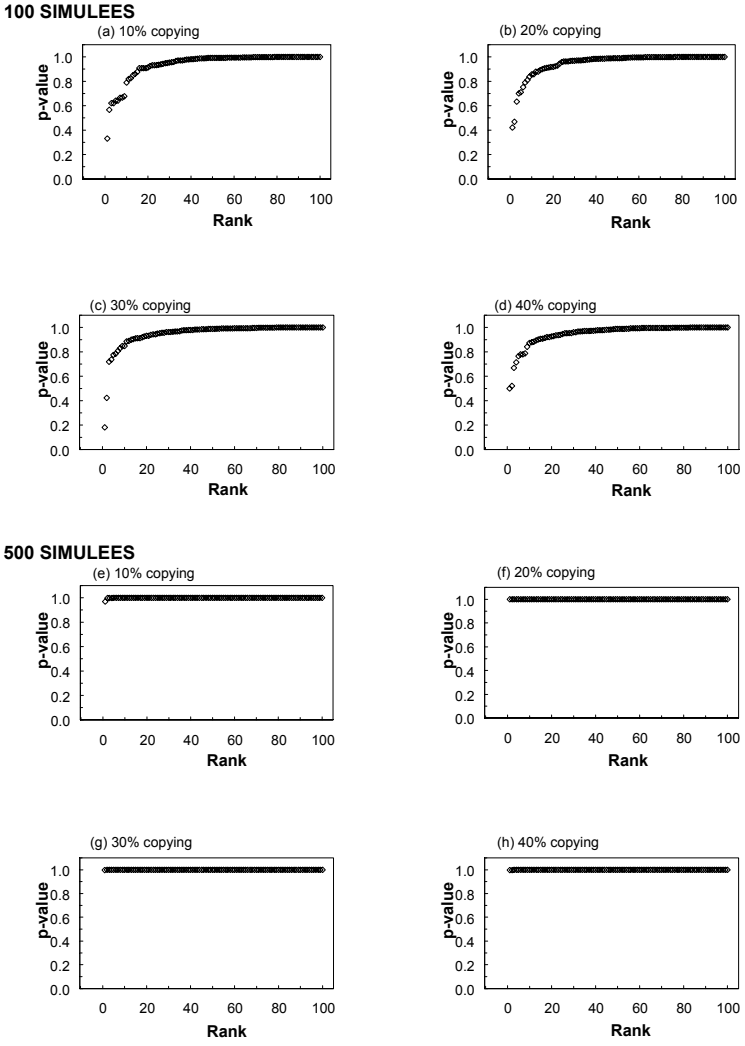


Figure 3.2: Scatter plots of 100 p-values of G^2 Statistics, Ranked in Increasing Order, for 40-item Test.

were, in most cases, lower than the Type I errors of S_2 and ω. S_1 was the most conservative index for $J = 100$ and $\overline{K}_2$ for $J = 500$.

3.5.3 Detection Rate

The detection rates for $\overline{K}_2$, S_1, S_2, and ω for different α-levels, percentages of copying, and test lengths are shown in Figure 3.4 for 100 simulees and Figure 3.5 for 500 simulees. The detection rates for all the indices increased with the percentage of copying. For example for 40 items and 100 simulees, the detection rates in Figure 3.4a (40% copying) are higher than the detection rates in Figure 3.4b (30% copying) which are both higher than that in Figure 3.4c (20% copying) and Figure 3.4d (10% copying). Similar trends were observed for other combinations of sample sizes and test lengths.

Consistent with the findings of Wollack (1997) and Sotaridona and Meijer (2002), the detection rate of ω increased with test length but not with sample size. For example, Figure 3.4 shows that for 100 examinees, the detection rate of ω was higher for the 80-item test than for the 40-item test. The same observation holds for 500 examinees (Figure 3.5). For a fixed test length, changing the sample size from 100 to 500 did not change the detection rate of ω (compare Figure 3.4 with Figure 3.5).

On the other hand, the test length, the sample size, or a combination of both test length and sample size affect the detection rates of $\overline{K}_2$, S_1 and S_2. In particular, increasing the test length (compare Figure 3.4 with Figure 3.5) or sample size (compare Figure 3.4a-d with Figure 3.4e-h, and Figure 3.5a-d with Figure 3.5e-h) resulted in increased detection rates.

Comparing the detection rates of the four indices, we should keep in mind that the empirical Type I errors are not exactly similar, though the differences are small. No index performs best in all testing conditions considered. The S_2 index has the highest detection rates for 20% and 10% copying regardless of test length and sample size (Figure 3.4c-d, g-h and Figure 3.5c-d, g-h), and for 30% or 40% copying with 40 items and 500 simulees (see Figure 3.5a-b).

Furthermore, for 30% or 40% copying, the detection rate for S_2 was approximately the same as the ω index for 40 items and 100 simulees (Figure 3.4a-b) and for 80 items and 500 simulees (Figure 3.5e-f), whereas the ω index has the highest detection rates for 80 items and 100 simulees (Figure 3.4e-f). The four indices are equally effective with almost 100% detection rates for $\alpha = .01$, 40% copying, and 80 items (see Figures 3.4e and 3.5e). In general, the $\overline{K}_2$ index had the lowest detection rate compared to the other indices.

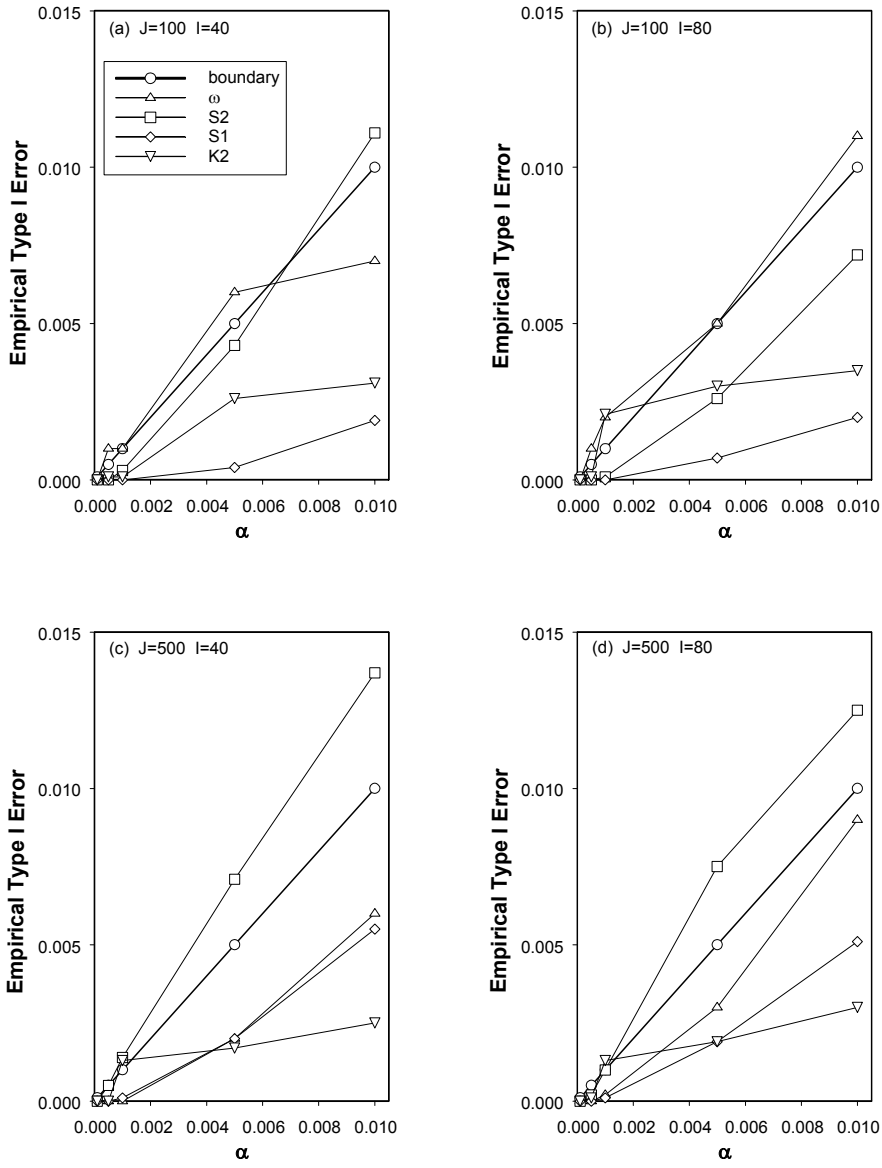


Figure 3.3: Nominal and Empirical Type I Error Rates as a Function of Simulee Size and Test Length.

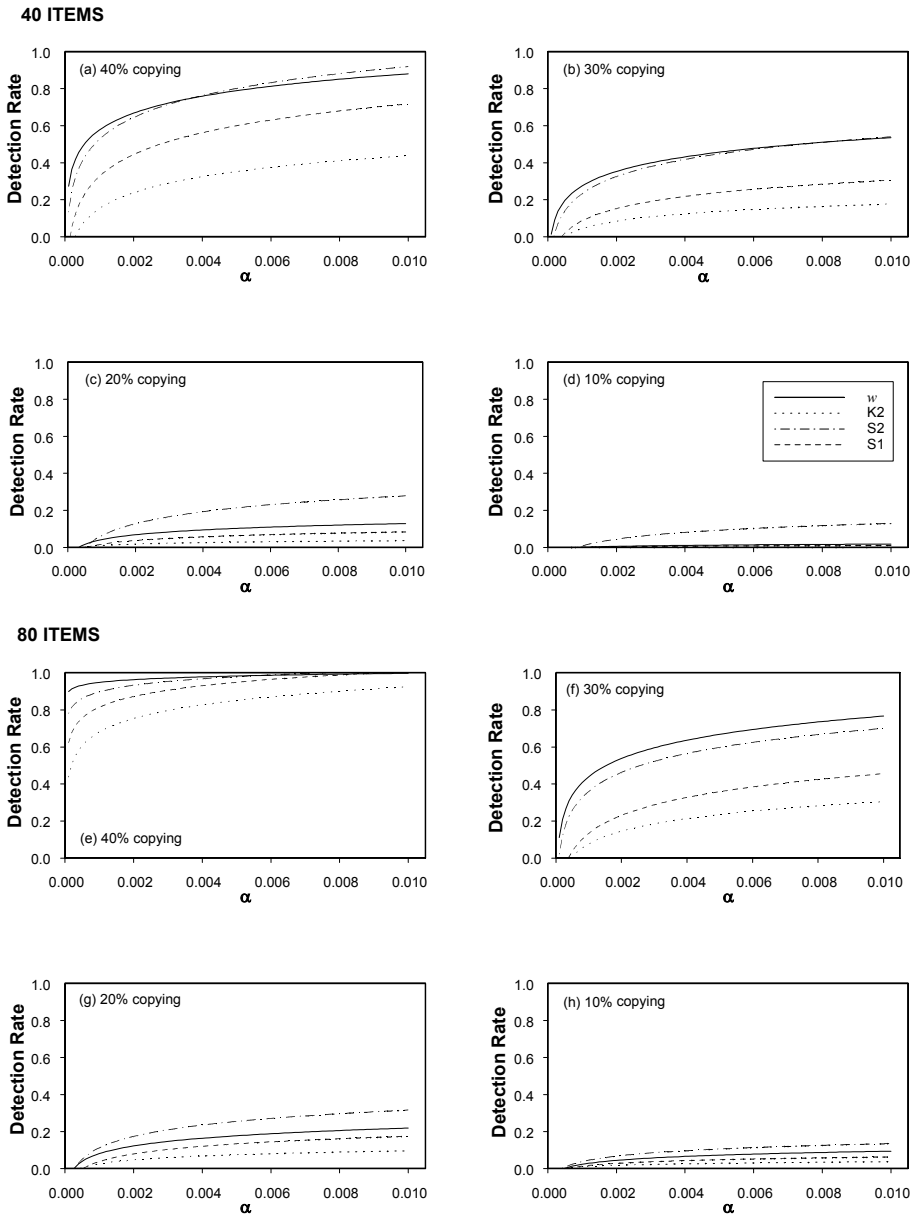


Figure 3.4: Detection Rates of ω, $\overline{K}_2$, S_1, and S_2 on a Test with 100 Simulees.

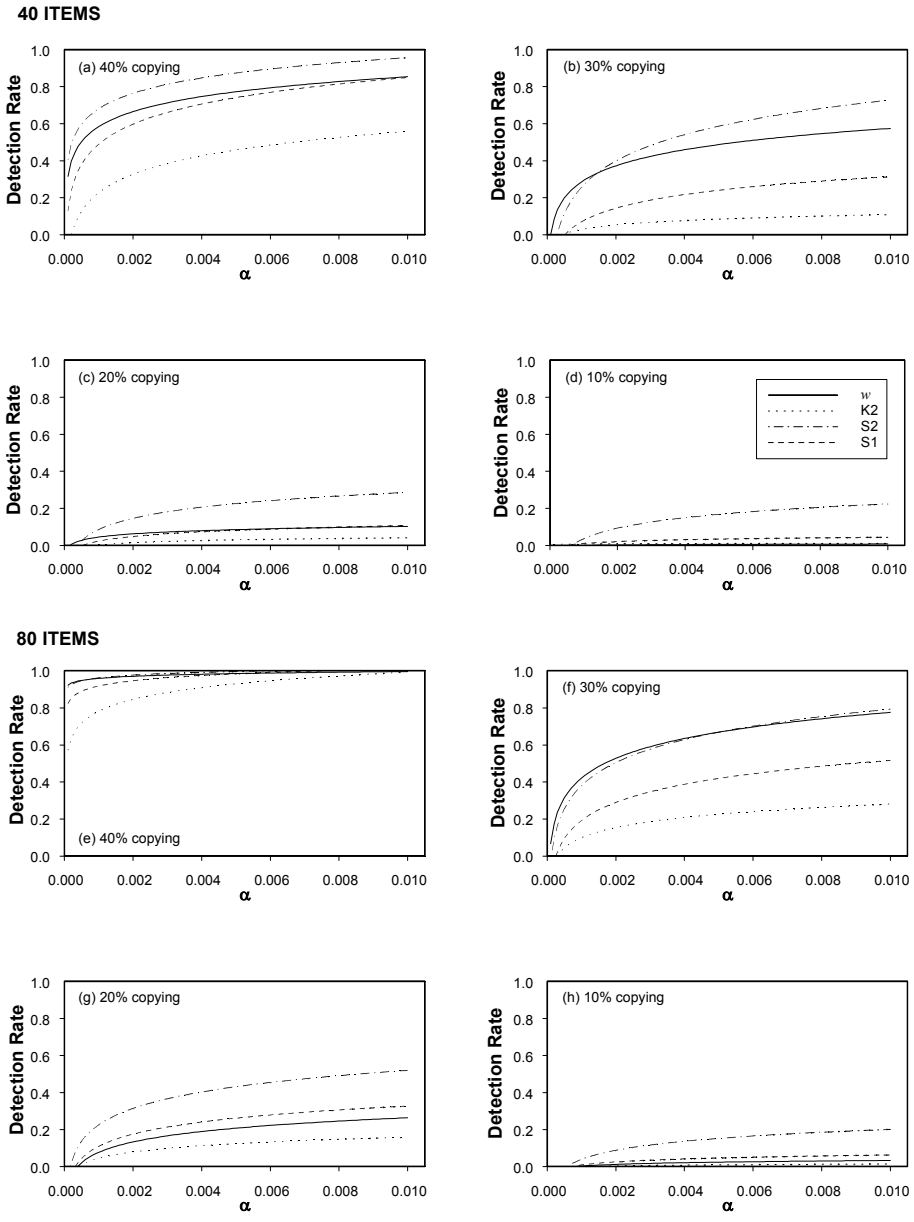


Figure 3.5: Detection Rates of ω, $\overline{K}_2$, S_1, and S_2 on a Test with 500 Simulees.

3.6 Discussion

We proposed the S_1 and the S_2 indices to detect answer copying as an alternative to the $\overline{K}_2$ index. In the S_1 index the Poisson distribution was used instead of the binomial distribution used in $\overline{K}_2$ for M. The S_2 index was proposed to overcome the limitation of $\overline{K}_2$ which is not sensitive to answer copying correct item scores. Crucial in using S_1 and S_2 is obtaining reliable estimates of the means of M and M^*. We used the loglinear model to estimate the mean. The Type I errors and detection rates of S_1 and S_2 were compared with the Type I errors and detection rates of $\overline{K}_2$ and ω.

Results suggest a good fit of the Poisson distribution for M and M^*. Using the Poisson distribution, instead of the binomial distribution, resulted in considerably higher detection rate for S_1 than for $\overline{K}_2$. The S_2 index, which incorporates information from the matching correct scores in addition to the matching incorrect–scores, resulted in a significant improvement in detection rate over S_1.

As shown in this study and in Sotaridona and Meijer (2002), if the item parameters in the nominal response model can be estimated reliably, ω seems to be the best choice for detecting answer copying because it is sensitive across all ability levels of the copier and can also be used to detect answer copying for small sample sizes. S_1 and S_2 cannot be used in this latter case. However**,** considering the computational simplicity and less restrictive assumptions imposed on S_2, the S_2 index may be a good alternative to use in practice. Results concerning the Type I errors of ω and $\overline{K}_2$ were also consistent with the results found in Sotaridona and Meijer (2002) which showed that the ω was slightly liberal at $J = 100$ whereas $\overline{K}_2$ was conservative.

The present study only considered five percent copiers and the items copied by the copiers were selected at random. The difference in the ability level of the source and the copier may affect the performance of a copying index. For future research, it might be interesting to study the Type I errors and detection rates of S_1 and S_2 for varying ways of answer copying and for different percentages of copiers, different percentage of correct answer copied, and different ability levels of the source.

Chapter 4

A Test Based on the Shifted Binomial

Abstract. A statistical test for the detection of answer copying on multiple-choice tests is presented. The test is based on the idea that the answers of examinees to test items may be the result of three possible processes: (1) knowing, (2) guessing, and (3) copying, but that examinees who do not have access to the answers of other examinees can arrive at their answers only through the first two processes. This assumption leads to a distribution for the number of matched incorrect alternatives between the examinee suspected of copying and the examinee believed to be the source that belongs to a family of "shifted binomials". Power functions for the tests for several sets of parameter values are analyzed. We show that an extension of the test to include matched numbers of correct alternatives would lead to improper statistical hypotheses.

This chapter has been accepted for publication as: van der Linden, W. J., & Sotaridona, L. S. (in press). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement.*

4.1 Introduction

Among the first to derive a statistical test to detect answer copying on multiple-choice tests were Angoff (1974) and Frary, Tideman, and Watts (1977). The g_2 index proposed by Frary et al. is an attempt to evaluate the number of matching alternatives between an examinee suspected to be a copier and another examinee believed to be the source against the expected number of matching alternatives. (For convenience, we will refer to these examinees just as "copier" and"source".) Two problems inherent working with such an index are obtaining the distribution of the index under the null hypothesis of no copying and evaluating the statistical power of the test based on it. Frary et al. attempted to solve the first of these problems by establishing a null model that assumes that the probability of selecting an alternative on an item is a certain function of the copier's number-correct score, the average number-correct score in the population, and the proportion of examinees in the population who selected the same alternative. Note that the first two quantities correct the probability of selecting an alternative for the examinee's ability relative to the abilities in the tested population. A correction of this type is needed to prevent confounding of an answer the examinee knows with correct answers that have been copied.

The K-index (Holland, 1996; Lewis & Thayer, 1998) is another attempt to correct for the examinee's ability. The index focuses only on the number of matching alternatives on the items that were answered incorrectly by the source. The null model is a binomial with a success parameter that is obtained by piecewise linear regression of the proportion of matching incorrect alternatives on the proportion-incorrect scores in a population of examinees. An alternative with quadratic regression is given by Sotaridona and Meijer (2002).

The most elaborate null model for a test to detect copying is the one on which Wollack's ω index is based (Wollack, 1997; Wollack & Cohen, 1998). Like the g_2 index, the ω index compares the observed number of matching alternatives against an estimate of the expected number. For the ω index. the estimate is derived under the assumption of the nominal response model (Bock, 1997) for the probabilities of selecting an alternative on an item for an examinee who does not copy. Note that the use of the nominal response model automatically involves conditioning of the probabilities of choosing an alternative on the examinee's ability.

In spite of the attempts to condition on the examinee's ability, a fundamental feature of all three tests is their dependence on the distribution of the item scores in the population of examinees. From a statistical point,

this is generally advantageous. The information in the response vector of a single examinee is limited. Access to collateral information can lead to procedures that are more robust as well as to inferences that tend to become more successful on average for the population. However, at the level of an individual examinee suspected of answer copying, the use of population-based statistical tests may be unfair. In principle, such tests can result in a statistically significant proof of answer copying for a pair of examinees in one population but acceptance of the null hypothesis of no copying if their responses vectors were included in the data set for another population, for example, a population taking the same test at a later occasion or at another site.

We present a statistical test to detect answer copying on multiple-choice tests that can be used when any reference to a population of examinees is undesirable. Obviously, statistical tests for answer copying cannot be derived without assumptions. The difference between the current test and existing tests lies therefore not in the fact that it is based on assumptions, but in the nature of them. The only assumptions we make are about the response behavior of the individual examinee suspected of copying. No assumptions are made about the existence of a score distribution in a population; neither is anything assumed about the response probabilities with which the examinee who may have served as a source to the copier answered the items.

In essence, the assumptions are based on the idea that an examinee who has access to the answers of a source can arrive at his/her own answer through three different processes: (1) knowing, (2) guessing, or (3) copying. Examinees who do not have access to a source can produce answers only through the first two processes. We do not claim that these assumptions are universally true. In fact, at best they will only hold approximately for an examinee suspected of copying. The approach we follow, however, is typical of model-based inference; we derive a test from a set of assumptions expected to hold approximately, study its behavior, and then assess if it will remain useful under likely violations of the assumptions. We return to this issue later in this paper.

4.2 Derivation of the Test

Like the K-index, the test focuses on the items for which the source has an incorrect answer. An extension of the test to include items with correct answers by the source will lead to improper statistical hypotheses (see below).

4.2.1 Assumptions

The test is derived from the following assumptions about the behavior of the copier on the items the source has answered incorrectly: First, if an examinee knows an item, he/she gives a correct answer. This assumption implies that if an examinee has access to the source but discovers that this examinee's answer is incorrect, he/she does not copy but gives his/her own answer. Second, if an examinee does not know an item but has access to the source, he/she accepts the answer by the source and copies. Third, if an examinee does not know an item and does not have access to a source, he/she guesses blindly among the response alternatives. Thus for each item incorrect by the source, we have three possible true states in which the copier can be, each characterized by a different probability of choosing the same alternative the source has chosen.

We use the following notation to present these probabilities. Let $i = 1, ..., I$ denote the items in the test and $a = 1, ..., k$ the response alternatives for these items. In addition, indices s and j are used for the source and the copier, respectively. The alternatives chosen by these two examinees on item i are denoted by random variables U_{si} and U_{ji}. The set of items for which s has chosen an incorrect alternative is denoted as W_s. The size of this set is denoted as w_s. Finally, a (random) indicator variable I_{jsi} is used to identify the items for which examinees s and j have chosen the same alternative. That is,

$$I_{jsi} = \begin{cases} 1 & \text{if} \quad U_{ji} = U_{si} \\ \\ 0 & \text{if} \quad U_{ji} \neq U_{si}. \end{cases}$$

The three possible probabilities that examinee j will choose the same alternative on the items in W_s as s are the following:

$$\Pr(I_{jsi} = 1) = \begin{cases} 0 & \text{if} \quad j \text{ knows the answer on } i \in W_s, \\ \\ k^{-1} & \text{if} \quad j \text{ guesses blindly on } i \in W_s, \\ \\ 1 & \text{if} \quad j \text{ copies from } s \text{ on } i \in W_s. \end{cases} \tag{4.1}$$

4.2.2 Hypotheses

The hypothesis to be tested is that j did not copy any of the items in W_s. We suggest testing this hypothesis against the alternative that j copied the answers for some of the items in W_s that he/she did not know. Observe that this alternative is less extreme than the hypothesis that j copied all

items in W_s and therefore covers a larger number of potential cases of answer copying. Under the current alternative hypothesis, it is still possible that j actually knows some of the items in W_s and for this reason did not copy them or that she/he did not have access to the answers by s for all of the items in W_s.

Let κ_{js} be the number of items in the set W_s examinee j knows and γ_{js} the number in this set examinee j copied from s. More formally, at the level of the set of items W_s, the hypothesis to be tested is:

$$\mathrm{H}_0 : \Pr(I_{jsi} = 1) = \begin{cases} 0 & \text{for } \kappa_{js} \text{ items in } W_s, \\ k^{-1} & \text{for } w_s - \kappa_{js} \text{ items in } W_s, \end{cases} \tag{4.2}$$

against

$$\mathrm{H}_1 : \Pr(I_{jsi} = 1) = \begin{cases} 0 & \text{for } \kappa_{js} \text{ items in } W_s, \\ k^{-1} & \text{for } \omega_s - \kappa_{js} - \gamma_{js} \text{ items in } W_s, \\ 1 & \text{for } \gamma_{js} \text{ items in } W_s, \end{cases} \tag{4.3}$$

with $\kappa_{js} \geq 0$, $\gamma_{js} > 0$, and $\kappa_{js} + \gamma_{js} \leq w_s$. The null hypothesis follows thus upon substituting $\gamma_{js} = 0$ in (4.3).

4.2.3 Distribution of Matching Incorrect Alternatives

The proposed test statistic is the number of matching incorrect alternatives between j and s on the items in set W_s:

$$M_{js} = \sum_{i \in W_s} I_{jsi}. \tag{4.4}$$

Both hypotheses imply distributions of M_{js} belonging to a family with probability function $p(m; w_s, \gamma_{js}, \kappa_{js}, k) \equiv p$

$$p = \begin{cases} 0 & \text{for } m < \gamma_{js}, \\ \binom{w_s - \kappa_{js} - \gamma_{js}}{m - \gamma_{js}} k^{-(m-\gamma_{js})} (1 - k^{-1})^{w_s - \kappa_{js} - m} & \text{for } \gamma_{js} \leq m \leq w_s - \kappa_{js}, \\ 0 & \text{for } w_s - \kappa_{js} < m, \end{cases} \tag{4.5}$$

with $\kappa_{js} \geq 0$, $\gamma_{js} \geq 0$, and $\kappa_{js} + \gamma_{js} \leq w_s$. The definition of this family follows from the fact that if j copies γ_{js} answers from W_s, the probabilities of observing numbers of matches smaller than γ_{js} are each equal to zero. Likewise, if j knows κ_{js} items in W_s, the probabilities of observing a number of matches larger than $w_s - \kappa_{js}$ are each equal to zero. For the subset of

$w_s - \kappa_{js} - \gamma_{js}$ items that j does not know and for which he/she has not copied any answer, however, the number of matches follows a binomial distribution with success parameter k^{-1}. Observe that the probability of $M_{js} = \gamma_{js}$ belongs to the compound event of j copying γ_{js} items and guessing none of the alternatives the source has chosen. Likewise, the probability of $M_{js} = w_s - \kappa_{js}$ belongs to the compound event of j copying γ_{js} items, knowing κ_{js} items, and guessing the alternatives the source has chosen on all remaining items.

The function in (4.5) can be presented more compactly as

$$p = \binom{w_s - \kappa_{js} - \gamma_{js}}{m - \gamma_{js}} k^{-(m-\gamma_{js})}(1 - k^{-1})^{w_s - \kappa_{js} - m} I_{\{\gamma_{js}, w_s - \kappa_{js}\}}(m), \quad (4.6)$$

where $I_{\{\gamma_{js}, w_s - \kappa_{js}\}}(m)$ is an indicator function that is equal to 1 if m is one of the integers $\gamma_{js}, \gamma_{js} + 1, ..., w_s - \kappa_{js}$ and equal to 0 otherwise. Because $p(m; w_s, \gamma_{js}, \kappa_{js}, k)$ is nonzero for $m \in \{\gamma_{js}, w_s - \kappa_{js}\}$, this function indicates the support of the family of distributions in (4.6). In spite of the presence of the binomial expression in the definition of (4.6), the family is *not* the binomial over the range of possible values of M_{js}. We will refer to this family as the "shifted binomial", because it can be viewed as a binomial with its support shifted from $\{0, w_s\}$ to $\{\gamma_{js}, w_s - \kappa_{js}\}$. The size of the shift is a critical quantity because it depends both on the (unknown) number of items j knows as well as the number j has copied.

Monotone Likehood Ratio

The binomial family has a monotone likelihood ratio, that is, for a fixed number of trials it holds for each pair of values for its success parameter $\pi_1 > \pi_0$ that the likelihood ratio $L(\pi_1)/L(\pi_0)$ is an increasing function of the number of successes. The family in (4.6) has a known success parameter but unknown additional parameters γ_{js} and κ_{js}. To proof the claims that the test in this paper is right sided and uniformly most powerful below, we will show that (4.6) has an increasing likelihood ratio with respect to γ_{js} but a decreasing ratio with respect to κ_{js}.

Actually, we will only use the fact that the ratio of (4.6) for $\gamma_{js} > 0$ and $\gamma_{js} = 0$ increases in m. For any value of κ_{js}, simplifying, omitting constants, and cancelling factors in this ratio, the ratio can be shown to lead to

$$\frac{m!}{(m - \gamma_{js})!}, \quad (4.7)$$

which is increasing in m. Likewise, we need the result for the ratio for $\kappa_{js} > 0$ and $\kappa_{js} = 0$. For any value of γ_{js}, this ratio can be reduced to

$$\frac{(w_s - m)!}{(w_s - \kappa_{js} - m)!}, \tag{4.8}$$

which is decreasing in m.

4.2.4 Statistical Test

Under the distribution in (4.6), the two hypotheses in (4.2) and (4.3) simplify to

$$\mathrm{H}_0 : \gamma_{js} = 0 \tag{4.9}$$

and

$$\mathrm{H}_1 : \gamma_{js} > 0. \tag{4.10}$$

The null distribution under which the (right-sided) test of the hypothesis in (4.9) has to be conducted still depends on the unknown parameter κ_{js}. We propose to conduct the test under the auxiliary assumption that j did not know any of the items in W_s, that is, $\kappa_{js} = 0$. This assumption gives us a test that tends to be more conservative than the one actually needed. This claim is easily be seen to hold as follows: Families with a monotone likelihood ratio are also stochastically ordered (Casella & Berger, 1990, page 390). From (4.8), it therefore follows that the upper tail of the distribution for $\kappa_{js} = 0$ is always further to the right than the upper tails of the distributions for $\kappa_{js} > 0$. As a result, ignoring the discreteness of m, setting $\kappa_{js} = 0$ always results in a critical value for the test larger than the one needed for the (unknown) true value of κ_{js} at the nominal level of significance.

We feel the auxiliary assumption is permitted because it does not harm the copier in any way. The one who may have to pay a price for this assumption is the testing agent because of a loss of power of the test to detect answer copying. For quantitative results on the extent to which the critical value of the test is larger than actually needed as well as the differences in power resulting from this increase, see the empirical section later in this paper.

A (nonrandomized) test of the hypothesis that j did not copy any answer against the alternative that j copied the answers of some of the items in W_s with nominal significance level not larger than α has as critical value for the test statistic M_{js} in (4.4) the smallest value of m^* for which the distribution in (4.6) under the null hypothesis $\gamma_{js} = 0$ yields

$$\Pr(M_{js} \geq m^*) \leq \alpha. \tag{4.11}$$

4.2.5 UMP Test

For a statistical test it is desirable to be uniformly most powerful (UMP) at the level of significance chosen. From the Karlin-Rubin theorem (e.g., Casella & Berger, 1990, sect. 8.3.2) it follows that the test above is a UMP test with level associated with the critical value in (4.11) provided the family in (4.6) has a monotone likelihood ratio in M_{js} and M_{js} is a sufficient statistic for the number of items copied, γ_{js}. We have already proved the first property in (4.7). The fact that M_{js} is a sufficient statistic for γ_{js} follows from the well-known factorization criterion. The factor

$$(1 - k^{-1})^{w_s - \kappa_{js} - m}$$

in (4.6) is independent of γ_{js}, whereas its remaining part is dependent on γ_{js} and m.

It is instructive to compare this result with those for a test of a point hypothesis for the success parameter in a regular binomial family, which also is UMP. In the current case, (4.6) is not the regular binomial and the parameter of interest is not a success parameter but a parameter that defines both the support of the distribution and the number of Bernoulli trials on which it is based. Observe also that (4.6) is not UMP with nominal level α but with the actual level of significance associated with (4.11). An exact level α test is possible only for a randomization version of (4.11).

Finally, we emphasize that the result above holds for the test in (4.11) that is based on the assumption $\kappa_{js} = 0$, but that it has not been shown that the test of (4.9) is UMP for an unknown value of κ_{js}. The impact of this parameter on the power of the test will be evaluated empirically in the next section.

4.2.6 Comparison with K Index

Both the null distribution of the K index (Holland, 1996) and the distribution in (4.6) for $\gamma_{js} = 0$ are related to the regular binomial family, but each in a different way. The null distribution of the K index is a parametric binomial; its success parameter is modeled as a function of other parameters that (partially) characterize the score distribution in a population of examinees. A sample from the score distribution is used to estimate these unknown parameters. On the other hand, the distribution in (4.6) is a binomial with a shift in its support, where the direction of the shift depends on two unknown parameters: the number of items the examinee has copied (γ_{js}) and the number the examinee actually knows (κ_{js}). The statistical

test based on this distribution does not involve any parameter estimation. Instead, κ_{js} is eliminated by the introduction of an auxiliary assumption that only leads to a more conservative test, whereas γ_{js} is eliminated under the null hypothesis of no copying.

The reason for the difference between the two distributions is that both tests are based on different assumptions. The K index is based on the assumption that, conditionally on W_s, the response behavior of examinees who do not cheat can be described by a series of Bernoulli trials with a probability of success given by piecewise linear regression of the proportion of matches on the proportion of incorrect scores in the population. Basically, this approach amounts to the idea of random sampling of examinees, parallel items, and curve fitting to obtain parameter estimates. On the other hand, the null distribution in (4.6) does not assume random sampling of examinees, assumes only those items on which the examinee guesses to be parallel, and derives the success parameter from the number of response alternatives on the item.

It is not our intention to discriminate between assumptions that are true and false. In fact, none of the assumptions on which these two tests rest is ever entirely true. The power of a model-based approach is that by making all assumptions explicit, we know that an inference can only be wrong if it violates one of these assumptions. More useful questions, therefore, are: How robust are the inferences with respect to possible violations of the assumptions? Are errors in inferences due to violations in a direction that harms any of our practical conclusions? As for this last question, our prejudice is that a statistical test of cheating that becomes more conservative due to a violation of any of its assumptions is to be preferred over one that becomes more liberal. The same point was made by Holland (1996) in his discussion of violations of the assumptions underlying the K index. We already used it to motivate the auxiliary assumption $\kappa_{js} = 0$ above, will also use it in the power analyses in the next section, and return to it in the discussion at the end of this paper.

4.3 Power of the Test

The actual power of the test above is a function of the unknown number of items j has copied from s, γ_{js}. The shape of the power function depends on (1) the number of alternatives per item, k, (2) the number of items s has incorrect, w_s, (3) the significance level chosen for the test, α, and (4) the number of items the examinee knows, κ_{js}.

We first present a set of power functions for the case $\kappa_{js} = 0$ for $k = 2, ..., 5$, $w_s = 20, 30, 40, 50$, and significance level $\alpha = .05$. The functions were calculated by first finding the critical value m^* for $\alpha = .05$ in (4.11) under the distribution given by (4.6) with $\gamma_{js} = 0$ and then calculating the probabilities $\Pr\{M_{js} \geq m^*\}$ under the same distribution for $\gamma_{js} = 0, ..., w_s$. The choice of these parameter values serves only as an example. In particular, the choice $\alpha = .05$ is conventional and should not be taken to imply any recommendation with respect to the control of Type I error in an actual test security context (for a discussion of the selection of α in such a context, see Lewis & Thayer, 1998). Because the power functions are easy to calculate, we encourage testing agencies to calculate them for their own choice of parameter values in actual test security checks.

The power functions are presented in Figure 4.1. From these functions it is clear that the test has considerable power to detect copying on multiple-choice tests, particularly if the number of response alternatives per item, k, goes up. This result confirms the result derived earlier, namely that the test is uniformly most powerful. But even for a test for three-choice items, the power to detect copying is already perfect if the examinee has copied approximately half of the items in W_s for $w_s = 20$ or one third for $w_s = 50$.

We noted earlier that the auxiliary assumption of $\kappa_{js} = 0$ leads to a test that tends to be conservative. Figure 4.2 illustrates the effect of the assumption of $\kappa_{js} = 0$ on the critical value of the test for the same sets of parameter values as in Figure 4.1 ($k = 2, ..., 5$; $w_s = 20, 30, 40, 50$; $\alpha = .05$). Also, it has already been shown that the current test is conservative; its actual level of significance in an application is likely to be lower than its nominal value.

The curves in Figure 4.2 show the critical values as a function of the number of items j knows, κ_{js}. For example, the lower-left plot shows that for a test with four-choice items ($k = 4$), the assumption $\kappa_{js} = 0$ leads to a critical value for the test equal to $m^* = 17$. If the examinee actually knows $k_{js} = 10$ of the 40 items in the set W_s, however, the critical value could have been lowered to $m^* = 14$ to realize the nominal significance level of $\alpha = .05$. Except for small horizontal pieces, which are due to the discreteness of test statistic M_{js}, all curves decrease with κ_{js}. This feature reflects the fact that the proposed statistical test is generally conservative, unless the assumption $\kappa_{js} = 0$ happens to be true, in which case it is exact.

We also noted that the price for using a conservative test is paid not by the examinee but by the testing agency in the form of less than optimal

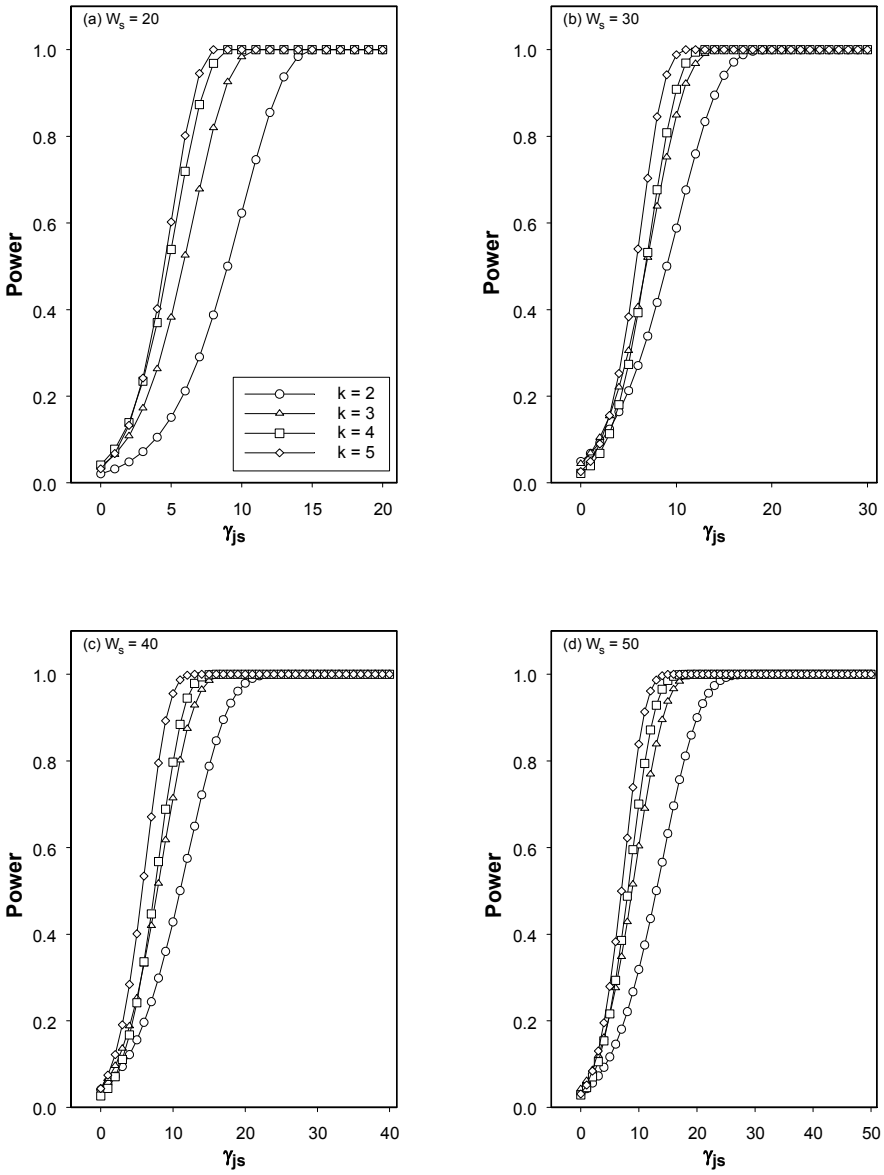


Figure 4.1: Power Functions for $k = 2, ..., 5$ and $w_s = 20, 30, 40, 50$ at Significance Level $\alpha = .05$.

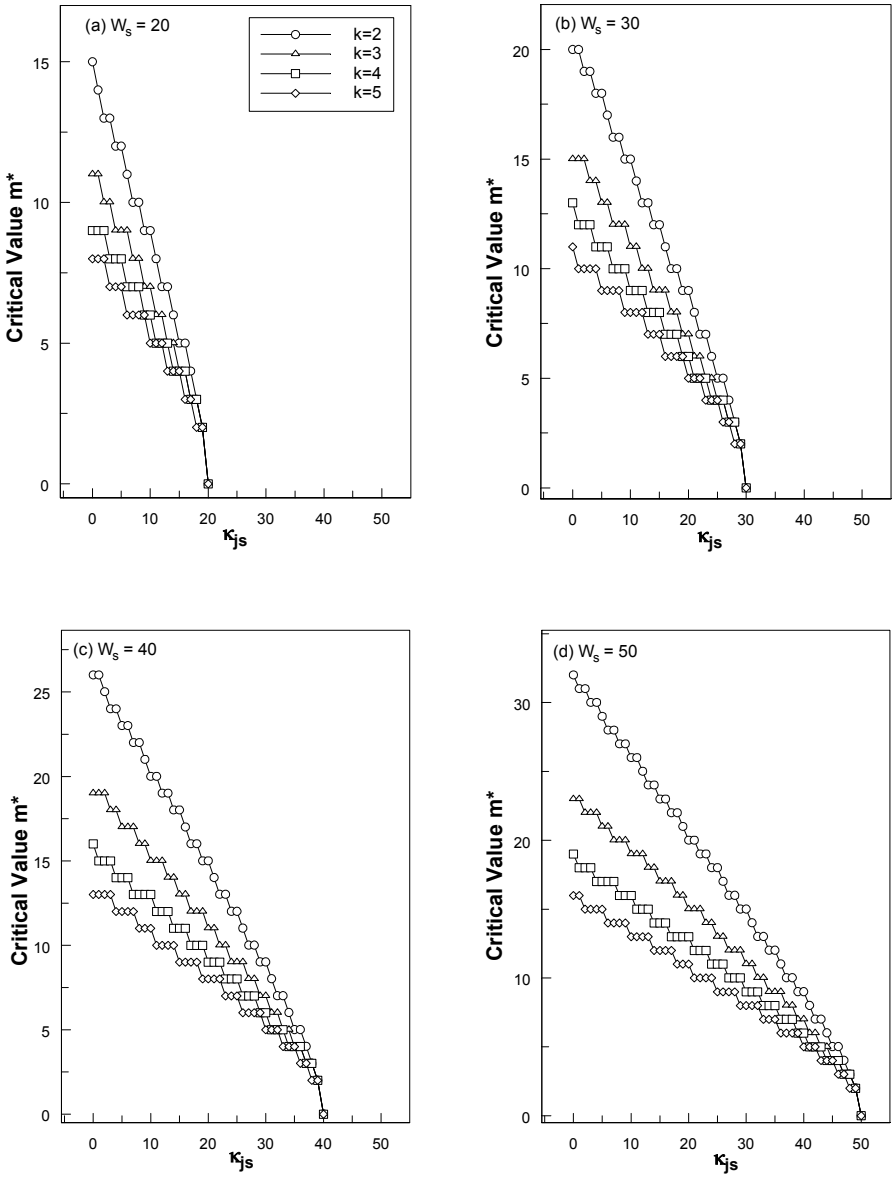


Figure 4.2: Critical Values as a Function of κ_{js} for $k = 2, ..., 5$ and $w_s = 20, 30, 40, 50$ at Significance Level $\alpha = .05$.

power to detect cheating. Figures 4.3-4.6 show how much larger the power of the test would have been if we had known the true value of κ_{js}. The curves in these figures show the increase in power relative to the power functions in Figure 4.1. That is, the increase in power was calculated as the difference between the power of the test for the true value of κ_{js} and $\kappa_{js} = 0$ divided by the power for $\kappa_{js} = 0$, and the result was plotted as a function of γ_{js}. Figures 4.3-4.6 show these functions for the same sets of parameter values as in the previous two figures ($k = 2, ..., 5$; $w_s = 20, 30, 40, 50$; $\alpha = .05$) and a selected set of values κ_{js}.

Each panel in these figures shows approximately the same pattern, which can be summarized as follows. First, knowing the true value of κ_{js} would lead to an increase of power only for small values of γ_{js}. Second, the increase would be larger, the smaller the number of alternatives per item, k. These two findings are consistent with the results in Figure 4.1 that show that the power curves are nearly equal to 1.0 for larger values of γ_{js} but approach this state at a somewhat lower rate for items with fewer alternatives. Once the power is close to one, there is hardly any space for improvement left. Third, the increase in power is generally larger for large values of κ_{js}, with an exception at the smallest values of γ_{js}, where for some of the larger values of κ_{js} the assumption $\kappa_{js} = 0$ actually appears to result in a small increase in power. These exceptions are due to the discrete nature of the null distribution of the test and the definition of the critical value in (4.11). For larger values of κ_{js} the actual level of Type I error can become smaller than α, and hence, for these values of κ_{js}, the power of the test can become lower than for $\kappa_{js} = 0$. A randomized version of the test would not suffer from this problem, but the use of randomization in a statistical test to detect cheating on multiple-choice tests does not seem ethical. Fourth, for smaller values of κ_{js} the power of the test appears to be remarkably robust and the assumption of $\kappa_{js} = 0$ does not involve much loss of power over the entire range of values of γ_{js}.

The proper way to use information as in Figures 4.1-4.6 in an actual test security check is (1) to identify the size of the set of items the source has wrong, w_s, (2) choose an appropriate significance level, α, (3) calculate the power function for the number of alternatives per item, k, for this value of w_s as in Figure 4.1, (4) calculate the function of the critical value m^* as in Figure 4.2 for the chosen values of α, w_s, and k to find out how much smaller the critical value should have been for the various possible numbers of items the examinee knows, κ_{js}, and (5) to produce a plot for the actual value of k as in Figures 4.3-4.6 to determine how much larger the power would have been if we had known κ_{js}.

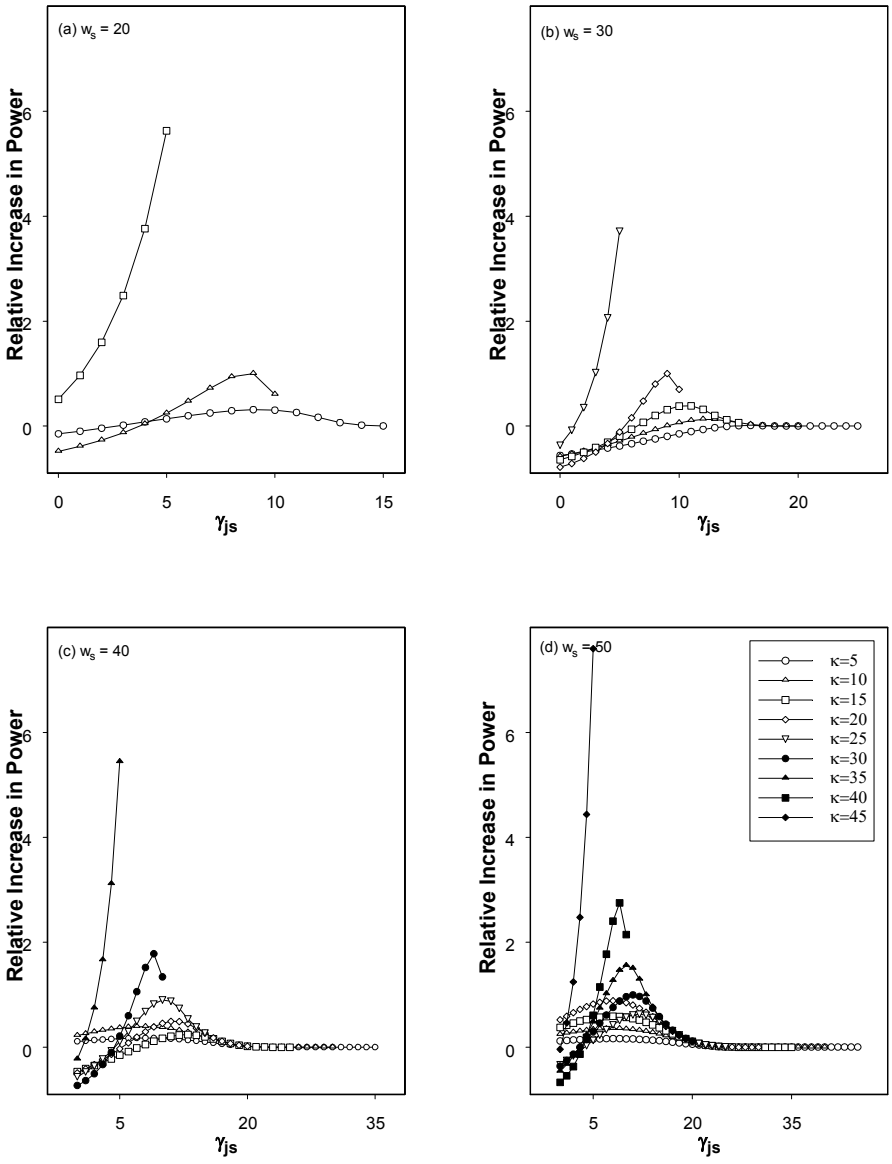


Figure 4.3: Relative Loss of Power Due to the Number of Items Known by the Examinee, κ_{js}, for $w_s = 20, 30, 40, 50$ at Significance Level $\alpha = .05$ and for $k = 2$.

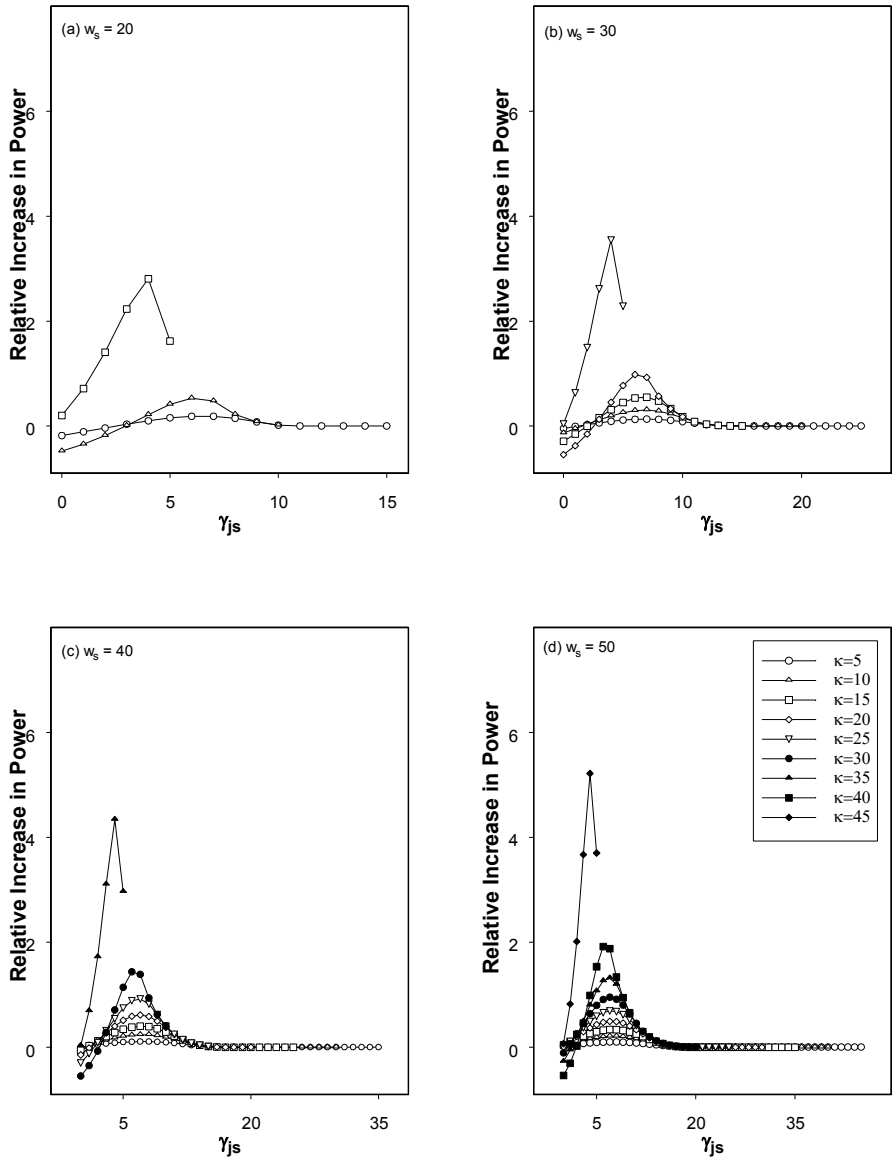


Figure 4.4: Relative Loss of Power Due to the Number of Items Known by the Examinee, κ_{js}, for $w_s = 20, 30, 40, 50$ at Significance Level $\alpha = .05$ and for $k = 3$.

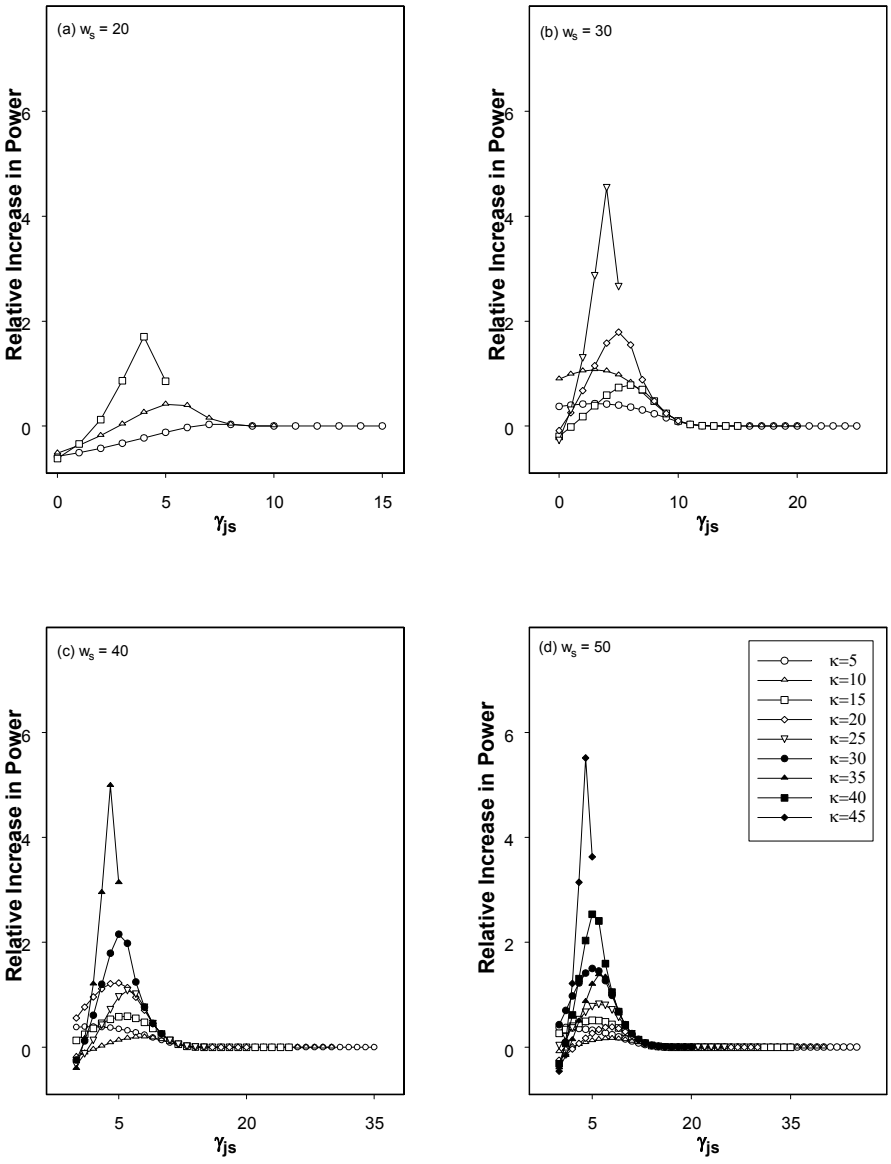


Figure 4.5: Relative Loss of Power Due to the Number of Items Known by the Examinee, κ_{js}, for $w_s = 20, 30, 40, 50$ at Significance Level $\alpha = .05$ and for $k = 4$.

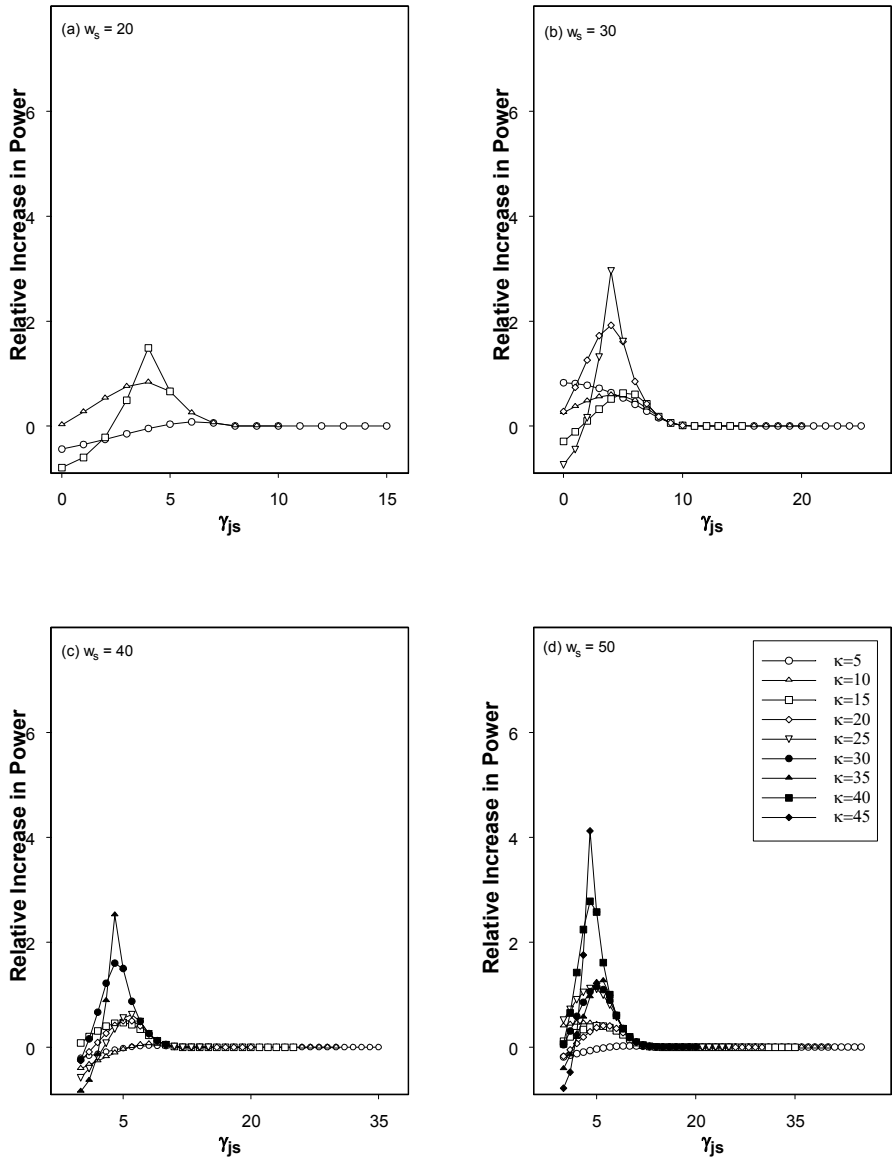


Figure 4.6: Relative Loss of Power Due to the Number of Items Known by the Examinee, κ_{js}, for $w_s = 20, 30, 40, 50$ at Significance Level $\alpha = .05$ and for $k = 5$.

4.4 Discussion

It can be asked if the test could be improved by having the statistic in (4.4) also include the items on which the source chooses the correct alternative. Just like the g_2 and w index, this option would allow the statistical test to derive its power from all items in the test rather than only those that s happens to answer correctly.

Let R_s be the subset of items the source has correct and J_{sji} an indicator variable for the event of a matching correct answer by s and j on item i. Analogous to (4.1), for the items in this set a copier can be in any of three possible true states, with the probabilities of a matching correct answer given by:

$$\Pr(J_{jsi}=1)=\begin{cases} 1 & \text{if } \; j \text{ knows the answer on } i \in R_s \\ k^{-1} & \text{if } \; j \text{ guesses blindly on } i \in R_s \\ 1 & \text{if } \; j \text{ copies from } s \text{ on } i \in R_s. \end{cases} \tag{4.12}$$

Unfortunately, the probabilities for the events of j knowing and copying the answer are equal. Extending the test proposed in this paper with the items in R_s would therefore result in a test confounding the difference between the events of j copying the answers s has correct and j knowing them.

In the current framework, the only possibility to further improve the power of the proposed test is to get more information about the number of items the examinee actually knows, κ_{js} from another source. For example, in a setting where an examinee retakes a test and shows an unusual increase in test scores, it may be possible to infer a lower bound to κ_{js} from the first test. Figures 4.2 and 4.3 show that, particularly for items with few alternatives or sources that have only a small number of items incorrect, conducting the test not at κ_{js}=0 but at a lower bound to κ_{js} deliberately chosen to be conservative is likely to result in an increase in power that should not be disregarded.

The statistical test in this paper is based on the assumption that an examinee who has no access to a source and does not know an item guesses. This assumption is widely used in test theory. It leads to the following probability statement

$$\Pr\{C\}=\Pr\{C \mid K\}\Pr\{K\}+\Pr\{C \mid G\}\Pr\{G\},$$

where C, K, and G represent the events of producing a correct answer to the item, knowing the answer, and guessing the answer. This statement implies

the well-known representation of the 3-parameter logistic (3PL) model in item response theory (Birnbaum, 1968):

$$\Pr\{C\} = \Pr\{C \mid G\} + [1 - \Pr\{C \mid G\}] \Pr\{K\},$$

with $\Pr\{K\}$ the response probability given by the 2PL model. It also underlies the correction for guessing on multiple-choice items known as "formula scoring" (Lord & Novick, 1968, sect. 14.3).

The truth of the probability statement above cannot be denied, but a critical issue is the determination of the probability of guessing the item correctly. In the 3PL model, the probability of guessing correctly is estimated from actual response data, whereas in the current paper, just as in the correction for guessing, it is assumed that the examinee guesses among the alternative with equal probabilities, k^{-1}.

This assumption has been criticized for two possible types of violations: (1) examinees may recognize some of the incorrect alternatives as wrong and guess blindly among the remaining alternatives or may have information that helps them to guess the correct alternative with a higher probability (partial knowledge) and (2) they may be attracted to an incorrect alternative and choose it with a probability larger than k^{-1} (negative knowledge). We evaluate the impact of these violations. In doing so, our point of reference is that, as discussed earlier, likely violations from the auxiliary assumption $\kappa_{js} = 0$, will always introduce a tendency to make the test conservative.

Generally, partial knowledge implies (1) a stronger tendency to choose the correct alternative and/or (2) a tendency to a decrease in the number of incorrect alternatives the examinee chooses from. The first tendency results in a larger expected number of items the examinee has correct relative to an examinee who guesses blindly. The net effect is equivalent to an increase in the number of items the examinee knows, κ_{js}, and hence a further increase in the conservativeness of the test. As for the second tendency, as shown by the plots in Figure 4.2, if, for a fixed value of κ_{js}, the number of alternatives decreases, the critical value of the test increases and the actual critical value has been set too low. We have thus one tendency to make the test more conservative and another to make it less conservative than the degree of conservativeness introduced by a likely violation of the assumption $\kappa_{js} = 0$.

Violations due to negative knowledge are likely to have a more uniform effect. Negative knowledge implies (1) a weaker tendency to choose the correct alternative and/or (2) a tendency to a decrease in the number of incorrect alternatives the examinee chooses from. Both implications point at a tendency to make the test less conservative.

Negative knowledge is thus a larger potential threat to the test in this paper than partial knowledge. However, as for the net effect of all these tendencies, we should remember that for a regular test with discrete items each addressing a different problem, the responses to the items are conditionally independent given the examinee. The same thus holds for the presence of negative or partial knowledge of the items. If the set W_s is not too small, the effects of negative and partial knowledge can thus be expected to average out. This does not hold for the tendency to conservativeness due to violations of the auxiliary assumption $\kappa_{js} = 0$.

As a general measure, it is always possible to make the test more conservative by choosing a smaller level of significance than the one we are actually interested in. In addition, we should remind ourselves of the fact that a proof of the guilt of individual examinees suspected of copying should never be based on a statistical test only but always be corroborated by evidence from other sources.

Chapter 5

A Test Based on Statistic Kappa

Abstract. A statistical test for detecting answer copying on multiple-choice tests based on Cohen's kappa (Cohen, 1960) is proposed. The test is free of any assumptions on the response processes of the examinees suspected of copying and having served as the source, except for usual assumption that these processes are probabilistic. Because the asymptotic null and alternative distributions of the kappa statistic are derived under the assumption of common marginal probabilities for all items, a recoding of the item alternatives is proposed to approximate this case. The results from a simulation study in this paper show that under this recoding the test approximates its nominal Type I error rates and has promising power functions.

5.1 Introduction

In educational testing, the multiple-choice format is often used because it provides an efficient and reliable way of scoring tests for large numbers of examinees. A serious problem with this format, however, is that, unless effective precautions are taken, copying of answers among examinees is easy.

This chapter was submitted for publication as: Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2003). Detecting answer copying using statistic kappa.

To detect answer copying on multiple-choice tests, both observational and statistical methods can be used (Cizek, 1999). Observational methods use a human observer to establish if answer copying has occurred. The evidence an observer can collect is observations of certain types of examinee behavior (e.g., one examinee talking to another during the test) or physical evidence (e.g., confiscated cheat sheets exchanged between two examinees). Statistical methods address cheating by modeling the response probabilities of examinees under the assumption of no cheating and looking for patterns of similar answers between examinees that are unlikely under the model.

Several copying statistics have been proposed to detect or back up allegations of answer copying. All these statistics are defined on the response vectors of the examinee suspected of copying and the examinee believed to have served as a source. For simplicity, we will call these examinees the copier and the source, respectively. Examples of copying indices are the K index (Holland, 1996; Lewis & Thayer, 1998) and its variants $\overline{K}_2$ (Sotaridona & Meijer, 2002), S_1, and S_2 (Sotaridona & Meijer, 2003), the B_m index (Bay, 1995), the g_2 index (Frary, Tideman, & Watts, 1977), and the ω-index (Wollack, 1997; Wollack & Cohen, 1998). For a comprehensive review of copying indices, see Cizek (1999).

Most of the statistics for detecting copying are based on the number of similar responses between the source and the copier, or on a standardized form thereof. Under the null hypothesis of no copying, these indices are assumed to follow a distribution, for example, the (generalized) binomial or Poisson distribution. To evaluate these statistics under the null distribution, parameter estimation may be required, for example, estimation of the item and examinee parameters in a response model.

A general problem with these statistics is that when the sample size is small, their parameters cannot be estimated reliably. Another problem is that the definitions of the parameters often involve a population of examinees. As a consequence, examinees suspected of copying would get a different value for the copying statistic, and hence a different likelihood of being suspected, if they had produced the same vector of responses but were included in a different population.

The aim of this study is to investigate a copying statistic with a null distribution that is independent of the behavior of any other examinees than the source and the copier. An advantage of such a statistic is not only that it never penalizes examinees for being "member of a population", but also that it can be used in small test settings, for example, classroom tests. For other statistical test based only on the response vectors of the source and the copier but with stronger assumptions on their response processes,

see van der Linden and Sotaridona (in press; 2003).

5.2 Assumptions on Response Process

Consider a test consisting of items $i = 1, \ldots, N$, each with response options $v = 1, \ldots, m$. Examinee index j takes the value c for the examinee suspected of copying and s for the examinee believed to have served as his/her source. In addition, U_{ji} denotes the response of j to i.

We do not assume anything specific about the multiple-choice format of the items, except that one alternative is correct and the examinees are instructed to choose the alternative which they believe is true. Particularly, it is possible that the items are allowed to have a nominal response format or a format that involves graded scoring. In addition, we assume that the response behavior of c and s is probabilistic, that is, can be characterized by a (possibly different) probability distribution over the alternatives of each item. However, we leave these probabilities unspecified and do not make the additional assumption that they follow a specific polytomous response model, as has been done, for example, in van der Linden and Sotaridona (2003) and Wollack (1997). In fact, the research reported in this paper was just motivated by the question of how much could be inferred about copying behavior of examinees without assuming any specific response model.

5.2.1 Independence and Agreement

The key observation on which the method in this chapter rests is that if the responses by c and s are probabilistic and c did not have access to the answers by s, the responses are statistically independent. If c did have access to some of these answers and copied them, the responses of c and s on these items would not only be dependent but even in perfect agreement. However, if c did not copy any answers from s, as their responses are probabilistic, it is still possible that some of them agree. The statistical problem we thus face is to decide how much agreement we should accept before rejecting the null hypothesis that c did not copy the answers for any of the items in the test.

The notion of agreement between the responses of c and s can be formalized as follows. Suppose all responses by c and s on the items in the test are collected in an mxm table. Response that are in agreement are classified in the main diagonal of this table, responses that do not agree in its off-diagonal area. The number of responses in the diagonal is invariant under permutations of the response alternatives over the rows/columns. This

type of table therefore seems appropriate for a statistical analysis of agreement between responses of pairs of examinees to items with a polytomous response format.

5.2.2 Conditioning

Some of the statistics for detecting answer copying in the literature are based on the idea to condition on some of the information in the response vector of the source. The reasons for this choice are that: (1) analyzing agreement between incorrect responses only may provide intuitively more powerful evidence of copying (Holland, 1996); (2) the model for the response process implies a conditional statistic, as is the case, for example, with the knowledge-copying-or-random-guessing model in van der Linden and Sotaridona (in press); and (3) the fact that the statistic was offered as an improvement on a predecessor that used conditioning (Wollack, 1997).

The method presented in this chapter is based only on the assumption that the responses are probabilistic. Under the null hypothesis of no copying, the only fixed quantity is the number of responses. For the mxm table introduced above, the hypothesis leads to a multinomial model with a fixed total number of observations but counts in the cells and margins of the table that are random. The only type of conditioning that may seem reasonable is on the incorrect responses by s. However, the consequence would be the loss of a row and a column in the table, and, hence, loss of information. We will therefore refrain from any type of conditioning.

5.3 Kappa

The hypothesis testing problem above exists also in other fields, e.g. psychiatry, when two raters are asked to classify a sample of subjects independently on a scale representing some construct in, e.g., a theory of mental health. In these fields, a standard approach is also to tabulate the joint responses of the raters into a two-way table, with one rater represented by the rows and the other by the columns of the table. The statistical test usually conducted to test the hypothesis of no agreement in the table is based on statistic kappa (Cohen, 1960). The statistical theory of the sampling distribution of kappa began with the derivation of its large-sample standard error in Fleiss, Cohen, and Everitt (1969). For a review of the theory and some applications of kappa in medical research, see Agresti (1990, 367-368).

Let π_{vv} denote the probability of a classification in cell (v, v) and π_{v+} and π_{+v} of a classification in row and column v, respectively. These probabilities

allow us to calculate a parameter for the true agreement between the two raters, which corrects for agreement by chance in the table:

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e}, \tag{5.1}$$

where

$$\pi_o = \sum_v \pi_{vv} \tag{5.2}$$

is the probability of an observed agreement between the two raters and

$$\pi_e = \sum_v \pi_{v+}\pi_{+v} \tag{5.3}$$

the probability of agreement due to chance if the raters operate independently.

5.3.1 Hypotheses

Using parameter κ, the hypothesis to be tested for agreement between raters beyond mere chance can be formulated as

$$\mathrm{H}_0 : \kappa = 0 \tag{5.4}$$

against

$$\mathrm{H}_1 : \kappa > 0. \tag{5.5}$$

5.3.2 Null Distribution of $\widehat{\kappa}$

Let $\widehat{\kappa}$ denote the statistics that is obtained by replacing π_o and π_e by the sample statistics $\widehat{\pi}_o = p_o$ and $\widehat{\pi}_e = p_e$, respectively, where

$$p_o = \sum_v p_{vv}, \tag{5.6}$$

$$p_e = \sum_v p_{v+}p_{+v}, \tag{5.7}$$

and p_{vv}, p_{v+}, and p_{+v} are the empirical proportions in cell (v, v) and row and column v, respectively.

Statistic $\widehat{\kappa}$ is asymptotically normally distributed (Agresti 1990, p. 366) with expected value

$$E(\widehat{\kappa}) = \kappa \tag{5.8}$$

and variance

$$\sigma^2(\widehat{\kappa}) = \frac{1}{N}\left\{\frac{\pi_o(1-\pi_o)}{(1-\pi_e)^2} + a + b\right\}, \tag{5.9}$$

with

$$a = \frac{2(1-\pi_o)\left(2\pi_o\pi_e - \sum_v \pi_{vv}(\pi_{v+}+\pi_{+v})\right)}{(1-\pi_e)^3},$$

$$b = \frac{(1-\pi_o)^2\left(\sum_v\sum_v \pi_{vv}(\pi_{v+}+\pi_{+v})^2 - 4\pi_e^2\right)}{(1-\pi_e)^4},$$

and where N is the number of ratings.

5.3.3 Statistical Test

The following standardization of $\widehat{\kappa}$ is defined

$$Z_{\widehat{\kappa}} = \frac{\widehat{\kappa} - \mathrm{E}(\widehat{\kappa})}{\sigma(\widehat{\kappa})}, \tag{5.10}$$

with $\mathrm{E}(\widehat{\kappa})$ the expected value of $\widehat{\kappa}$ in (5.8) and $\sigma(\widehat{\kappa})$ the standard deviation of $\widehat{\kappa}$ which is the square root of the variance in (5.9). Under the null hypothesis in (5.4) it holds for the expected value of $\widehat{\kappa}$ that

$$\mathrm{E}(\widehat{\kappa}) = 0,$$

whereas its variance in (5.9) becomes equal to

$$\begin{aligned}\sigma^2(\widehat{\kappa}) &= \frac{1}{N(1-\pi_e)^2}\left\{\pi_e(1-\pi_e) + \sum_v\sum_v(\pi_{v+}\pi_{+v})(\pi_{v+}+\pi_{+v})^2\right. \\ &\quad \left. -2\sum_v(\pi_{v+}\pi_{+v})(\pi_{v+}+\pi_{+v})\right\}\end{aligned} \tag{5.11}$$

To obtain a test statistic for the hypotheses in (5.4) and (5.5), it is customary to replace $\sigma^2(\widehat{\kappa})$ in (5.10) by its sample equivalent, absorbing its sampling variance in the argument that leads to asymptotic normality of (5.10). The resulting statistic

$$Z_{\widehat{\kappa}} = \frac{\widehat{\kappa}}{\widehat{\sigma}(\widehat{\kappa})}, \tag{5.12}$$

with $\widehat{\sigma}(\widehat{\kappa})$ denoting the square root of the sample equivalent of (5.11), has asymptotic null distribution $Z_{\widehat{k}} \sim N(0,1)$. The test of the null hypothesis under this distribution is right sided with critical value z^* defined by

$$\Pr(Z_{\widehat{\kappa}} \geq z^*) = \alpha. \tag{5.13}$$

5.4 Application to Detection of Copying

The analogy between the problem of detecting answer copying and agreement between ratings seems to suggest that we can also use statistic $Z_{\widehat{k}}$ to test if the response vectors of c and s agree beyond chance. We will do so, but, in so doing, are aware of a potential problem in this application due to the fact that the probabilities that c and s choose an alternative generally differ across items. As a consequence, the joint response by c and s follows a different multinomial distribution for each item, whereas the sampling model of $Z_{\widehat{k}}$ used to obtain the asymptotic results in Fleiss, Cohen and Everitt (1969) is based on the same multinomial distribution for each observation.

The presence of nonidentically distributed observations does not challenge the use of the central limit theorem in the derivation of asymptotic normality of $Z_{\widehat{k}}$ in Fleiss, Cohen and Everitt (1969). If the sum of the variances for the cells in the table does not tend to a finite limit, an assumption which is reasonable for our application, a version of the theorem for nonidentically distributed observations holds (Lehmann, 1999, Corollary 2.7.1). However, there is a problem in using the probabilities π_{v+} and $\pi_{+\nu}$ in (5.3) and (5.11). These probabilities are based on the assumption of common response probabilities by c and s across all items, whereas in fact they do vary.

We will study the impact of this variation on the Type I error in the simulation study below. Also, we will present the results of a partial solution to this problem, which consists of permuting the alternatives of the items before pooling their information in the table. As noted earlier, these permutations do not change the number of agreements between the responses in the diagonal of the table but do have an impact on the marginal probabilities for the table. We capitalize on this fact by choosing a recoding of the alternative that leads to lower variation of response probabilities across items.

5.4.1 Discussion

Though never discussed in the literature, the same problem is likely to exists for statistical tests of rater agreement based on $\widehat{\kappa}$ in other fields. For example, in studies of clinical diagnosis, it seems hard to believe that clinicians maintain their *a priori* probabilities of classifying cases in the categories on the scale. Rather, we expect these probabilities to change during the rating process as a function of the actual cases they have already rated.

In addition to differences in response probabilities between items, we also have differences between c and s due to their different abilities. The effect of these differences is different marginal probability distributions for c and s, and, hence, a maximum possible value of κ smaller than one (Cohen, 1960). But this fact does not constrain the conclusions from the statistical test based on statistic $\widehat{\kappa}$ in any way. Because N is the only fixed quantity in the test, lack of agreement between classifications manifests itself not only by a smaller number of joint responses in the diagonal of the table but also by differences between the marginal distributions. The lower maximum for κ is thus a consequence of the lack of agreement, as it should be, and not an undesirable cause of it.

5.5 Power Analysis

The asymptotic distribution of $\widehat{\kappa}$ under the alternative hypothesis is $N(\kappa, \sigma^2(\widehat{\kappa}))$, with $\sigma^2(\widehat{\kappa})$ given by (5.9). It is thus possible to estimate the (asymptotic) power function of the test, which is given by the probabilities

$$\Pr\left(\frac{\widehat{\kappa}-\kappa}{\sigma(\widehat{\kappa})} \geq z^*\right). \tag{5.14}$$

The only thing needed to for this estimate is to calculate an estimate of $\sigma^2(\widehat{\kappa})$ in (5.9) from the sample proportions. Because the power depends on the probabilities $\pi_{\nu\nu}$, $\pi_{\nu+}$, $\pi_{+\nu}$, and π_o, the power functions for the test can be approximated by generating tables with joint responses of c and s under various levels of copying and plotting (5.14) as a function of these levels. In practice, this type of power analysis is thus only possible if we have estimates of the response probabilities of the two examinees under a polytomous model. An example of this type of power analysis under the nominal response model will be given in the simulation studies in the next section.

5.6 Simulation Study

The purpose of these studies were threefold: (1) to study the impact of the differences between the response probabilities of the items on Type I error of the statistical test based on kappa, (2) to show the effects of recoding the alternatives of the items before pooling their information in the table, and (3) to explore the power of the test using the estimate in the previous section. We used multiple-choice tests consisting of 30 and 60 items, each with 5 response alternatives per item. The ability levels of s and c were varied over the whole range of values that can be met in an application. The significance level used was $\alpha = .05$, and the Type I error was thus evaluated using critical value of $\widehat{\kappa}$ equal to 1.645. This choice of significance level was conventional and is not necessarily the right choice in a practical application, where it should be chosen carefully balancing between the Type I and expected Type II errors acceptable to the testing agency and examinees.

5.6.1 Response Model

The item responses were simulated using the nominal response model. Under this model the probability of examinee j with ability level θ_j responding to option v of item i, where i have intercept and slope parameters ζ_{ih} and λ_{ih}, is given by

$$\pi_{i\nu}(\theta_j) = \frac{\exp(\zeta_{i\nu} + \lambda_{i\nu}\theta_j)}{\sum_{h=1}^{m} \exp(\zeta_{ih} + \lambda_{ih}\theta_j)}. \tag{5.15}$$

Further details of the model can be found in Bock (1972, 1997).

5.6.2 Parameter Values

The impact of the abilities of the source and the copier on κ was controlled by using a grid of possible (θ_c, θ_s) values and then fixing the number of examinee pairs for each grid. The grid was defined on the interval $[-2, 2]$ with an increment of .5 in each component. The number of response vectors generated for each grid point was 2,000. One condition had identical parameter values for all items. Under this condition, the item was chosen to have a probability of .50 for the correct alternative at $(\theta_c, \theta_s) = (0, 0)$. The second condition had different parameter values for the items. The slope and intercept parameters of each item were drawn from $U(-1, 1)$ and $U(-1.5, 1.5)$, respectively.

5.6.3 Generation of Response Vectors

The observed response of examinee j to item i was obtained by drawing a sample from the set $v = \{1, \ldots, 5\}$, where each element of v has a probability of being drawn equal to $\pi_{i1}(\theta_j), \pi_{i2}(\theta_j), \ldots, \pi_{i5}(\theta_j)$ respectively, with $\pi_{iv}(\theta_j)$ computed using (5.15). As in other studies (Thissen & Steinberg, 1997; Sotaridona & Meijer, 2002) the response alternative of an item with the largest value of the slope parameter was chosen as its correct alternative.

5.6.4 Results

Impact of Different Probabilities between Items.

Table 5.11 shows the empirical Type I error rates for the case of identical response probability for all items in the test. Most of the rates were close to their nominal value of $\alpha = .05$. The exception were the rates for the combinations of extreme θ values in opposite directions, which appeared to be systematically smaller than expected. Our explanation is that for these combinations we have probability distributions for the response alternatives for s and c that are extremely skewed in opposite directions. It is a well-known fact that for such probabilities the normal approximation is quite slow. In fact, the approximation is so slow that lengthening of the test from $N = 30$ to $N = 60$ items has a negligible effect on the most extreme combinations of θ values. Fortunately, use of the test for these θ values would be conservative, that is, lead to probabilities of Type I errors that are smaller than the nominal values of α. Our conclusion is that if all items have response probabilities close to each other, a test based on statistics κ can be used safely for all examinees, but some instances of answer copying among examinee with extreme ability levels in opposite directions may escape the attention of the testing agency.

The results for the case of items with different response probabilities for the items are in Table 5.2. These results show that variation in the probabilities would create a huge problem for a test based on the kappa statistic. All rates were extremely higher than required, except for combinations of θ values at the opposite end of the scale. Thus, if the items differ largely in their response probabilities for the examinees, and nothing is done to decrease this variation, using a test based on statistics κ cannot be recommended for most of the combinations of θ values, except for a few combinations of extreme values in opposite directions. To get more favorable conditions for the application of κ we propose to recode the items.

Table 5.1: Empirical Type I Error Rates for Case of Items With Identical Response Probabilities ($\alpha = .05$)

Test Length			θ_s								
			-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
N=30		-2.0	.02	.04	.03	.01	.00	.00	.00	.00	.00
		-1.5	.03	.04	.03	.02	.01	.00	.00	.00	.00
		-1.0	.02	.03	.04	.03	.02	.01	.00	.00	.00
		-0.5	.02	.02	.03	.04	.04	.03	.02	.00	.00
	θ_c	0.0	.00	.01	.02	.03	.05	.03	.02	.02	.01
		0.5	.00	.00	.01	.02	.04	.04	.03	.03	.02
		1.0	.00	.00	.00	.02	.02	.04	.04	.03	.03
		1.5	.00	.00	.00	.00	.01	.04	.03	.04	.04
		2.0	.00	.00	.00	.00	.01	.03	.02	.04	.04
N=60		-2.0	.04	.04	.03	.01	.00	.00	.00	.00	.00
		-1.5	.03	.04	.04	.02	.01	.00	.00	.00	.00
		-1.0	.02	.03	.04	.02	.02	.02	.00	.00	.00
		-0.5	.01	.02	.05	.04	.04	.02	.01	.00	.00
	θ_c	0.0	.00	.01	.02	.04	.04	.04	.03	.02	.01
		0.5	.00	.00	.01	.02	.03	.04	.04	.03	.02
		1.0	.00	.00	.01	.01	.03	.04	.04	.04	.04
		1.5	.00	.00	.00	.00	.01	.03	.04	.05	.02
		2.0	.00	.00	.00	.00	.01	.02	.04	.04	.04

Recoding the Items

As already noted, for the nominal response format, there is no unique way to assign the response alternatives to the rows/columns in the table. Also, each possible assignment will result in the same counts in the diagonal. We use this freedom to recode the response alternatives to have more acceptable differences in probabilities between the items. In this study, we explored two different methods.

Recoding. The first method was based on the following recoding: (1) the correct alternative was assigned to the first row/column of the table and (2) the other alternatives were assigned in decreasing order of their a-value (proportion of examinees who chose the alternative) to the remaining rows/columns. The empirical Type I error rates after use of this method for the same data set as in Table 5.2 are given in Table 5.3. The impact

Table 5.2: Empirical Type I Error Rates for Case of Items With Nonidentical Response Probabilities ($\alpha = .05$)

Test Length		θ_s								
		-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
N=30		-2.0	.90	.83	.70	.48	.26	.12	.04	.01	.00
		-1.5	.81	.77	.66	.49	.28	.15	.06	.02	.01
		-1.0	.70	.64	.58	.44	.32	.20	.12	.05	.03
		-0.5	.47	.47	.43	.38	.33	.26	.18	.15	.11
	θ_c	0.0	.26	.30	.32	.32	.33	.31	.31	.31	.28
		0.5	.11	.13	.18	.24	.35	.29	.43	.48	.51
		1.0	.02	.06	.10	.19	.29	.44	.56	.64	.72
		1.5	.01	.02	.05	.13	.30	.48	.62	.77	.83
		2.0	.00	.01	.03	.11	.28	.49	.70	.87	.92
N=60		-2.0	.99	.98	.89	.68	.36	.11	.02	.00	.00
		-1.5	.97	.93	.83	.64	.39	.17	.06	.02	.01
		-1.0	.89	.85	.72	.58	.40	.26	.14	.07	.03
		-0.5	.65	.63	.58	.54	.44	.34	.26	.19	.13
	θ_c	0.0	.34	.37	.42	.43	.45	.45	.43	.36	.36
		0.5	.12	.18	.23	.37	.47	.55	.60	.62	.66
		1.0	.02	.05	.13	.26	.42	.59	.71	.79	.82
		1.5	.01	.01	.07	.18	.39	.59	.78	.88	.93
		2.0	.00	.00	.03	.14	.34	.65	.86	.93	.97

of this simple method was already dramatic, particularly for the θ values of the c and s in the middle of the scale. For these values, all error rates were near their nominal value, and a statistical test of answer copying based on statistic κ became possible. Also, the normal approximation seems already to be stable after $N = 30$ items; the increase to $N = 60$ did not introduce any noticeable change in results.

Recoding conditional on θ. The second method was identical to the first method, except that it was applied conditional on the ability level of c. That is, the a-values of the items for c were estimated by simulating $1,000$ response patterns for each ability level of c and then using the responses to calculate the sample proportions. Table 5.4 shows the empirical Type

Table 5.3: Empirical Type I Error Rates for Case of Items With Nonidentical Response Probabilities After Recoding of Alternatives ($\alpha = .05$)

Test Length		θ_s								
		-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
N=30		-2.0	.24	.17	.11	.05	.02	.00	.00	.00	.00
		-1.5	.18	.15	.11	.06	.02	.01	.00	.00	.00
		-1.0	.11	.08	.09	.06	.02	.02	.02	.00	.00
		-0.5	.04	.07	.05	.04	.05	.04	.02	.01	.01
	θ_c	0.0	.03	.03	.03	.05	.06	.06	.06	.05	.03
		0.5	.00	.01	.01	.04	.07	.10	.12	.11	.12
		1.0	.00	.00	.01	.02	.05	.12	.14	.18	.17
		1.5	.00	.00	.00	.02	.05	.11	.17	.27	.31
		2.0	.00	.00	.00	.01	.03	.10	.21	.33	.38
N=60		-2.0	.53	.33	.16	.05	.01	.00	.00	.00	.00
		-1.5	.35	.23	.13	.05	.02	.01	.00	.00	.00
		-1.0	.14	.14	.07	.06	.04	.01	.01	.00	.00
		-0.5	.05	.05	.06	.04	.05	.04	.03	.02	.02
	θ_c	0.0	.01	.02	.03	.04	.06	.08	.10	.07	.06
		0.5	.00	.00	.02	.04	.08	.15	.20	.20	.20
		1.0	.00	.00	.00	.02	.08	.18	.26	.34	.35
		1.5	.00	.00	.00	.02	.08	.18	.32	.44	.47
		2.0	.00	.00	.00	.01	.05	.18	.36	.49	.57

I error rates after the recoding of the items conditional on the θ value of c. It is clear that there are significant improvements over the first recoding method. The error rates were in general very close to their nominal level across the whole range of the θ values of c and s. For $N = 60$ the Type I error rates are somewhat higher than for $N = 30$, and for high positive θ values for both the source and the copier they are somewhat inflated. We have not been able to find an explanation for this inflation, which made the test slightly liberal.

Recoding conditional on number-correct score. In practice, when an estimate of the ability of c in a parametric may not be available, the number-correct score can be used as an estimate. In this type of conditioning, the response patterns of examinees with similar number-correct scores as the copier can be used to estimate conditional a-values for the items, but the

Table 5.4: Empirical Type I Error Rates for Case of Items With Nonidentical Response Probabilities After Recoding of Alternatives Conditional on the Ability of the Copier($\alpha = .05$)

Test Length		θ_s								
		-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
N=30		-2.0	.06	.06	.06	.06	.04	.04	.03	.02	.02
		-1.5	.05	.05	.06	.05	.04	.04	.04	.03	.02
		-1.0	.06	.06	.06	.06	.06	.04	.04	.04	.04
		-0.5	.05	.06	.06	.06	.06	.06	.05	.06	.04
	θ_c	0.0	.05	.05	.05	.06	.05	.08	.05	.04	.04
		0.5	.04	.04	.06	.07	.06	.06	.05	.06	.07
		1.0	.04	.04	.04	.04	.06	.07	.07	.06	.04
		1.5	.02	.03	.04	.06	.05	.04	.06	.06	.07
		2.0	.02	.03	.03	.04	.04	.06	.06	.07	.06
N=60		-2.0	.08	.10	.07	.07	.06	.06	.04	.03	.02
		-1.5	.07	.06	.09	.06	.07	.06	.04	.03	.04
		-1.0	.08	.07	.08	.07	.06	.08	.06	.05	.04
		-0.5	.08	.08	.08	.07	.07	.06	.07	.07	.05
	θ_c	0.0	.08	.07	.08	.08	.10	.08	.09	.07	.05
		0.5	.07	.08	.07	.08	.08	.09	.09	.10	.08
		1.0	.05	.05	.08	.06	.08	.08	.10	.10	.11
		1.5	.04	.06	.06	.06	.08	.10	.12	.10	.11
		2.0	.04	.04	.06	.08	.09	.08	.12	.11	.12

recoding remains otherwise identical. The results of a simulation applying this approach are shown in Table 5.5.

In this simulation, we first sampled 1,000 θ values from $N(0, 1)$ and generated item score patterns using the nominal response model in Equation (5.15) with $a \sim U(-1, 1)$ and $c \sim U(-1.5, 1.5)$. Then we grouped the score patterns into six number-correct score categories and calculated the conditional a-values conditional on these grouped scores. The ranges of score in each group are shown in 5.5. The minimum number of simulees in each group was 145. The Type I error was computed for each cell using $2,000$ pairs of response vectors where the pairs were sampled without replacement.

The results in Table 5.5 show that the conditional recoding approach using the number-correct score resulted in empirical Type I errors that were

Table 5.5: Empirical Type I Error Rates for Case of Items With Nonidentical Response Probabilities After Recoding of Alternatives Conditional on the Number-Correct Score of the Copier ($\alpha = .05$)

Test Length		Score Group					
		0-6	*7-10*	*11-14*	*15-18*	*19-21*	*22-30*
N=30	*0-6*	.10	.06	.03	.08	.00	.00
	7-10	.06	.08	.04	.02	.00	.00
	11-14	.03	.05	.07	.06	.02	.00
	15-18	.01	.04	.06	.08	.05	.02
	19-21	.00	.01	.03	.05	.07	.07
	22-30	.00	.00	.01	.02	.06	.08
		0-13	*14-21*	*22-27*	*28-34*	*35-40*	*41-60*
N=60	*0-13*	.12	.10	.03	.01	.00	.00
	14-21	.10	.11	.09	.04	.00	.00
	22-27	.06	.10	.08	.08	.04	.01
	28-34	.02	.05	.09	.08	.08	.04
	35-40	.00	.01	.04	.07	.09	.07
	41-60	.00	.00	.02	.03	.05	.08

close to those for conditioning on a known value of θ. For extreme values and in the same direction, the test remains somewhat too liberal; for values in the opposite direction, the test becomes more conservative.

Power Study

In this study, we conducted the power analysis described earlier under the nominal response model with recoding conditional of the true ability of c. Let $\gamma = 1, 2, \ldots, N$ denote the number of items copied by c from s. For each pair of response patterns of c and s, (U_c, U_s), copying was simulated by generating a response pattern $U_{c\gamma}$ from U_c by replacing the response of c with those of s in U_c on γ randomly selected items. The pairs of response

patterns $(U_{c\gamma}, U_s)$ were used to calculate the power in (5.14) as a function of γ. The power was calculated for an actual Type I error rate $\alpha = .05$. The process was repeated for all pairs of selected θ values of c and s.

The results for $N = 30$ and $N = 60$ are shown in Figures 5.1 and 5.2 for $\theta = -1.5, -0.5, 0.5$, and 1.5. These figures show that the test had high power (at least .80) when the examinee copied a minimum of 12 items or 40 percent of the items for $N = 30$ or at least 18 items or 30 percent of the items for $N = 60$. The power is somewhat higher when the ability of the source is near the middle of the scale (-0.5 and 0.5) than in the extreme of the scale (-1.5 and 1.5).

5.7 Discussion

A statistical test based on statistic κ to detect answer copying in a multiple-choice test was proposed. Our basic idea was that answer copying results in more agreement between the response of two examinees than can be expected by chance. The main advantage of choosing κ as test statistic is that it is a measure of agreement with known asymptotic properties and a proven record in numerous types of applications.

Another feature of κ is that its value is calculated only from the responses of the two examinees suspected of copying and having served as a source. However, κ showed to be sensitive to differences in response probabilities between items for examinees c and s, who under the null hypothesis of no copying, have different ability levels. We showed that this sensitivity can be removed by coding conditional on the observed number-correct score. In the empirical example, the test became somewhat conservative for extreme scores in the same direction and liberal for extreme scores in opposite direction. In our opinion, for a statistical test on answer copying, the former is no problem, and the latter can easily be remedied by choosing a smaller than nominal significance levels for pairs of examinees with such scores. More experience with a larger variety of test lengths, response probabilities, and score ranges conditional on which we should recode is needed to be able to generalize the results in the simulation study.

Though we were attracted to using κ because it is a measure of agreement independent of the response vectors of any other examinees than c and s, we ended with a procedure that assumes the presence of a population of examinees. The result is not disappointing, though. The scores in this population are only used to recode the response alternatives on the items for c and s. More importantly, however, the recoding procedure uses the

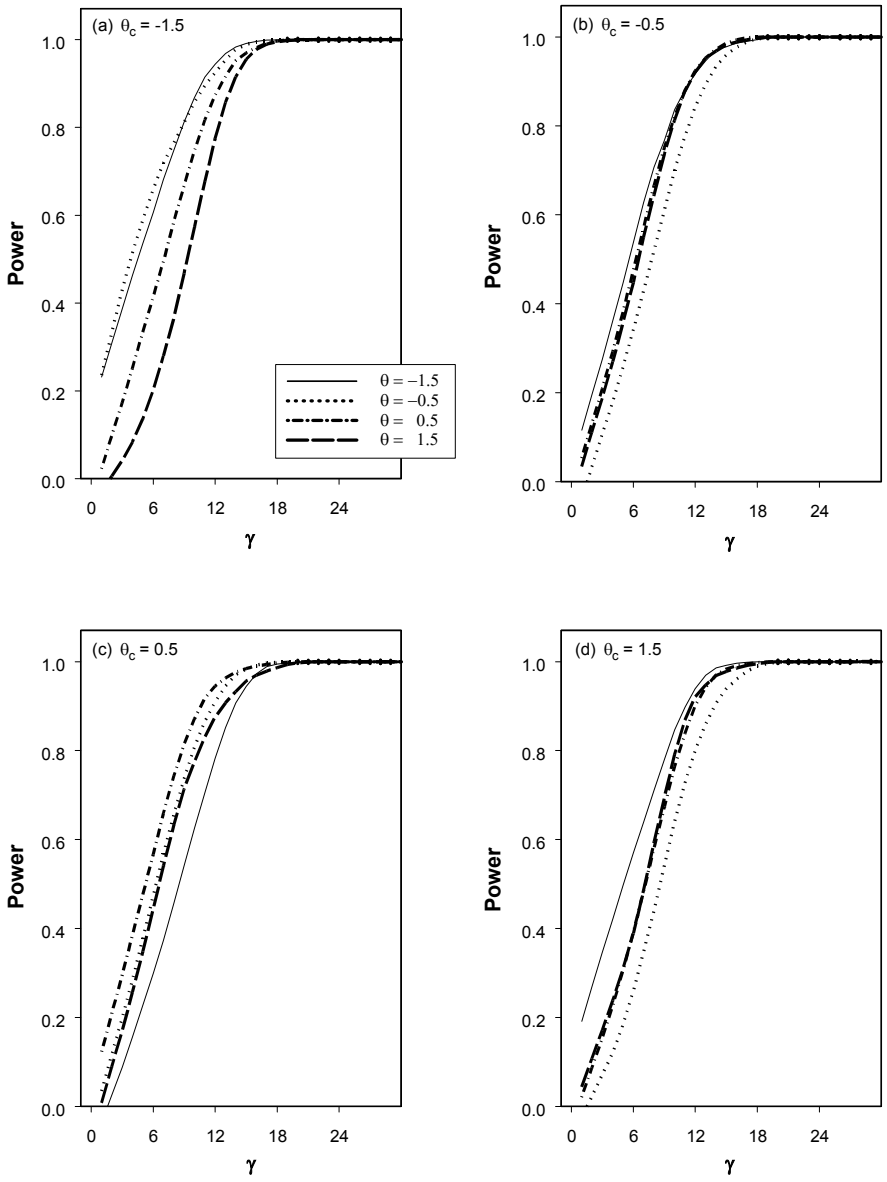


Figure 5.1: Power Functions of the Test for $\theta_s = -1.5, -0.5, 0.5, 1.5$ and $\theta_c = -1.5, -0.5, 0.5, 1.5$ for $N = 30$.

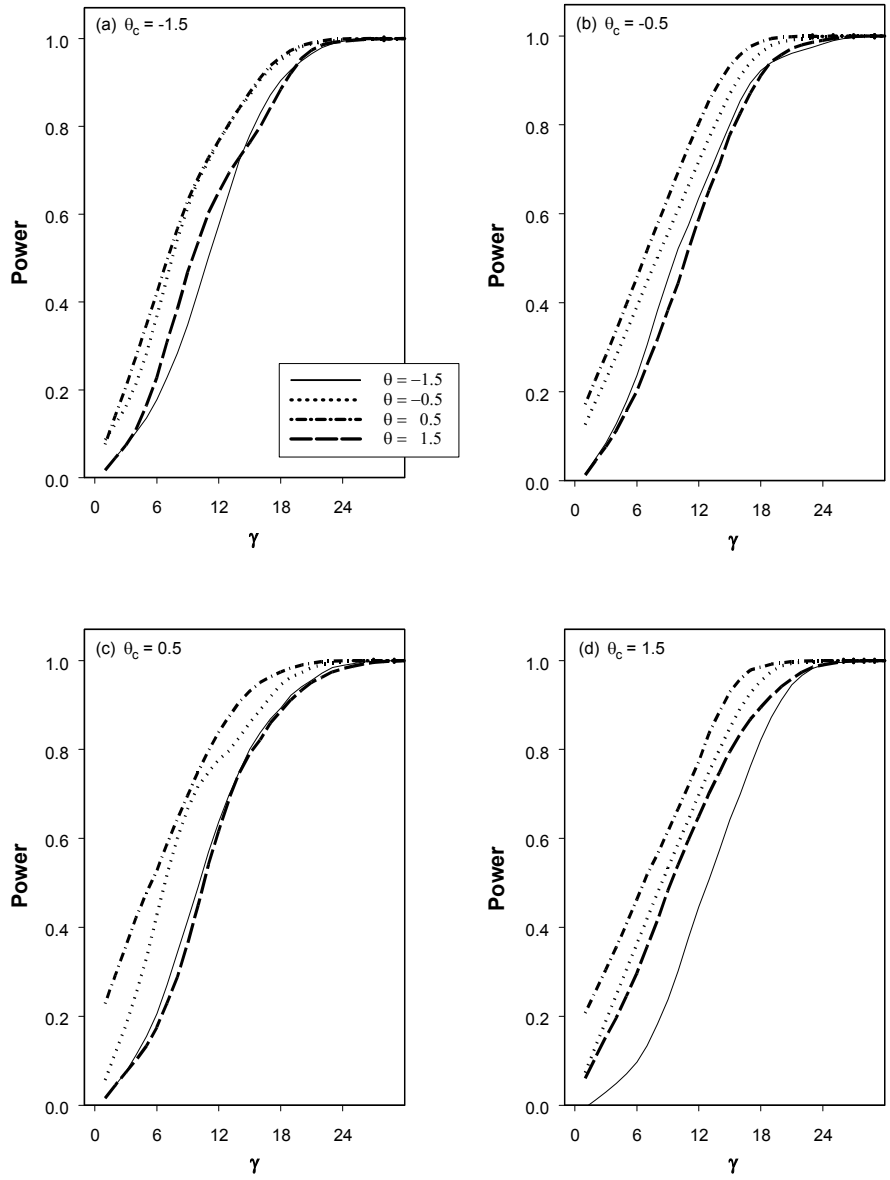


Figure 5.2: Power Functions of the Test for $\theta_s = -1.5, -0.5, 0.5, 1.5$ and $\theta_c = -1.5, -0.5, 0.5, 1.5$ for $N = 60$.

number-correct score, which has been proven to be monotone in the ability parameter for a large variety of polytomous IRT models (Hemker, Sijtsma, Molenaar & Junker, 1996, 1997). No assumption on the response behavior of the examinees in the population is made other than that their response probabilities can be described by an (unknown) model in this collection.

Chapter 6

Screening Using Neural Networks

Abstract. A conceptually new approach to screening data from high-stakes tests for possible cases of cheating using a neural network was proposed. The idea is to use this approach as a supplement to existing screening practices in order to further improve test security measures. The paper gives an overview of the basic principles of neural networks and discussed how this technique can be applied to identify cheaters. The results of a small simulation study showed that a neural network approach has high power when the configuration of the item scores of a cheater has similar characteristics as the configuration used to train the network. Some directions for future research are discussed.

This chapter was submitted for publication as: Sotaridona, L. S.(2003). Screening of Cheating on high-stakes tests using neural networks.

6.1 Introduction

A thorough investigation of cheating on high-stakes tests is usually initiated by a trigger. A trigger could be (a) an unusually large increase in the score relative to the recent score obtained for the same test, (b) a very large decrease in the score relative to the recent score obtained for the same test, (c) a report of suspicious behavior during the test, or (d) other events, such as reports of lost test booklets, that may arouse suspicion that cheating had occur (see for example Cizek, 1996, p. 142). Once a trigger is identified, additional evidence may be obtained and a decision has to be made whether the suspicion is worth pursuing for further investigation. Additional evidence to proceed with an investigation will often involve the use of a statistical test of answer copying as one of the piece of evidence.

Cizek (1996, p. 146) pointed out that reliance on triggers (c) and (d) is insufficient because it often leads to underestimation of the number of cases identified for further statistical analysis. A check of the possible existence of triggers (a) and (b) is usually done routinely by testing companies as one of the aspects of a complex test security processes to insure the validity of the reported test scores. An example of such processes that are practiced at the Educational Testing Service (ETS) is described in Cizek (1996, p. 156).

In the absence of other evidence, triggers (a) and (b) become the sole device for screening out possible cheaters. Although triggers (a) and (b) are very useful in practice, they are not sufficient as a screening device. For example, in order to use this device an examinee should have taken the same test at least once in the past. This means that the cheaters who were very successful in their first attempt to take the test are left undetected, assuming that no other triggers are present. In this paper, a conceptually new approach for screening possible cheaters is proposed. The idea is to use this approach as a supplement to existing screening practices in order to further improve test security measures.

The new approach is an application of a neural network to classification problem which involves assigning input patterns to one of a set of discrete classes (Bishop, 1997). The classification task involve two steps. The first step is to train the neural network to identify a set of response patterns we are interested in using representative samples of these response patterns as training samples. The second stage is application, that is use the trained neural network to identify similar, both seen and unseen, cases of response patterns. For a neural network to be successful in this task, it should be able to generalize what it learned from the few representative samples to other similar response patterns that were not used during the training.

To identify the cheaters, we restrict ourselves to the simplest case, namely we assume that we have two classes of examinees, cheaters and noncheaters, and that the input patterns are the response patterns of known cheaters and noncheaters, for example obtained in an earlier administration of the test. These response patterns can be used to train the neural network. The aim of this study is to investigate if a neural networks can be used to identify cheating.

This study is organized as follows. First, we introduce the basic principles of neural networks. Second, we discuss how this technique can be applied to identify cheaters. Third, we illustrate the use of neural networks in a small simulation study, where we discuss in some detail the use of a training algorithm. Finally, we discuss directions for future research.

6.2 Overview on Neural Networks

Neural networks (NN) are collections of mathematical models that emulate some of the properties of biological nervous systems and draw on the analogies of adaptive biological learning. The key element of the NN paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements that are analogous to neurons and are tied together with weighted connections that are analogous to synapses (Freeman & Skapura, 1991).

Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of NN as well. Learning typically occurs by example through training, or exposure to a set of input vector $\mathbf{x}$ and output vector $\mathbf{y}$ where the *training algorithm* iteratively adjusts the connection weights ($\mathbf{w}$). The values of $\mathbf{w}$ can be interpreted as parameters (e.g., like the values of b_0 and b_1 in $\widehat{y} = b_0 + b_1 x$) that have to be estimated and the learning process specifies the algorithm used to estimate the parameters (Abdi, Valentin, & Edelman, 1999). The set of ordered vector pair $(\mathbf{x}, \mathbf{y})$ is called the training set or examples.

Basically, the task during training a neural network is simply to search for $\mathbf{w}$ so that using some function $f(\mathbf{w}; \mathbf{x})$ we can map input $\mathbf{x}$ to output $\mathbf{y}$ with minimal error. In most cases it will not be possible to determine a suitable value of $\mathbf{w}$ without the help of training data. The mapping is therefore modeled in terms of $f(\mathbf{w}; \mathbf{x})$ which contains a number of adjustable parameters $\mathbf{w}$, whose values are determined with the help of the training data.

The importance of neural networks is that they offer a very powerful and

very general framework for representing nonlinear mappings from several input variables to several output variables. Simple techniques for representing multivariate nonlinear mappings in one or two dimensions rely on linear combinations of fixed basis functions. Such methods have severe limitations when extended to spaces of many dimensions; a phenomenon known as the curse of dimensionality. The key contribution of neural networks in this respect is that they employ basis functions which are themselves adapted to the data, leading to efficient techniques for multidimensional problems (Bishop, 1997).

6.3 Screening of Response Vectors

The mapping problem discussed in the previous section can be adapted to the problem of screening examinees who are cheating on a test. In this context, $\mathbf{x}$ is a response score vector of an examinee with item score x_i, $i = 1, 2, ..., N$, for dichotomous scoring, x_i is a 1/0 indicator that correspond to correct/incorrect score to an item i. $\mathbf{y}$ is a vector that identifies whether or not $\mathbf{x}$ is classified as a cheater. Below is an example of how the data may look like.

$$\text{noncheater:} \quad \mathbf{x} = (11010000) \rightarrow \mathbf{y} = (0)$$

$$\text{cheater:} \quad \mathbf{x} = (11010011) \rightarrow \mathbf{y} = (1)$$

6.3.1 Implementation

The implementation of the neural network involves three major stages–data collection, selection of network architecture, and training.

Collection of Data

Collect a set of training patterns that are representative for the application envisaged. These training patterns can be thought of as a set of ordered vector pairs $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_p, \mathbf{y}_p), \ldots (\mathbf{x}_P, \mathbf{y}_P)\}$ where each $\mathbf{y}_p$ represents the output pattern vector associated with the input vector $\mathbf{x}_p$. For training purposes, it is sometimes convenient to form an input matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P)$ and an output matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_P)$.

Selection of Network Architecture

Select an architecture that is suitable for the application at hand. There are several network architectures that are available, each is designed for a specific purpose. Choosing the right architecture for a specific application is important. We will not go into detail, instead the interested reader is referred to Skapura(1996) and Zurada (1992).

For the screening problem in this study, we will propose a neural network known as the *back propagation network (BPN)* (Freeman, 1994, Chap. 3; Abdi, Valentin, & Edelman, 1999, Chap. 5) that is capable of learning a nonlinear mapping of the input $\mathbf{X}$ on the output $\mathbf{Y}$. A nonlinear mapping is necessary because the response matrix $\mathbf{X}$ is often consists of patterns that are linear combination of other response patterns. For example with dichotomous scoring in a four-item test, three of the $4^2 = 16$ response patterns could be (0,0,1,1), (0,0,1,0), and (0,0,0,1). Notice that the first pattern can be obtained as the sum (linear combination) of the remaining two patterns.

Although the input patterns are linear combinations of the other input patterns, the outputs are not necessarily linear combinations of the other outputs. Hence, the inputs are said to be not linearly separable and thus a nonlinear mapping is necessary (Skapura, 1996, p. 34).

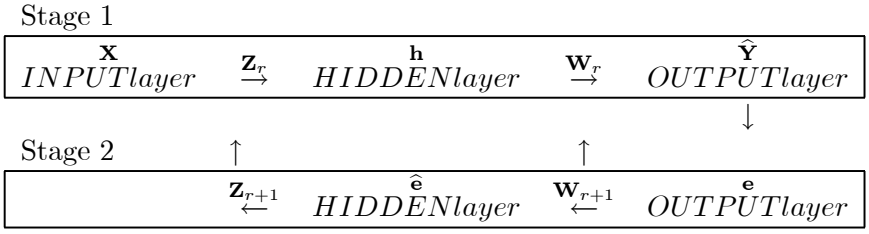
The backpropagation network is currently the most general-purpose and commonly used neural network paradigm. The BPN can be applied in many different problem situations because of the gradient-descent technique that can be used to train the network. Gradient-descent is analogous to an error minimization process that attempt to fit a closed-form solution of empirical data points that deviates from the exact value by minimal amount (Skapura, 1996).

Figure 6.1 shows the training process of a BPN. Typically, BPN has three layers—input, hidden, and the output. The response pattern $\mathbf{X}$ is presented at the input layer and propagated through the output layer via a hidden layer. The strength of the connection between the input layer and the hidden layer is represented by a weight matrix $\mathbf{Z}$ and between the hidden layer and the output layer by a weight matrix $\mathbf{W}$. As noted earlier, the goal during training is to find suitable values of $\mathbf{Z}$ and $\mathbf{W}$ that will generate a satisfactory mapping of $\mathbf{X}$ to $\mathbf{Y}$, that is the difference between the output produced by the network, $\widehat{\mathbf{Y}}$, and the desired output $\mathbf{Y}$ is minimal.

The learning process during training has two stages. In Stage 1, the input pattern $\mathbf{X}$ is converted into an activation level $\mathbf{a}$ as a weighted sum with the weight given by $\mathbf{Z}$. This activation is then converted as a response, $\mathbf{h}$, to the hidden layer using a nonlinear transfer function f (usually a logistic

function). The response $\mathbf{h}$ is then converted into an activation level $\mathbf{b}$ as a weighted sum with the weight given by $\mathbf{W}$. The activation level $\mathbf{b}$ is then used to produce a network output $\widehat{\mathbf{Y}}$. In Stage 2, the error $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ is back propagated and the connection weights are modified slightly in a direction that reduces the error. The whole process is repeated until an acceptable level of error is attained. The next section explains the more technical details of the training.

Figure 6.1. Training Process of Back Propagation Networks



Training

The training algorithm for the BPN using a logistic transfer function is discussed. Other nonlinear transfer functions is possible, for example, hyperbolic tangent and Gausian.

As noted earlier, training requires an iterative search for a set of weights that can map the input vectors to the desired output vectors. For multilayer network there are two sets of weights, $\mathbf{Z}$ for the input layer, and $\mathbf{W}$ for the hidden layer. The training proceeds as follows:

1. Initialize the values of the weight $\mathbf{Z}$ and $\mathbf{W}$. The values are obtained randomly to avoid imposing any of our own prejudices.

2. Compute the activation level $\mathbf{a}$ of the input layer,

$$\mathbf{a} = \mathbf{Z}^T\mathbf{X}, \tag{6.1}$$

where T is a matrix transpose operation.

3. Compute the response of the hidden layer with $\mathbf{a}$ as input using the logistic function, that is $\mathbf{h} = f(\mathbf{a})$ where

$$f(\mathbf{a}) = (1 + \exp(-\mathbf{a}))^{-1}. \tag{6.2}$$

The response is then forwarded to the output layer.

4. Compute the activation level $\mathbf{b}$ of the output layer,

$$\mathbf{b} = \mathbf{W}^T\mathbf{h}, \tag{6.3}$$

which becomes an input to the output layer.

5. Compute the response of the output layer with $\mathbf{b}$ as input using the logistic function

$$\widehat{\mathbf{Y}} = f(\mathbf{b}) = (1 + \exp(-\mathbf{b}))^{-1}. \tag{6.4}$$

6. Learning begins at this step. The error is computed as the difference between the computed response $\widehat{\mathbf{Y}}$ and the desired response $\mathbf{Y}$, that is,

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}. \tag{6.5}$$

7. Back propagate the error signal of the output layer

$$\boldsymbol{\delta}_{out} = f'(\mathbf{b}) \circledast \mathbf{e}, \tag{6.6}$$

where f' represents the derivative of the activation function f and $\circledast$ the elementwise product. For f chosen to be a logistic function,

$$f'(\mathbf{b}) = \widehat{\mathbf{Y}} \circledast \left(\mathbf{1} - \widehat{\mathbf{Y}}\right) \tag{6.7}$$

thus,

$$\boldsymbol{\delta}_{out} = \left(\widehat{\mathbf{Y}} \circledast \left(\mathbf{1} - \widehat{\mathbf{Y}}\right)\right) \circledast \mathbf{e} \tag{6.8}$$

where $\mathbf{1}$ is a unit matrix of the same dimension as $\mathbf{Y}$.

8. Back propagate the error signal of the hidden layer

$$\boldsymbol{\delta}_{hid} = f'(\mathbf{a}) \circledast \widehat{\mathbf{e}} \tag{6.9}$$

where $\widehat{\mathbf{e}} = \mathbf{W}\boldsymbol{\delta}_{out}$ is the estimated error of the hidden layer. With f a logistic function,

$$f'(\mathbf{a}) = \mathbf{h} \circledast (\mathbf{1} - \mathbf{h}), \tag{6.10}$$

thus,

$$\boldsymbol{\delta}_{hid} = (\mathbf{h} \circledast (\mathbf{1} - \mathbf{h})) \circledast \widehat{\mathbf{e}}. \tag{6.11}$$

9. The iterative process starts here. The weights at iteration r, $\mathbf{W}_r$ and $\mathbf{Z}_r$, are updated and the weights for the next iteration are

$$\mathbf{W}_{r+1} = \mathbf{W}_r + \lambda \mathbf{h} \left(\boldsymbol{\delta}_{out}\right)^T \tag{6.12}$$

$$\mathbf{Z}_{r+1} = \mathbf{Z}_r + \lambda \mathbf{X} \left(\boldsymbol{\delta}_{hid}\right)^T \tag{6.13}$$

where λ is small learning constant.

10. Repeat steps 1 through 9 until the output response $\widehat{\mathbf{Y}}$ is close to the desired response $\mathbf{Y}$ or until the error $\mathbf{e}$ is less than some prespecified value $\boldsymbol{\varepsilon}$.

6.4 Simulation Study

To investigate if the back propagation network can be used to identify cheating examinees we conducted a simulation study. In this simulation study we first trained the network to recognize the response patterns of a number of simulated cheaters and noncheaters. The error threshold was set to .05 and this choice is related to the criterion used to classify cheaters and and noncheaters which will be discussed shortly. The number of hidden units is usually obtained from experience, that is, by experimenting with several numbers and then judge them with respect to the training time, learning, and learning rate. No general guideline for this choice but largely depends, among others, on the type of data and the type of application at hand; see for example, Freeman and Skapura (1991, pp. 103-104). The number of hidden units used in this study was 4.

Second, we presented to the network a number of different cheaters and noncheaters that were not used in the training sample, and we investigated whether the neural networks were able to classify the patterns as cheaters and non cheaters, respectively.

6.4.1 Method

Response Model and Parameters

A test consisting of 10 items was considered. The two-parameter logistic (2PL) model (Hambleton, Swaminathan, & Roger, 1991; van der Linden & Hambleton, 1997) was used to generate the response vector. According to this model, the probability of a correct score $P_i(\theta)$ is given by

$$P_i(\theta) \equiv \Pr(X_i = 1) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]} \tag{6.14}$$

where X_i with realization x_i is a response indicator (1 for correct and 0 for incorrect), and a_i and b_i are the item discrimination and difficulty parameter, respectively. The probability of an incorrect score is $Q_i(\theta) = 1 - P_i(\theta)$.

The discrimination parameters a_i = (.63, .64, .76, .78, .92, .94, 1.02, 1.05, 1.19, 1.27) were sampled from $U(.6, 1.4)$ and the difficulty parameters equalled b_i = (-2.0, -1.6, -1.1, -0.7, -0.2, 0.2, 0.7, 1.1, 1.6, 2.0). 1000 θ values were sampled from $N(0, 1)$ and 1000 item score patterns were simulated according to the 2PL model.

Training Samples

To simulate a cheating examinee we selected the response patterns of the simulees with $\theta \leq -1.0$ and the item scores on the three most difficult items were changed such that the probability of a correct response was equal to .9. The noncheaters answered the items according to (6.14).

An item score x_i was simulated by drawing a random sample from the set $\{0, 1\}$ with chances $Q_i(\theta)$ and $P_i(\theta)$, respectively. These draws were repeated for all θ values (cheaters and noncheaters) and for all the 10 items. The response patterns of all simulated cheaters and 30% samples of the response patterns of noncheaters were used as input to train the neural network.

Application Samples, Detection Rates, and Type I Errors

In the application stage, we generated data in a similar way as in the training stage. 100 replicated samples were used to determine the detection rates. First, compute the detection rate in every replication as the proportion of correctly classified cheaters and noncheaters and then the minimum, average, and maximum detection rates across the 100 replications were determined.

When training a BPN using a logistic transfer function with binary outputs, it is standard practice to train the network to produce outputs values of {0.1,0.9} instead of the desired response values of {0,1}. Also, we should interpret output values in excess of 0.8 as active (e.g., cheaters) and values less than 0.2 as inactive (e.g., nocheaters) after training has been completed. These constraints are related to the computation of the error signal of the output layer in (6.6) which is proportional to the derivative of (6.2) evaluated at **b**. This derivative is given in (6.7) and the error signal is given in (6.8). You will notice in (6.8) that the error signal approaches 0 if the output value saturates (approaches 1 or 0). This implies that the error signal will be very small even if the actual output ($\widehat{\mathbf{Y}}$) is opposite the desired output ($\mathbf{Y}$). Thus, saturated logistic units in a BPN adapt very slowly. By setting the values of the output patterns to {0.1,0.9} reduces the training time and by allowing a margin of error around these values improves the network's ability to generalize to new input patterns (Skapura, 1996, pp. 34-36). This study adapted Skapura's suggestion: a pattern is classified to belong to a cheater when the output was greater than .8, and a pattern is classified to belong to a noncheater when the output was less than .2. An output within the interval $[.2, .8]$ was considered in transition. A simple interpretation for a pattern in transition is that the neural network is undecided given what it learned. This may be due to an insufficiently trained neural network because of small training samples, or due to training samples that are not representative of the patterns in the actual application.

The Type I error rate in every replication was computed as the proportion of misclassified cheaters and noncheaters.

6.4.2 Results

The values of the weights $\mathbf{Z}$ and $\mathbf{W}$ are shown below after $200,000$ iterations. With this sets of weights, we tested the performance of the trained networks to classify response patterns of simulated cheaters and noncheaters from replicated datasets. The set of weights are as follows:

$$\mathbf{Z} = \begin{pmatrix} -8.2020641 & -6.963136 & 8.9717280 & -11.240910 \\ -7.2407110 & -3.386511 & 4.2239189 & -2.259890 \\ -10.6518100 & -11.597602 & 18.4870411 & -9.075678 \\ -1.6791263 & -1.730923 & 7.4556612 & -9.235846 \\ -7.3057677 & -4.491081 & 0.4399706 & -12.837222 \\ -0.7995295 & -3.090211 & 3.0331101 & -7.832709 \\ 4.0820178 & -9.638773 & 1.5443495 & -3.464852 \\ 8.9167785 & 10.040429 & -11.3079910 & 1.588692 \\ -17.1608280 & 8.232890 & -10.0344857 & 16.380543 \\ -18.7282075 & 15.868412 & 24.3978266 & 35.367935 \end{pmatrix}$$

$$\mathbf{W} = \begin{pmatrix} -6.371597 \\ 2.393118 \\ -2.214164 \\ 1.929745 \end{pmatrix}$$

Given an input pattern $\mathbf{X}$ and the set of weights $\mathbf{Z}$ and $\mathbf{W}$, the predicted output of the networks can be written compactly as

$$\widehat{\mathbf{Y}} = f(\mathbf{W}^T f(\mathbf{Z}^T \mathbf{X})) \tag{6.15}$$

where f is defined in (6.2).

Type I Errors/Detection Rates

Table 6.1 shows the minimum, the mean, and the maximum Type I error and detection rates across 100 replications for different situations.

The results revealed that the Type I errors were low and the detection rates were high and very similar between cheaters and noncheaters.

Robustness

To investigate the robustness of the results, we used the same patterns as simulated in the training stage but we changed the patterns in the application stage. We first simulated cheaters from a population with a higher ability distribution than the ability distribution of the cheaters in the training samples. The cheaters were selected with $\theta \leq 0$ (instead of $\theta \leq -1$) and the probability of a correct response to the 3 most difficult items for these cheaters were set to .9 as before.

Table 6.1: Type I Error and Detection Rates

	Detection Rate			Type I Error Rate		
Situations	min.	mean	max.	min.	mean	max.
Noncheaters	.89	.91	.93	.01	.03	.06
Cheaters	.89	.92	.95	.01	.03	.04
Robustness						
(a)	.83	.86	.87	.04	.05	.07
(b)	.78	.83	.87	.02	.05	.07
(c)	.48	.55	.62	.16	.19	.22

The detection and Type I error rates were indicated as situation (a) in Table 6.1. As expected, a decrease in the detection rates and an increase in the Type I error rates were observed, although the differences were not very large.

Second, we also investigated the robustness by changing the difficulty level and the number of items copied. Two cases were investigated. In the first case, the probability of a correct response to the 2 most difficult items were set to .9. Results are in situation (b) in Table 6.1. In the second case, the items were first sorted in increasing order of difficulty and then the probability of a correct response to the 8th and 9th items were set to .9. Results are in situation (c) in Table 6.1. In the first case, the Type I errors were acceptable and the detection rates were still high. In the second case, the detection rates decreased substantially and the Type I errors were inflated.

6.5 Discussion

In this study we investigated the usefulness of neural networks as a tool to screen for cheating. Although the scope of simulation study was limited, the results provide some insights about the potential usefulness of neural networks for cheating detection. The results suggest that a neural network is able to recognize cheating behavior when the response patterns in the training and the application stage have a similar configuration. The small

robustness study showed that the power may decrease when the configuration of item scores in the training and the application stage differ. This implies that we should have sufficient training data so that the network can recognize different types of cheating behavior.

The screening of cheating by means of neural networks seems to be useful when we have little or no information about the identity of the source and the copier or when cheating is the result of, for example, item preknowledge. Furthermore, data should be available to train the network. That is, data from the same test with information about possible cheating behavior should be available. This seems realistic when a test is administered more than once to groups of examinees.

A future research area may be the relation between person-fit statistics and neural networks to detect misfitting item score patterns (Meijer, 2003; Emons, 2003). A drawback of most current person fit research is that misfitting response behavior is defined based on an IRT model, that is, misfitting or aberrant behavior is defined as behavior that is unexpected given that an IRT model fits the data. This is not the case using a neural network approach. One may in advance train the neural network to recognize a particular type of aberrant response behavior of interest, and then screen the data to recognize such a pattern.

Finally, when applying neural networks to identify particular types of response vectors sometimes a large number of possible response vectors should be presented to the system. Feeding all possible vectors into the network is often too time consuming. Because the number grows exponentially in the number of items in the test, for larger tests this approach is even infeasible. However, some of these vectors are more critical than others. An important question to be addressed in future studies is how to optimize the sample of response vectors used in the training stage.

Chapter 7

Summary

This thesis concerns the development and evaluation of statistical methods for the detection of answer copying on achievement tests.

At the Educational Testing Service (ETS), the largest testing company in the United States, the K-index is used as statistical evidence when pursuing cases of cheating in tests, such as, for example, the Scholastic Aptitude Test (SAT). There is not much known about the statistical properties of the K-index, therefore, after a short introduction in Chapter 1 in Chapter 2 a simulation study was conducted to investigate the statistical properties of the K-index in particular its Type I error and detection rate for relatively small sample sizes (100 persons) and relatively large sample sizes (500 persons). The results presented in Chapter 2 showed that the K-index can be used for small sample sizes because its Type I error rate was close to the nominal level but the detection rate was low when the number of items copied was 20% or less. Results further showed that the detection rate of the K-index for sample sizes of 100 and 500 simulees was smaller compared to the detection rate of another copying index based on item response theory, ω. An approximation to K-index, $\overline{K}_2$, that used a quadratic function to estimate the parameter of the binomial distribution improved the detection rate of the K-index to a considerable extent.

The K-index is computed based on the number of similar incorrect responses of the source and the copier that is assumed to follow a binomial distribution. Two additional modifications to the K-index were proposed in Chapter 3. The first modification is called the S_1 index and it uses a Poisson distribution to model the number of similar incorrect responses instead of the binomial distribution that was used in the original version of the K-index. The second modification is called the S_2 index and is computed

using the number of similar incorrect and weighted correct responses of the source and the copier. The motivation behind the S_2 index is to incorporate additional information about copying that is contained in the similar correct response. However, instead of using a uniform weight of 1 for all similar correct responses, a weight is introduced that depends on the probability of a correct response. The key idea is simple: when a copier answers an item correctly that was also answered correctly by the source, the weight is high when the estimate of the probability of a correct response is low, and the the weight is low when the estimate of the probability of a correct response is high. This idea leads to a transformation function that is monotonely decreasing with the probability of a correct response. In addition, the S_2 index was proposed in order to overcome a limitation of the K-index, namely, its insensitiveness to copying correct answers. Results of a simulation study revealed that the detection rate of S_1 index was considerably higher than the detection rate of $\overline{K}_2$. The S_2 index which incorporates information from the similar correct scores in addition to the similar incorrect–scores yielded a significant improvement in detection rate over the S_1 index.

Chapters 4 and 5 present two statistical tests for the detection of answer copying on multiple-choice tests. Both tests only use information from the response patterns of the source and the copier. They differ in the way the response process is modeled and also on the type of conditioning employed.

The statistical test presented in Chapter 4 is based on the idea that the answers of examinees to test items may be the result of three possible processes: (1) knowing, (2) guessing, and (3) copying, but that examinees who do not have access to the answers of other examinees can arrive at their answers only through the first two processes. This assumption leads to a distribution for the number of matched incorrect alternatives between the examinee suspected of copying and the examinee believed to be the source that belongs to a family of "shifted binomials". An analyses of the power functions of the test for several sets of parameter values showed that the test has considerable power to detect copying on multiple-choice tests, particularly if the number of response alternatives per item goes up. But even for a test for three-choice items, the power to detect copying in general is high.

The key observation on which the statistical test in Chapter 5 rests is that if the responses by the copier and the source are probabilistic and the copier did not have access to the answers by the source, the responses are statistically independent. If the copier did have access to some of these answers and copied them, the responses of the copier and the source on these items would not only be dependent but even in perfect agreement. The

statistical test is based on Cohen's kappa (Cohen, 1960). The test is free of any assumptions on the response processes of the examinees suspected of copying and having served as the source, except for usual assumption that these processes are probabilistic. Because the asymptotic null and alternative distributions of the kappa statistic are derived under the assumption of common marginal probabilities for all items, a recoding of the item alternatives is proposed to approximate this case. The results from a simulation study in this paper show that under this recoding the test approximates its nominal Type I error rates and has promising power functions. More experience, however, is needed with a larger variety of test lengths, response probabilities, and score ranges conditional on which we should recode to be able to generalize the results in the simulation study.

In chapter 6, a conceptually new approach to screening data from high-stakes tests for possible cases of cheating using a neural network was proposed. The idea is to use this approach as a supplement to existing screening practices in order to further improve test security measures. The chapter gives an overview of the basic principles of neural networks and discussed how this technique can be applied to identify cheaters. It appears that the screening of cheating by means of neural networks seems is useful when we have little or no information about the identity of the source and the copier or when cheating is the result of, for example, item preknowledge. Furthermore, data should be available to train the network. That is, data from the same test with information about possible cheating behavior should be available. This seems realistic when a test is administered more than once to groups of examinees.. The results of a small simulation study showed that a neural network approach has high power when the configuration of the item scores of a cheater has similar characteristics as the configuration used to train the network. Some directions for future research are discussed.

Chapter 8

Samenvatting

In deze studie worden verschillende statistische methoden om bedrog op examens op te sporen voorgesteld en geëvalueerd, met name die vorm van bedrog waarbij een persoon (de "copier") de antwoorden van iemand anders (de "source") overschrijft. Er bestaan verschillende statistische indices en tests om afkijkgedrag op te sporen. Een veelgebruikte index is de K-index. Deze wordt onder andere gebruikt bij Educational Testing Service (ETS) in de Verenigde Staten om afkijkgedrag op de Scholastic Aptitude Test (SAT) op te sporen. Er is echter nog niet veel bekend over de statistische eigenschappen van deze index.

Nadat in Hoofdstuk 1 een algemene inleiding is gegeven, wordt in Hoofdstuk 2 uitgebreid simulatieonderzoek gerapporteerd naar de eigenschappen van de K-index. Zowel de fout van de eerste soort (type I fout) als het onderscheidend vermogen (power) van de K-index worden onderzocht. Bovendien wordt de power vergeleken met een andere bestaande index, ω. Dit wordt zowel voor relatief kleine (100 personen) als grote steekproeven (500 personen) uitgevoerd. Resultaten van de simulatiestudie laten zien dat de K-index geschikt is om te worden gebruikt in relatief kleine steekproeven: de empirische type I fout komt overeen met de nominale type I fout. Het onderscheidend vermogen van de index is echter laag wanneer op minder dan 20% van de items het antwoord is overgeschreven van een andere examinandus. Resultaten laten verder zien dat het onderscheidend vermogen van de K-index in veel situaties lager is dan ω. In Hoofdstuk 2 wordt verder nog een benadering van de K-index voorgesteld die een kwadratische functie gebruikt om de geschatte parameter van de binomiale verdeling te verbeteren.

In Hoofdstuk 3 worden twee nieuwe statistische methoden voorgesteld gebaseerd op de K-index. De K-index is gebaseerd op het vergelijken van het

aantal overeenkomstige foute antwoorden tussen de "source" en de "copier". De eerste methode is de S_1 index die een Poisson verdeling gebruikt om het aantal overeenkomstige foute antwoorden te modelleren in plaats van de binomiale verdeling die wordt gebruikt voor de K-index. De tweede methode die wordt voorgesteld is de S_2 index die naast het aantal overeenkomstige foute antwoorden, het aantal overeenkomstige goede antwoorden modelleert. De S_2 index wordt voorgesteld om additionele informatie over afkijkgedrag mee te modelleren. Wanneer een "copier" een goed antwoord geeft op een item waarvan de kans op een goed antwoord in de populatie groot is, dan krijgt deze response een laag gewicht. Wanneer echter een "copier" een goed antwoord geeft waarbij de kans op een goed antwoord klein is, dan krijgt dit antwoord een relatief groot gewicht. Aan de hand van gesimuleerde data wordt aangetoond dat de S_1 en S_2 indices een hogere power hebben dan de K-index.

In de Hoofdstukken 4 en 5 worden twee statistische toetsen voorgesteld om afkijkgedrag op te sporen die alleen gebaseerd zijn op informatie die kan worden gehaald uit het response patronen van de "source" en de "copier". Er wordt dus geen informatie gebruikt die afhankelijk is van de populatie waartoe de "source" en de "copier" behoren. De twee methoden verschillen in de wijze waarop het antwoordgedrag is gemodelleerd.

De statistische toets die wordt voorgesteld in Hoofdstuk 4 is gebaseerd op het idee dat de antwoorden op een examen het resultaat kunnen zijn van drie verschillende mechanismen: (1) kennis van het goede antwoord (2) gokken van het goede antwoord, en (3) afkijken. Verder wordt uitgegaan van de assumptie dat personen die niet afkijken het goede antwoord alleen kunnen geven door middel van de eerste twee opties. Deze assumptie leidt tot een familie van "shifted" binomiale verdelingen voor het aantal overeenkomstige foutieve antwoorden voor de "copier" en de "source". Een analyse van de power functies voor verschillende configuraties van parameter waarden laat zien dat de toets een hoog onderscheidend vermogen heeft, met name wanneer het aantal response alternatieven per item relatief groot is.

In Hoofdstuk 5 wordt ervan uitgegaan dat als de antwoorden van de "-copier" and de "source" het gevolg zijn van een stochastisch proces en dat wanneer de "copier" geen kennis heeft van de antwoorden van de "source" de antwoorden van de copier en de "source" statistisch onafhankelijk zijn. Als de "copier" wel toegang heeft tot sommige antwoorden van de "source", dan zijn de antwoorden van de copier en de "source" niet alleen afhankelijk, maar zelfs in perfecte overeenstemming. De statistische toets die wordt gebruikt in Hoofdstuk 5 is gebaseerd op Cohen's kappa. De toets is niet gebaseerd op enige assumpties van het response proces, behoudens dat deze probabilis-

tisch is. Omdat de asymptotische nul en alternatieve verdelingen van kappa zijn afgeleid onder de assumpties van gemeenschappelijke marginale kansen voor alle items, wordt een hercodering van de item alternatieven voorgesteld om deze situatie te benaderen. De resultaten van een simulatiestudie laten zien dat gegeven deze hercodering de empirische en nominale type I fout met elkaar in overeenstemming zijn. Verder onderzoek dient uit te wijzen in hoeverre deze resultaten zijn te generaliseren voor verschillende tests en kansen op een goed antwoord.

In hoofdstuk 6 wordt een conceptueel nieuwe benadering, het gebruik van neurale netwerken, voorgesteld om bedrog op te sporen binnen "high stakes" test situaties. Het idee is dat deze benadering kan dienen als aanvulling op bestaande statistische toetsen. Dit hoofdstuk geeft een overzicht van de principes waar de techniek van neurale netwerken op gebaseerd is en geeft aan hoe deze techniek kan worden gebruikt om bedrog op te sporen. Het blijkt dat het gebruik van neurale netwerken nuttig kan zijn wanneer weinig of geen informatie over de identiteit van de "copier" en de "source" bekend is of wanneer bedrog het gevolg is van voorkennis. Bovendien dienen data aanwezig te zijn om het netwerk te trainen. Dit is realistisch wanneer een toets meerdere malen wordt afgenomen aan verschillende groepen van respondenten. Resultaten van een simulatiestudie laten zien dat de neurale netwerk benadering hoge power heeft wanneer de configuratie van item scores van een bedrieger dezelfde eigenschappen heeft als de configuratie van item scores die wordt gebruikt om het netwerk te trainen.

Chapter 9

References

Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks* (Sage University Papers Series on Quantitative Application in the Social Sciences, series no.07-124). Thousand Oaks, CA: Sage.

Agresti, A. (1990). *Categorical data analysis.* NY: Wiley.

Agresti, A. (1996). *An introduction to categorical data analysis.* NY: Wiley.

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

Bay, M. L. G. (1994). Detection of copying on multiple-choice examinations (Doctoral dissertation, Southern Illinois University, 1987). *Dissertation Abstracts International, 56(3-A)*, 899.

Bay, L. G. (1995). *Detection of cheating on multiple-choice examinations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bishop, C. M. (1997). Neural networks: a pattern recognition perspective, *Handbook of neural computing* (online), IOP Publishing Ltd and Oxford University Press.

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

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 33-49). New York: Springer-Verlag.

Casella, G., & Berger, R. L. (1990). *Statistical inference.* Pacific Grove, CA: Brooks/Cole.

Cizek, G. J. (1999). *Cheating on tests: how to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum.

Cohen, J. (1960). A coeficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

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

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large-sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 6,* 152-165.

Freeman, J. A. (1994). *Simulating neural networks with mathematica.* Menlo Park, California: Addison-Wesley Publishing Company.

Freeman, J. A., & Skapura, D. M. (1991). *Neural networks: algorithms, applications, and programming techniques.* Menlo Park, California: Addison-Wesley Publishing Company.

Good, P. I. (2001). *Applying statistics in the courtroom: A new approach for attorneys and expert witnesses.* Chapman & Hall/CRC: Boca Raton, Florida.

Hambleton, K. H., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, California: Sage.

Hanson, B. A. (1994). *Statistical indexes of response similarity derived from the compound binomial distribution.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Hemker, B. T., Sijtsma, K., Molenaar, W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of total score. *Psychometrika, 61*, 679-693.

Hemker, B. T., Sijtsma, K., Molenaar, W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331-347.

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

Kamp, Y., & Hasler, M. (1990). *Recursive network for associative memory.* Baffins Lane, Chichester: Wiley.

Lehmann, E. (1999). *Elements of large-sample theory.* New York: Springer Verlag.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

MathSoft. (2000). S-Plus 2000 [Computer software and manual]. Seatle, WA: Author.

Meijer, R. R. (1998). Consistency of test behavior and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71*, 147-160.

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

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

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.), Boston: McGraw-Hill.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 543-570.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: some powerful and practicable procedures. *Psychological Bulletin, 110,* 577-586.

Skapura, D. M. (1996). *Building neural networks.* New York: ACM Press.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement,* 39, 115-132.

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

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2003). Detecting answer copying using statistic kappa. Manuscript submitted for publication.

Sotaridona, L. S. (2003). Screening of cheating on high-stakes test using neural networks. Manuscript submitted for publication.

Teamsters v. U.S., 431 U.S. 324 (1977).

Thissen, D. (1991). MULTILOG *user's guide* (Ver. 6). Chicago: Scientific Software, Inc.

Thissen, D., & Steinberg, L. (1997). A response model for multiple choice items. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer-Verlag.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

van der Linden, W. J., & Sotaridona, L. S. (in press). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement.*

van der Linden, W. J., & Sotaridona, L. S. (2003). Detecting answer copying when the regular response process follows a known response model. Manuscript submitted for publication.

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

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

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

Zurada, J. M. (1992). *Introduction to artificial neural system.* St. Paul, MN: West Publishing Company.