THE POWER OF THE K-INDEX (OR PMIR) TO DETECT COPYING

Charles Lewis
Dorothy T. Thayer

ETS
Educational Testing Service
Princeton, New Jersey

December 1998
THE POWER OF THE K-INDEX
(or PMIR)
TO DETECT COPYING

Charles Lewis
and
Dorothy T. Thayer
Educational Testing Service
Princeton, NJ 08541
Abstract

At Educational Testing Service, the K-Index is used to assess unusual agreement between the incorrect responses of two test takers on a multiple-choice test. Since the Fall of 1996, the value of the K-Index used for this assessment has been adjusted, based on the Bonferroni inequality. The resulting value is referred to as the Probability of Matching Incorrect Responses (PMIR). If a test taker copies from someone (a Source) with no incorrect responses, the PMIR is useless \( i.e., \) has zero power for detecting unusual agreement. Similarly, if a Source has relatively few incorrect responses, the power of the PMIR will be low. We propose a framework within which to study the power of the PMIR. The power function derived within this framework exhibits quite complex behavior; however, we conclude that the power of the PMIR to detect substantial amounts of copying is quite low, even when the Source has a relatively large number of incorrect responses.

Keywords: power; detection of copying; Bonferroni; detecting unusual agreement; binominal; significance testing
A Primer on Detection of Copying at ETS

The copy detection process at ETS begins when a test taker (referred to as the Subject) is identified for attention by the Test Security Office. This may occur for any of a number of reasons. One such reason is the existence of a large score difference between the Subject's score on a test and their score on a previous administration of the same test. Once this identification has taken place, the Subject's responses are compared to those seated nearby (referred to as potential Sources), as determined by a seating chart and/or test booklet numbers. The statistic used to compare the Subject with a potential Source is the number of items on the test (or portion of the test, here referred to as a section) for which both test takers gave identical (or matching) incorrect responses.

A 'model-based' probability (the Kling Index, or K-Index) is computed for each potential Source and test section. It is an estimate of the chance that at least the observed number of matching incorrect responses would occur if the two test takers being compared were working independently on the section of the test being analyzed. If this probability is sufficiently small, it is taken as an indication of 'unusual agreement' between the two sets of incorrect responses being compared.

A recent modification involves multiplying the K-Index by the number of comparisons being made (i.e., the number of potential Sources times the total number of tests/sections compared per Source). The result is referred to as the PMIR: Probability of Matching Incorrect Responses. This modification is intended to control the 'per Subject' Type I error rate, making use of the Bonferroni inequality. (A Type I error in this case consists of rejecting the null hypothesis that two test takers were working independently when the hypothesis is actually true.)
Using a critical value of 0.01 for the PMIR, this means that, if a Subject working independently of other test takers had their responses analyzed by this procedure, the chances should be less than 0.01 that they would be identified as having ‘unusual agreement’ of incorrect responses with any of the potential Sources.

Suppose that, instead of using the PMIR, we were to work with the K-Index, using a critical value of 0.0001. This would be equivalent to using the PMIR with a critical value of .01 and making 100 comparisons per subject. In what follows, we shall limit our discussion to the K-Index, with appropriate adjustment to the critical value, reflecting the number of comparisons being made.

A primer on the K-Index

The reference for details on the derivation and properties of the K-Index is Holland (1996). The ‘model’ on which the K-Index is based is that the number of matching incorrect responses (WM) for two test takers (A and B) who are working independently has a binomial sampling distribution. The sampling modeled by this distribution is conditional on the numbers of incorrect responses for the two test takers (WA and WB) and on the test (or section) under consideration, including the number of items (N). For clarity, we assume WA ≥ WB. Then it is clear that WM ≤ WB. In the model, we take WB as the number of trials for the binomial.

As described in Holland (1996, p.17), the formula used to find the “success” probability (P) for the K-Index “was originally developed and studied empirically by F. Kling.” This formula is a continuously piecewise linear function of WA/N, the proportion of incorrect responses made by test taker A. It has a “slope” parameter (b) that controls how steeply the success probability increases with WA/N. The formula is
\[ P = \begin{cases} 0.085 + b(WA/N), & \text{if } 0 < (WA/N) \leq 0.3. \\ \left[0.085 + b(0.3)\right] + 0.4b((WA/N) - 0.3), & \text{if } 0.3 < (WA/N) \leq 1.0 \end{cases} \] \tag{1}

This function is illustrated in Figure 1 with the slope \( b \) set equal to 0.5, the value used with the SAT I verbal test. More generally, a value of \( b \) must be specified for every test with which the K-Index is used. Holland (1996, section 5) studied the choice of \( b \) for several ETS testing programs by empirical evaluation of agreement probabilities. He found that, with appropriate values for \( b \), that the formula (1) generally provided an upper bound on the empirical agreement proportions.

(Insert Figure 1 here)

The formula for computing the K-Index is given by

\[ K = \sum_{i=0}^{WB} \binom{WB}{i} P^i (1-P)^{WB-i}. \] \tag{2}

This is just the binomial probability of observing at least \( WM \) successes out of \( WB \) trials, with probability of success \( P \) given by (1). Holland (1996) argues that the value of the K-Index based on (1) and (2) will be an over-estimate of the 'actual' probability of at least \( WM \) matches for two test takers having \( WA \) and \( WB \) incorrect responses if they are working independently. This conservatism is good for controlling the Type I error rate but, as we shall see, bad for power.

We can compare the binomial distribution on which the K-Index and PMIR are based with an empirical distribution to illustrate the discrepancy between the two. For our example, we chose the 11/94 administration of the SAT I and looked at a section of the Verbal test having \( N=36 \) items. We compared all 11,287 test takers having \( WA=11 \) incorrect responses for this section to the 11,367 test takers with \( WB=10 \) incorrect responses. This gave us a total of 128,299,329 values of \( WM \). The resulting relative frequency distribution is displayed in Figure 2, together with the corresponding binomial distribution of the Kling model.
Clearly, the binomial model is displaced to the right of the empirical distribution. To allow us to see details of the upper tails of the two distributions, they are plotted again in Figure 3, this time using a logarithmic scale for the relative frequencies (probabilities). Since the K-Index is the sum of the probabilities in the upper tail of the distribution, the fact that these probabilities all exceed the corresponding relative frequencies implies that the K-Index in this case provides an over-estimate of the upper tail relative frequencies, consistent with Holland's analysis.

What about power?

To study power, we need a model for 'copying.' (The K-Index is based on a model for working independently.) Our model is very simple, and is a direct extension of the Kling model. We assume that the Subject (A) copies a **randomly** selected set of NC items from the Source (B).

To introduce the notation for this model, consider the three “2x2 plus 1” contingency tables (following Holland, 1996) represented in Figure 4. The first table classifies all N items according to whether they were correctly answered (or omitted) or incorrectly answered by the Subject (A) and by the Source (B). In addition, the (Incorrect, Incorrect) classification is divided into two cells, depending on whether or not the two incorrect answers match. (This is what the “plus 1” terminology refers to.) The only labeled cell in the table is the number of matching incorrect items, WM. The marginal numbers of incorrect items are identified by WA and WB, respectively. This first table in Figure 4 summarizes the notation used to describe the application of the K-Index.
The second and third tables in Figure 4 provide the same classifications for the NC items copied by A from B and for the NI (= N - NC) items answered independently by A. In the second table, note that all WC items in the (Incorrect, Incorrect) classification have matching incorrect answers (since A copied them from B).

First we consider the distribution of WC, conditioned on fixed values for WA and WB (WA ≥ WB). If there were no further restrictions on WC, it would have a hypergeometric sampling distribution given by

\[ p(\text{WC}|\text{N}, \text{NC}, \text{WB}) = \binom{\text{WB}}{\text{WC}} \frac{\binom{\text{N} - \text{WB}}{\text{NC} - \text{WC}}}{\binom{\text{N}}{\text{NC}}} , \]  

(3)

with WC taking on integer values from 0 to WB. However, by comparing the first two tables in Figure 4, we can see that there is one additional restriction on WC, namely that the number of copied responses that are not incorrect (NC - WC) must be less than or equal to the total number of A's responses that are not incorrect (N - WA):

\[ \text{NC} - \text{WC} \leq \text{N} - \text{WA} \]

or

\[ \text{WC} \geq \text{WA} - \text{N} + \text{NC} . \]  

(4)

Note that this restriction is only relevant when WA + NC exceeds the total number of items on the test. With this restriction, WC has a truncated hypergeometric distribution, given WA, WB, N, and NC:

\[ p(\text{WC}|\text{N}, \text{NC}, \text{WA}, \text{WB}) = \frac{p(\text{WC}|\text{N}, \text{NC}, \text{WB})}{\text{Pr}(\text{WC} \geq \text{WA} - \text{N} + \text{NC})} . \]  

(5)

In addition, we suppose that there are WMI matching incorrect responses due to chance, as represented in the third table in Figure 4, so that the total number of matching incorrect responses is given by
\[ WM = WC + WMI. \]  \hspace{1cm} (6)

Given \( N, WA, WB, NC \) and \( WC \), we assume that \( WMI \) follows the 'Kling' binomial distribution with number of trials given by

\[ WBI = WB - WC, \]  \hspace{1cm} (7)

and probability of success using the Kling formula with the slope \( (b) \) for the test, and proportion incorrect \( WAI/NI \), where

\[ WAI = WA - WC \]  \hspace{1cm} (8)

and

\[ NI = N - NC. \]  \hspace{1cm} (9)

In other words, the \( NI \) items that \( A \) answers independently of \( B \) are taken to comprise a 'mini-test,' with numbers of incorrect responses for \( A \) and \( B \) given by \( WAI \) and \( WBI \), respectively, and number of matching incorrect responses given by \( WMI \). The analogy with the Kling model for the full test can be seen by comparing the first and third tables in Figure 4.

Taken together, the distributions for these two components \((WC \text{ and } WMI)\) allow us to write an expression for the probability of observing \( WM \) matching incorrect responses, given that \( A \) copies a random \( NC \) of the \( N \) items on a test and given that \( A \) and \( B \) have \( WA \) and \( WB \), incorrect responses, respectively:

\[
p(WM|N, NC, WA, WB) = \sum_{WC} p(WMI|NI, WAI, WBI)p(WC|N, NC, WA, WB), \hspace{1cm} (10)
\]

where the sum is over all permissible values of \( WC \) and, from (6), \( WMI \) is given by \( WM - WC \).

Now let us return to our example \((N=36, WA=11, WB=10)\). Suppose our total number of comparisons is 40. Then, to control the overall significance at .01, our per comparison alpha level becomes 0.01/40=.000250. The K-Index for \( WM=9 \) is \( K=.000018 (<0.01/40) \). For \( WM=8 \), we have \( K=.000272 (>0.01/40) \), so we would take as our critical value \( WM^* = 9 \). In other words, based on the Kling model, we
would only declare 'unusual agreement' in this case when \( WM \) is at least 9. The corresponding critical value based on our empirical distribution is \( WM^* = 7 \), with a relative frequency of 7 or more matches of 0.000083 (<0.01/40). Consequently, we may expect a substantial loss in power in this case associated with basing the test of significance on the theoretical (Kling binomial model) rather than the empirical sampling distribution.

If the Subject (A) is copying half of the Source's (B's) answers \( (NC=18) \), the theoretical power, based on (10), to detect this amount of copying is only 0.04, based on the Kling critical value of \( WM^* = 9 \), while it increases to 0.42, based on the empirical critical value of 7. When 3/4 of the answers are copied \( (NC=27) \), the corresponding two theoretical power values are 0.40 and 0.94, respectively. These results support the concern regarding a loss of power due to the use of theoretical, rather than empirical, critical values for our tests.

There is another point to consider: If the K-Index values are over-estimates of upper tail probabilities, what does this say about our theoretical values for the power? Since they are also upper tail probability estimates based on the Kling model, presumably they may be over-estimates as well. To investigate this possibility, we went back to our example and simulated different amounts of random copying.

Specifically, we again took as Sources all examinees with 10 incorrect responses on the verbal section with 36 items (11,367 examinees). For each of these Sources, we sampled 14 examinees from the total population. We called these examinees our potential Subjects. For each potential Subject, we took a random selection of \( NC \) items (with \( NC = 18 \) or 27) and changed these responses to match the responses of the Source. Next, we counted the new total number of incorrect responses for this potential Subject (new \( WA \)), the number of incorrects matching those of the Source (\( WM \)) and accumulated these in an array with Subject incorrects indexing rows and matching incorrects indexing columns. We
repeated the random selection of NC copied items for each potential Subject 15 times. This gave us approximately 200 draws (14 x 15 = 210) for each Source. After carrying out this procedure for all our Sources, we used the resulting array to compute the empirical power for Subjects with (new) WA = 11, using 7 or 9 matching incorrects as critical values for the test.

For NC = 18, the proportion of cases with matching incorrects greater than or equal to 7 was equal to 0.34 and the proportion of cases with matching incorrects greater than or equal to 9 was equal to 0.02. (Compare these to the corresponding theoretical power values of .42 and .04 cited above.) For NC = 27, the proportion of cases with matching incorrects greater than or equal to 7 was equal to 0.95 and the proportion of cases with matching incorrects greater than or equal to 9 was equal to 0.38. (Again, these should be compared to the corresponding theoretical values of .94 and .40) Table 1 contains these comparisons and extends them to a range of values for WA and WB.

(Insert Table 1 here)

As we anticipated, in most of the cases given in Table 1, the theoretical power does exceed the empirical power. The exceptions (for instance, in the case where NC = 27, WB = 10, WA = 11, WM* = 7) all occur when both theoretical and empirical power exceed .90. The largest discrepancy between theoretical and empirical power among the tabled values occurs for NC = 18, WB = 25, WA = 26, WM* = 17. This discrepancy is equal to .11.

To expand on the results for this example, we carried out a similar empirical power analysis, this time for the total SAT I Verbal test (N = 78 items). These results are contained in Table 2. Once again, we see a general pattern of theoretical power values exceeding corresponding ones for empirical power (with a maximum discrepancy of .19). Based on these tables, we might tentatively conclude that the theoretical power values may be used, but with considerable
caution, recognizing that some over-estimation is likely to occur, and substantial over-estimation may occur.

(Insert Table 2 here)

The two tables also support the earlier conclusion that considerable power may be lost by using the model-based, rather than empirical, critical values for carrying out tests of the hypotheses that examinees are working independently of those around them. The only (trivial) exceptions seem to be cases where the power for the theoretical value is already close to unity.

Another important point that can be made with these tables is that our test procedure has limited power to detect substantial amounts of copying. Considering empirical power and empirical critical values, to have a probability of at least .80 of detecting a Subject who copies 18 out of 36 items, the Source from whom they are copying must have more than 20 incorrect responses (again, out of 36 items). To be able to identify a Subject who copies 36 out of 78 items with probability .80, their Source must also have more than 20 incorrect responses. Note that, if we are using model-based critical values for these tests, the empirical power to detect these amounts of copying never exceeds .50 for any of the tabled values.

**Conclusion**

The PMIR procedure has relatively low power to detect substantial amounts of 'random' copying. This is partly a function of the over-estimation of tail probabilities by the Kling model. It is also partly a function of the large numbers of comparisons that are typically made in the effort to identify a Source for a given Subject.
Reference

Table 1

Empirical and Theoretical Power of PMIR Test for Alpha = 0.01/40
SAT I Verbal Section: 36 Items

<table>
<thead>
<tr>
<th>NC = 18</th>
<th>NC = 27</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_B = 5$, $W_A = 6$</td>
<td>$W_B = 10$, $W_A = 11$</td>
</tr>
<tr>
<td>$W_M^* = 5$</td>
<td>$W_M^* = 7$</td>
</tr>
<tr>
<td><strong>Empirical</strong></td>
<td><strong>Theoretical</strong></td>
</tr>
<tr>
<td>Empirical</td>
<td>0.0216</td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.0555</td>
</tr>
<tr>
<td>$W_B = 15$, $W_A = 16$</td>
<td>$W_B = 20$, $W_A = 21$</td>
</tr>
<tr>
<td>$W_M^* = 10$</td>
<td>$W_M^* = 11$</td>
</tr>
<tr>
<td><strong>Empirical</strong></td>
<td><strong>Theoretical</strong></td>
</tr>
<tr>
<td>Empirical</td>
<td>0.3360</td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.4241</td>
</tr>
<tr>
<td>$W_B = 25$, $W_A = 26$</td>
<td>$W_B = 25$, $W_A = 26$</td>
</tr>
<tr>
<td>$W_M^* = 15$</td>
<td>$W_M^* = 17$</td>
</tr>
<tr>
<td><strong>Empirical</strong></td>
<td><strong>Theoretical</strong></td>
</tr>
<tr>
<td>Empirical</td>
<td>0.8335</td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.8763</td>
</tr>
<tr>
<td>Power</td>
<td>NC = 36</td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>Empirical</td>
<td>W_M = 6</td>
</tr>
<tr>
<td>Empirical</td>
<td>0.3482</td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.5060</td>
</tr>
<tr>
<td>Empirical</td>
<td>W_M = 8</td>
</tr>
<tr>
<td>Empirical</td>
<td>0.5741</td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.7307</td>
</tr>
<tr>
<td>Empirical</td>
<td>W_M = 10</td>
</tr>
<tr>
<td>Empirical</td>
<td>0.7603</td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.8731</td>
</tr>
<tr>
<td>Empirical</td>
<td>W_M = 12</td>
</tr>
<tr>
<td>Empirical</td>
<td>0.8781</td>
</tr>
<tr>
<td>Theoretical</td>
<td>0.9420</td>
</tr>
</tbody>
</table>
Figure 1:
K-Index Model for Probability of Matching Incorrect Response
(b = 0.5)
Figure 2:
Number of Matching Incorrect Responses for an
SAT I Verbal Section with 36 Items
(WA = 11, WB = 10)

- □ Empirical Relative Frequency
- ▼ Binomial Probability (K-Index)
Figure 3:
Number of Matching Incorrect Responses for an
SAT I Verbal Section with 36 Items
(WA = 11, WB = 10)

- □ Empirical Relative Frequency
- ● Binomial Probability (K-Index)

Relative Frequency/Probability

Number of Matches

0 1 2 3 4 5 6 7 8 9 10

10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^{0}
Figure 4
Notation used to describe the application of the K-Index

All items classified in 2 x 2 plus 1 Table

<table>
<thead>
<tr>
<th>Subject A Responses</th>
<th>Source B Responses</th>
<th>Correct + Omit</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct + Omit</td>
<td>-</td>
<td>-</td>
<td>N - WA</td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>-</td>
<td>WM / -</td>
<td>WA</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>N - WB</td>
<td>WB</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

Copied items classified in 2 x 2 plus 1 Table

<table>
<thead>
<tr>
<th>Subject A Responses</th>
<th>Source B Responses</th>
<th>Correct + Omit</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct + Omit</td>
<td>NC - WC</td>
<td>0</td>
<td>NC - WC</td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>WC / 0</td>
<td>WC</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>NC - WC</td>
<td>WC</td>
<td>NC</td>
<td></td>
</tr>
</tbody>
</table>

Independent items classified in 2 x 2 plus 1 Table

<table>
<thead>
<tr>
<th>Subject A Responses</th>
<th>Source B Responses</th>
<th>Correct + Omit</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct + Omit</td>
<td>-</td>
<td>-</td>
<td>NI - WAI</td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>-</td>
<td>WMI / -</td>
<td>WAI</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>NI - WBI</td>
<td>WBI</td>
<td>NI</td>
<td></td>
</tr>
</tbody>
</table>