A New Scoring Algorithm for Multiple-Choice Tests: Conditional Knowledge Model

Alex Strashny [astrashn@uci.edu]
Dept. of Economics, UCI, 3151 Social Science Plaza A, Irvine, CA 92697-5100.
28 August 2002

This paper uses basic rules of probability to develop a new scoring method. The method accounts for guessing, partial knowledge, and misinformation; it also differentiates between incorrect responses and omits. Aside from multiple-choice tests, the method can be used to score short-answer tests. Test scores and confidence intervals are found using simple formulas. Accounting for omits increases test score in almost all cases. Students who guess on questions that they should have omitted are almost always penalized. A counterintuitive finding of this paper is that tests with two answers per question are better able to differentiate between students than tests with higher number of answers per question. In the course of the paper, two new probability density functions are constructed. Their expected values and variances are given.

1. Introduction

This paper develops a new method for scoring multiple-choice tests. Because the method is based on the probability distribution of knowledge conditional on different types of responses, I call this method the *Conditional Knowledge Model* (CKM). CKM accounts for guessing, partial knowledge, misinformation, and omitted responses. Despite its versatility, the model produces simple formulas for estimating test score and its variance.

Salvucci et al (1997) define validity of a score as its lack of bias and reliability as lack of variance. CKM estimates test score as the expected value of score conditional on student's responses. Thus, by definition, test score estimates are valid. CKM also calculates score's variance, from which reliability can be gauged.

Section 2 defines some terminology and sets up the basic model. Sections 3 through 6 solve this model under four sets of assumptions:

1. No partial knowledge, no omits;
2. No partial knowledge, with omits;
3. Partial knowledge, no omits; and finally,
4. Partial knowledge, with omits.

Section 7 examines the effects of accounting for omits versus treating omits as incorrect responses. Section 8 considers the effect of CKM correction on scores of students who do not follow the directions given on the test and guess on questions which they should have omitted. Finally, Section 9 shows that tests with only two answers per question are better able to differentiate between students than tests with a higher number of answers per question.

Appendix I gives sample calculations. In order to just learn how to apply CKM, see Table 1 (if you assume no partial knowledge) or Table 2 (if you assume partial knowledge), and read Section 2 and Appendix I. Appendix II gives some formulas that were too lengthy to put in the main body of the paper. Finally, Appendix III discusses some advantages of CKM over the Three Parameter Item Response Model.

2. The Setup

Let *N* be the total number of questions on a multiple-choice test. N_C, N_I, and N_O are the number of questions answered correctly, incorrectly, and omitted. P_C, P_I, and P_O are the proportions of correct, incorrect, and omitted responses. In conventional practice, test score is set equal to the proportion of correct responses.

Let *n* be the number of answers per question; one of these answers is right, the rest are wrong. R_i is the type of the response that the student gives to question *i*. *R* can be correct (*C*), incorrect (*I*), or omitted (*O*). K_i is the amount of knowledge of question *i* that the student has. Since amount of knowledge is not directly observed, it is a random variable that is conditional on the type of response. *K* is somewhere between negative one and one. When *K* is one, it indicates full knowledge: the student clearly knows that the right answer is right and that the wrong answers are wrong; when *K* is zero, it indicated zero knowledge: the student does not prefer one of the answers to another; when *K* is negative one, there is full misinformation: the student thinks he knows what the right answer is, but that answer is in fact wrong.

The test score *S* is the average amount of knowledge K_i across all the questions. To keep test score between zero and one hundred percent, I rescale this average by dividing by two and adding a half. Let *D* be observable data from a test: $D = \{N, P_C, P_I, P_O\}$. Then test score is

$$S \mid D = \frac{1}{2} + \frac{1}{2N}\sum_{i=1}^{N} K_i \mid R_i \qquad (1).$$

That is, test score conditional on the observed data is the rescaled average amount of knowledge across all questions; the amount of knowledge on a question is conditional on the type of response to that question.

Since the type of response *R* can only take on three different values, the amount of knowledge *K* can have three distinct distributions. Rewrite equation (1):

$$S \mid D = \frac{1}{2} + \frac{1}{2N}\left[\sum K \mid (R = C) + \sum K \mid (R = I) + \sum K \mid (R = O)\right]$$

$$= \frac{1}{2} + \frac{1}{2}\left[P_C \overline{K \mid (R = C)} + P_I \overline{K \mid (R = I)} + P_O \overline{K \mid (R = O)}\right]$$

$$S \mid D = \frac{1}{2} + \frac{1}{2}\sum_r P_r \overline{K \mid (R = r)} \qquad (2).$$

The expected value of the test score is

$$E[S \mid D] = 0.5 + 0.5\sum_r P_r E[K \mid R = r] \qquad (3).$$

That is, the expected value of the test score is the rescaled sum of the expected amount of knowledge conditional on each type of response, times the proportion of questions that

have that response. Using this expected value as the point estimator of test score makes the estimator valid. Expected value is used as point estimator throughout the paper.

Assume that the amount of knowledge on each question is independent of the amount of knowledge on the other questions. This is a reasonable assumption if no two questions cover closely related topics.[1] Under this assumption, the variance of the test score is

$$Var(S \mid D) = (4N)^{-1} \sum_{r} P_r Var(K \mid R = r) \qquad (4).$$

Variance is a measure of reliability. The higher the variance, the lower the reliability. As expected, variance decreases (and reliability increases) as the number of questions on a test increases. This means that by putting more questions on a test, the teacher can determine a student's score more accurately.

If the number of questions on a test is large, then, by Central Limit Theorem, test score S is approximately Normally distribution. Assuming test score's distribution is actually close to the Normal distribution, its 95% confidence interval is approximately the expected value plus or minus two times the standard deviation.

$$CI(S) \approx E[S] \pm 2\sqrt{Var(S)} \qquad (5).$$

The confidence interval in equation (5) is another way to assess reliability. The tighter the confidence interval, the more reliable the test score is. From this, we again see that putting more questions on a test increases reliability.

To find the expected value and variance of test score, we need to know the distributions of amount of knowledge K conditional on different types of responses R. Below I develop these distributions under four different sets of assumptions.

3. No Partial Knowledge, no Omits

Assume that there are no omitted questions. Also assume that there is *no partial knowledge*. That is, for each question, the student is either fully knowledgeable ($K = 1$), has zero knowledge ($K = 0$), or is fully misinformed ($K = -1$).

The probabilities of correct and incorrect response given the three amounts of knowledge are given below.

$$\begin{aligned} &\Pr(R = C \mid K = 1) = 1 && \Pr(R = I \mid K = 1) = 0 \\ &\Pr(R = C \mid K = 0) = \frac{1}{n} && \Pr(R = I \mid K = 0) = \frac{n-1}{n} \\ &\Pr(R = C \mid K = -1) = 0 && \Pr(R = I \mid K = -1) = 1 \end{aligned} \qquad (6).$$

If there is full knowledge, the response is always correct. If the student has zero knowledge, he guesses and gets the right answer with probability of one divided by the number of answers on the question. If the student is fully misinformed, he picks a wrong answer thinking that it's the right one.

Reverse the conditioning in these probabilities using Bayes's Theorem:

$$\Pr(K = k \mid R) = \frac{\Pr(R \mid K = k)\Pr(K = k)}{\sum_{v} \Pr(R \mid K = v)\Pr(K = v)} \qquad (7).$$

The unconditional probability on the right hand side is the prior probability of amount of knowledge *K*. This probability reflects the teacher's beliefs about the student's amount of knowledge *before* the teacher sees that student's test. Assume that the teacher has no prior information about the student's amount of knowledge. In that case, the prior probability is *non-informative*; that is, $\Pr(K = k)$ is the same for all values of *k*. Under this assumption,

$$\Pr(K = 1 \mid R = C) = \frac{n}{n+1} \qquad \Pr(K = 1 \mid R = I) = 0$$
$$\Pr(K = 0 \mid R = C) = \frac{1}{n+1} \qquad \Pr(K = 0 \mid R = I) = \frac{n-1}{2n-1}$$
$$\Pr(K = -1 \mid R = C) = 0 \qquad \Pr(K = -1 \mid R = I) = \frac{n}{2n-1} \qquad (8).$$

For example, suppose a test has four answers per questions and the student gets a questions correct. The probability that he has full knowledge of this question is 80%.

From equation (8), the expected values and variances of amount of knowledge are:

$$E[K \mid R = C] = \frac{n}{n+1} \qquad E[K \mid R = I] = \frac{-n}{2n-1}$$
$$Var(K \mid R = C) = \frac{n}{(n+1)^2} \qquad Var(K \mid R = I) = \frac{n^2 - n}{(2n-1)^2} \qquad (9).$$

As the number of answers per question increases, the power of guessing decreases. A correct response implies full knowledge more and more surely. Thus, the expected value of amount of knowledge conditional on correct response approaches one; the variance approaches zero.

When each question has few answers, zero knowledge can lead to a correct response with a relatively high probability. However, full misinformation always leads to an incorrect response. Thus, expected knowledge given an incorrect response is close to negative one (full misinformation). As the number of answers per question increases, both zero knowledge and full misinformation lead to incorrect response. Since it's impossible to distinguish between the two, expected knowledge approaches the average of zero knowledge and full misinformation, which is negative half. Again, because it's not possible to distinguish between zero knowledge and full misinformation, the variance remains positive and approaches one quarter.

[** Figures 1, 2 **]

Figure 1 shows test scores as a function of the percent of correct responses. It is constructed using equations (3) and (9). A short-answer test can be thought of as a

multiple-choice test with an infinite number of answers per question. Scores for this type of test are also shown in Figure 1.

Figure 2 shows test scores along with two confidence intervals, one if there are thirty questions on the test, and the other if there are one hundred questions on the test. In addition to equations (3) and (9), this figure also uses equations (4) and (5).

4. No Partial Knowledge with Omits

Now, consider the case when some of the responses are omits. In general, the probability of omits can be written as a function of (a) the probability of omits conditional on amount of knowledge and (b) the prior probability of the amount of knowledge:

$$\Pr(R = O) = \sum_k \Pr(R = O \mid K = k)\Pr(K = k) \qquad (10).$$

If the teacher has no prior information about a student's amount of knowledge, then, as before, $\Pr(K = k)$ is non-informative, and is equal for all k. Since amount of knowledge can only take on three values, $\Pr(K = k) = \frac{1}{3}$.

A student will not omit if he has full knowledge or if he is fully misinformed. Thus, on the right hand side, the probability of omit conditional on K being either one or negative one is zero. From this, we see that

$$\Pr(R = O \mid K = 0) = 3\Pr(R = O) \qquad (11).$$

Estimate the unconditional probability of omits with the proportion of questions omitted. Then

$$\Pr(R = O \mid K = 0) \approx 3P_O \qquad (12).$$

When the amount of knowledge is zero, the student omits with probability about $3P_O$. Obviously, the approximation is only valid when the proportion of omits is small.

The probabilities of correct and incorrect responses conditional on the student picking an answer (not omitting) are already given in equation (6). Thus, if we account for omits, equation (6) becomes

$$\begin{array}{lll}
\Pr(R = C \mid K = 1) = 1 & \Pr(R = I \mid K = 1) = 0 & \Pr(R = O \mid K = 1) = 0 \\
\Pr(R = C \mid K = 0) = \frac{1}{n}(1 - 3P_O) & \Pr(R = I \mid K = 0) = \frac{n-1}{n}(1 - 3P_O) & \Pr(R = O \mid K = 0) = 3P_O \\
\Pr(R = C \mid K = -1) = 0 & \Pr(R = I \mid K = -1) = 1 & \Pr(R = O \mid K = -1) = 0
\end{array} \qquad (13).$$

Now use Bayes's Theorem from equation (7) to reverse the conditioning in these probabilities.

$$\Pr(K=1\mid C)=\frac{n}{n+1-3P_O} \qquad \Pr(K=1\mid I)=0 \qquad \Pr(K=1\mid O)=0$$

$$\Pr(K=0\mid C)=\frac{1-3P_O}{n+1-3P_O} \qquad \Pr(K=0\mid I)=\frac{(n-1)(1-3P_O)}{2n-1-3P_O(n-1)} \qquad \Pr(K=0\mid O)=1$$

$$\Pr(K=-1\mid C)=0 \qquad \Pr(K=-1\mid I)=\frac{n}{2n-1-3P_O(n-1)} \qquad \Pr(K=-1\mid O)=0$$

(14).

When there are no omits, that is, when $P_O=0$, equation (14) reduces to equation (8). When there are omits, both (a) the probability of full knowledge conditional on correct response increases; and (b) the probability of full misinformation conditional on incorrect response increases. This is because omits always correspond to zero knowledge. The removal of zero-knowledge questions through omits makes it more likely that correct responses correspond to full knowledge and incorrect responses to full misinformation. For example, suppose a test has four answers per questions. A student omits 10% of the questions. If he gets a question correct, the probability that he has full knowledge of that question is 85%.

Based on equation (14), the expected values and variances are:

$$E[K\mid R=C]=\frac{n}{n+1-3P_O} \qquad Var(K\mid R=C)=\frac{n(1-3P_O)}{(n+1-3P_O)^2}$$

$$E[K\mid R=I]=\frac{-n}{2n-1-3P_O(n-1)} \qquad Var(K\mid R=I)=\frac{n(n-1)(1-3P_O)}{(2n-1-3P_O(n-1))^2}$$

$$E[K\mid R=O]=0 \qquad Var(K\mid R=O)=0 \qquad (15).$$

Though omits do not directly make a contribution to the test score, they affect the score by changing the proportions of correct and incorrect responses, and by changing the weights placed on these proportions. In the presence of omits, the weight placed on correct responses becomes more positive and the weight placed on incorrect responses becomes more negative.

Though it's not readily apparent, in the presence of omits, as the number of answers increases, variance due to correct responses goes to zero faster. The variance due to incorrect responses is lower in the presence of omits than without omits.

[** Figure 3 **]

Figure 3 shows how test score changes with amount of questions omitted. In almost all cases, test score increases when omits are accounted for. This premium for omits is in line with what the literature says should happen (see Kurz 1999).

[** Table 1 **]

Table 1 calculates equations (9) and (15) for various values of answers per question n and proportion of omits. Use it together with equation (3) to calculate test score. Use it together with equations (4) and (5) to find confidence intervals of score. See Appendix I for sample calculations.

5. Partial Knowledge without Omits

Assume now that students can have *partial knowledge*. Partial knowledge is usually defined as the ability to eliminate some, but not all, of the wrong answers (Frary 1980). Extend this definition and define partial misinformation as the ability to eliminate some of the answers, one of which could be the right answer.

The relationship between the amount of knowledge K and the probability of correct response is clarified by the thought experiments below. The first thought experiment is based on Frary's definition and applies only when the amount of knowledge is positive. The second thought experiment, which applies only when the amount of knowledge is negative, is analogous, and allows for the possibility of eliminating the right answer.

Here is the first thought experiment, which applies only when amount of knowledge K is positive. Suppose there are n answers per question. First, the student covers up the right answer. This leaves $n-1$ wrong answers. He then crosses out some of these wrong answers, in proportion to his amount of knowledge. If the amount of knowledge is zero, he is unable to cross out any of the wrong answers. If the amount of knowledge is one, he eliminates all the wrong answers. Thus, he eliminates $K(n-1)$ of the wrong answers; $(1-K)(n-1)$ of the wrong answers are left.[2]

The student then uncovers the right answer. Thus, he sees $1+(1-K)(n-1)$ possible answers. He now chooses one of these answers with equal probability. So the probability of choosing the right answer is one out of $1+(1-K)(n-1)$.

As mentioned before, partial misinformation means that the student might think that the right answer is actually wrong. Use this property to extend the above thought experiment to negative values of the amount of knowledge.

When the amount of knowledge is negative, the student again begins by covering up one of the answers. However, he now covers up a wrong answer. This leaves $n-1$ answers, one of which is right. The student now crosses out some of these answers, in proportion to the absolute value of the amount of knowledge. When the student is fully misinformed, he eliminates all of the $n-1$ answers, including the right one. Thus, the student eliminates $|K|(n-1)$ of the answers. This leaves $(1-|K|)(n-1)$ answers.

After the student finishes eliminating the answers, the probability that the right answer is still available is $1-|K|$. The student now uncovers the wrong answer that he initially covered up. This means that he sees $1+(1-|K|)(n-1)$ answers in all. If the right answer is still available, the probability that the student picks it is one in $1+(1-|K|)(n-1)$.

Based on these two thought experiments, the probabilities or correct and incorrect response are:

When $K \geq 0$:

$$\Pr(R = C \mid K) = \frac{1}{1 + (1 - K)(n - 1)} \qquad \Pr(R = I \mid K) = \frac{(1 - K)(n - 1)}{1 + (1 - K)(n - 1)}.$$

When $K < 0$:

$$\Pr(R = C \mid K) = \frac{1 - |K|}{1 + (1 - |K|)(n - 1)} \qquad \Pr(R = I \mid K) = \frac{1 + (1 - |K|)(n - 2)}{1 + (1 - |K|)(n - 1)} \qquad (16).$$

When amount of knowledge *K* takes on values of one, zero, and negative one, these probabilities reduce to the ones in equation (6). Thus, no partial knowledge is a just a special case of the partial knowledge assumption.

[** Figure 4 **]

Figure 4 shows the probability of correct response as a function of knowledge for different values of *n*. As the number of answers per questions increases, the probability of correct response for $K < 1$ goes to zero.

As before, the prior distribution for amount of knowledge is non-informative. Since amount of knowledge *K* is now a continuous variable, rewrite Bayes's Theorem as

$$p(K \mid R) = \frac{\Pr(R \mid K)}{\int_{-1}^{1} \Pr(R \mid K) dK} \qquad (17).$$

The numerator in (17) is taken right from equation (16). The denominator is just a constant that is needed so that the conditional distribution of knowledge integrates to one.

[** Figures 5, 6 **]

Figures 5 and 6 show probability density of knowledge conditional on correct and incorrect response. As the number of answers per question increases,

- The probability of knowledge conditional on correct response shifts towards one; and
- The distribution of knowledge conditional on incorrect response approaches the uniform distribution.

Based on (17), the expected values of amount of knowledge are:

$$E[K \mid R = C] = \frac{(2 \ln n - 3)n^2 + 4n - 1}{2(n - 1)\left[(n - 2)\ln n + n - 1\right]}$$

$$E[K \mid R = I] = \frac{(2\ln n - 3)n^2 + 4n - 1}{2(n-1)\left[-2n^2 + 5n - 3 + (n-2)\ln n\right]} \qquad (18).$$

As the number of answers per question increases, expected knowledge given a correct response approaches one. In the limit, guessing correctly is impossible and only full knowledge allows the student to respond correctly. Expected knowledge given incorrect response approaches zero. Again, in the limit, only full knowledge allows for a correct response. All other amounts of knowledge, from full misinformation to almost full knowledge, result in an incorrect response. Since distribution of knowledge approaches the uniform from negative one to one, the expected value approaches zero.

Because of their length, the expressions for variance are given in the Appendix II. As the number of answers per question increases, variance of knowledge given correct response increases at first. However, once the number of answers is large enough, the variance starts approaching zero. When there are only two answers per question, there are two ways to get the correct response: the student either knows the right answer, or he guesses it. As the number of answers per question increases, the number of ways to guess correctly also increases: the student can eliminate one answer and guess correctly, or he can eliminate two answers and guess correctly, etc. However, each of these ways of guessing correctly implies a different amount of knowledge. Thus, the variance of knowledge increases. However, once there are enough answers per question, it becomes harder and harder to guess correctly. Correct response starts implying full knowledge more and more surely. Thus, variance of knowledge starts approaching zero.

As the number of answers per question increases, variance of knowledge given incorrect response approaches one third. This is because the distribution of knowledge approaches the uniform distribution from negative one to one.

[** Figures 7, 8 **]

Figure 7 shows test scores as a function of correct responses. Figure 8 shows test scores along with two confidence intervals.

6. Partial Knowledge with Omits

Now, allow for omits. Since amount of knowledge *K* is now a continuous random variable, rewrite equation (10) as

$$\Pr(R = O) = \int_{-1}^{1} \Pr(R = O \mid K)\, p(K)\, dK \qquad (19).$$

Maintain the non-informative prior for amount of knowledge. This means that $p(K) = 0.5$. Assume that, for some constant *a*, the student always omits if amount of knowledge is between *-a* and *a*, and never omits otherwise. This means that the unconditional probability of omits is *a*. As before, estimate the unconditional probability of omits with the proportion of the questions omitted. Thus,

$\Pr(R = O \mid K) = 1$ when $K \in (-a, a)$, and 0 otherwise

$\Pr(R = O \mid K) \approx I(-P_O, P_O)$ (20).

From equation (17), the distribution of amount of knowledge given an omit is

$$p(K \mid R = O) = \frac{I(-P_O, P_O)}{2P_O} = U(-P_O, P_O) \quad (21).$$

The expected value and variance are therefore

$$E[K \mid R = O] = 0 \qquad Var(K \mid R = O) = \frac{P_O^2}{3} \quad (22).$$

Conditional probabilities of correct and incorrect response remain as shown in equation (16), except that now the equation applies only for knowledge greater than a or less than $-a$.

[** Figures 9, 10 **]

In the presence of omits, the ordinate of the distribution of knowledge is higher because the denominator in equation (17) is now lower. See Figures 9 and 10.

Because of their length, the expected values conditional on correct and incorrect response are given in Appendix II; the variances are too long to give even in the appendix. In the presence of omits, expected value of knowledge given correct response approaches one slightly faster. Expected value of knowledge given incorrect response approaches zero slower. That's because incorrect response in the presence of omits points to misinformation with a higher probability.

When there are omits, variance given both correct and incorrect response is higher. That is because there is now a break in the probability distribution of knowledge between $-a$ and a.

[** Figure 11 **]

Figure 11 shows how test score changes with amount of questions omitted. As when no partial knowledge is assumed, accounting for omits increases test score.

[** Table 2 **]

Table 2 gives expected values and variances of knowledge for various values of answers per question n and proportion of omits. Use it together with equation (3) to calculate student's score. Use it together with equations (4) and (5) to find confidence intervals of score. See Appendix I for sample calculations.

7. Effect of accounting for omits

In the education literature, it has long been maintained that omits should add to test score more than incorrect responses do. For example, this is the basis for the formula scoring correction (Kurz 1999).

CKM is not designed with the explicit goal of benefiting omits as opposed to incorrect responses. However, as it turns out, omits do increase test score in virtually all cases (see Figures 3 and 11). Only when the proportion of correct responses is extremely low, does accounting for omits decrease test score very slightly.

The effect of omits is that they make expected knowledge given correct response more positive and expected knowledge given incorrect response more negative. However, expected knowledge given an omit is always greater than expected knowledge given incorrect response. That's why differentiating between omits and incorrect responses generally bumps up the score.

Knowledge conditional on an omit has lower variance than knowledge conditional on an incorrect response. However, in the presence of omits, the variance of knowledge conditional on both correct and incorrect responses increases. Because of this, the overall effect of omits on the variance of test score is unclear. In the example given in Appendix I, variance of score in the presence of omits decreases only very slightly.

8. Effect on High Risk Takers

On many multiple-choice tests, the directions tell students that if their knowledge of a question is close to zero, they should omit it. This is done to increase the reliability of the test. These tests are often graded using formula scoring, which either penalized incorrect responses or awards omits.

If a test is graded using formula scoring, the *expected* score of a student who does not follow test directions is the same as of the student who follows these directions. Thus, on average, a student is neither awarded nor penalized for ignoring the directions. However, formula scoring has been criticized because not following the directions increases a student's score half the time (Kurz 1999, Angoff 1989).

Let us see what effects CKM has on the scores of students who do not follow these directions. Suppose there are two students who are identical in every way except that L is a low risk taker and follows the directions given on the test, while H is a high risk taker who does not follow these directions. The proportion of questions on which these students have zero knowledge is P_Z. L answers all the questions except for the ones on which he has zero knowledge. So for L, $P_O = P_Z$. L's proportion of correct and incorrect responses are P_C and P_I. H, on the other hand, guesses on the P_Z questions.

[**Reworking this section. What follows may not be 100% correct.**]

Consider how components of equation (3) change for H relative to L. For H, both the proportion of correct responses and proportion of incorrect responses increase. The

expected value of this increase is $\frac{1}{n}P_Z$ for proportion of correct responses and $\frac{n-1}{n}P_Z$ for proportion of incorrect responses. However, the weights placed on these proportions change towards zero (see Table 1 or Table 2).

Let w_r be the weights placed on proportion of correct and incorrect responses for student L. That is,

$$w_r = E[K \mid R = r, P_O = P_Z] \qquad (22).$$

Let Δw_r be the difference in weights used for student H versus student L. That is,

$$\Delta w_r = E[K \mid R = r, P_O = 0] - E[K \mid R = r, P_O = P_Z] \qquad (23).$$

Then, the expected difference in H's score minus L's score is

$$E[\Delta S] = E[S_H - S_L] = 0.5\left[P_C \Delta w_C + P_I \Delta w_I + P_Z \left\{\frac{1}{n}(w_C + \Delta w_C) + \frac{n-1}{n}(w_I + \Delta w_I)\right\}\right] \quad (24).$$

[** Figure 12 **]

Figure 12 plots this expected score difference assuming that there is partial knowledge, for four answers per question. In almost all cases, the high risk taker, that is, the student who does not follow the directions, is penalized. This student only benefits when he makes very few correct responses.

9. Ability of tests to differentiate between students

One of the purposes of a test is to differentiate between different students. From this perspective, on an ideal test, for every unit change in the proportion of correct responses, the valid estimate of test score increases by one unit as well. Thus, ideally, the graph of expected score versus proportion of correct responses is a forty-five degree line.

On the other hand, if a test is unable to differentiate between students at all, then the valid estimate of test score does not change, no matter how much the proportion of correct responses changes. In this worst-case scenario, the graph of expected score versus proportion of correct responses is a horizontal line.

In this vein, define test quality as the derivative of expected value of score with respect to proportion of correct responses:

$$Q = \frac{dE[S]}{dP_C} \qquad (25).$$

In the ideal situation, this test quality Q is one[3]; if expected value of score never changes, test quality is zero. Test quality can be gauged by eye from Figures 1 and 7. It is also given in Table 3 for various values of answers per question n.

[** Table 3 **]

Intuitively, if there are more answers per question, then the test is better in some sense. This intuition is supported under the assumption of no partial knowledge. When there are two answers per question, test quality is 67%; it steadily approaches 75% as the number of answers per question increases.

The rationale for this result is simple. The more answers per question, the harder it is to guess correctly; thus, a correct response corresponds to full knowledge more and more surely. Because of this, expected value of score changes with proportion of correct responses almost one for one.

However, assuming that partial knowledge exists, test quality does not change in such a straightforward manner. When there are two answers per question, test quality is 27%; it approaches 50% as number of answers per question approaches infinity. However, before increasing, test quality actually *decreases*. Quality is at a minimum when there are six answers per question. It starts increasing slowly after that. But even when there are ten answers per question, test quality is still less than when there are two, and even three, answers per question. Thus, for practical purposes, multiple-choice tests with two answers per question are most able to differentiate between students. Only short-answer tests, which can be thought of as multiple-choice tests with an infinite number of answers per question, do a better job at this.

The logic that it is harder to guess when there are more answers per question still applies. That is why eventually test quality increases to above what it is when there are two answers per question. The reason that test quality decreases at first is that expected knowledge given correct response increases slower than expected knowledge given incorrect response. Correct response can mean that there is full knowledge, *or* that knowledge is less than full and the student simply guesses correctly. Thus, expected knowledge stays well below one even for large numbers of answers per question (see Table 2). On the other hand, when there are even a few answers per question, incorrect response can reasonably mean that knowledge is somewhere between full misinformation to almost full knowledge. Thus, expected knowledge is very close to zero.

10. Conclusion

This paper develops a new valid scoring algorithm for multiple-choice tests called CKM. Though the method is easy to apply, it is derived from solid probability foundations. The method allows finding the test score based on the proportion of responses that are correct, incorrect, and omitted. It can also quantify the reliability of a test.

Accounting for omits generally increases test score, which is in line with education literature. Also, unlike in formula scoring, students who do not follow directions and guess on questions that they should have omitted are generally penalized.

A surprising finding of this paper is that multiple-choice tests with two answers per question are better able to differentiate between students than tests with higher number of answers per question.

Acknowledgments
Thanks to Rich Brown for helpful input.

Endnotes

[1] This assumption is only needed for finding the variance and confidence interval of test score. Even if this assumption does not hold, the calculation for expected value of test score still holds.

[2] Throughout, amount of knowledge K is treated as a continuous variable. To see why this should be so, consider that when the student eliminates an answer, he could know that (a) just one element of the answer is false; or (b) everything in the answer is false; or (c) everything in the answer is false, plus, other information, that is not part of the answer, is also false. Each of these three amounts of knowledge leads to the elimination of the same single answer. Thus, even though the number of eliminated answers is discrete, amount of knowledge is fully continuous.

[3] This assumes that test score can only take on values between zero and one.

Bibliography

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement*, 26. 323-336.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, 4. 79-90.

Kutz, T. B. (1999). A review of scoring algorithms for multiple-choice tests. *Southwest Educational Research Association*.

Salvucci, S.; Walter, E.; Conley, V.; Fink, S.; & Saba, M. (1997). *Measurement error studies at the National Center for Education Statistics*. Washington D.C.: U.S. Department of Education.

Table 1
Expected Value and Variance of Knowledge given Correct and Incorrect Responses
Assumes that there is no partial knowledge

n	No Omits				Omit 5%			
	Expected		Variance		Expected		Variance	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
2	0.667	-0.667	0.222	0.222	0.702	-0.702	0.209	0.209
3	0.750	-0.600	0.188	0.240	0.779	-0.638	0.172	0.231
4	0.800	-0.571	0.160	0.245	0.825	-0.611	0.145	0.238
5	0.833	-0.556	0.139	0.247	0.855	-0.595	0.124	0.241
6	0.857	-0.546	0.122	0.248	0.876	-0.585	0.109	0.243
7	0.875	-0.539	0.109	0.249	0.892	-0.579	0.097	0.244
8	0.889	-0.533	0.099	0.249	0.904	-0.574	0.087	0.245
9	0.900	-0.529	0.090	0.249	0.914	-0.570	0.079	0.245
10	0.909	-0.526	0.083	0.249	0.922	-0.567	0.072	0.246
Inf	1.000	-0.500	0.000	0.250	1.000	-0.541	0.000	0.248

n	Omit 10%				Omit 20%			
	Expected		Variance		Expected		Variance	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
2	0.741	-0.741	0.192	0.192	0.833	-0.833	0.139	0.139
3	0.811	-0.682	0.153	0.217	0.882	-0.790	0.104	0.166
4	0.851	-0.656	0.127	0.226	0.909	-0.769	0.083	0.178
5	0.877	-0.641	0.108	0.230	0.926	-0.758	0.069	0.184
6	0.896	-0.632	0.094	0.233	0.938	-0.750	0.059	0.188
7	0.909	-0.625	0.083	0.234	0.946	-0.745	0.051	0.190
8	0.920	-0.620	0.074	0.236	0.952	-0.741	0.045	0.192
9	0.928	-0.616	0.067	0.236	0.957	-0.738	0.041	0.194
10	0.935	-0.614	0.061	0.237	0.962	-0.735	0.037	0.195
Inf	1.000	-0.588	0.000	0.242	1.000	-0.714	0.000	0.204

Table 2
Expected Value and Variance of Knowledge given Correct and Incorrect Responses
Assumes that there is partial knowledge

n	No Omits				Omit 5%			
	Expected		Variance		Expected		Variance	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
2	0.273	-0.273	0.259	0.259	0.287	-0.287	0.269	0.269
3	0.305	-0.193	0.275	0.275	0.318	-0.204	0.283	0.288
4	0.328	-0.155	0.284	0.282	0.341	-0.164	0.291	0.297
5	0.347	-0.132	0.291	0.286	0.360	-0.140	0.297	0.302
6	0.362	-0.116	0.296	0.290	0.375	-0.123	0.301	0.306
7	0.375	-0.105	0.299	0.293	0.387	-0.111	0.304	0.309
8	0.386	-0.096	0.302	0.295	0.398	-0.101	0.307	0.312
9	0.396	-0.088	0.304	0.297	0.408	-0.094	0.308	0.314
10	0.404	-0.082	0.305	0.299	0.417	-0.087	0.310	0.316
Inf	1.000	0.000	0.000	0.333	1.000	0.000	0.000	0.351

n	Omit 10%				Omit 20%			
	Expected		Variance		Expected		Variance	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
2	0.303	-0.303	0.278	0.278	0.339	-0.339	0.299	0.299
3	0.333	-0.216	0.291	0.303	0.368	-0.245	0.307	0.334
4	0.356	-0.174	0.299	0.313	0.389	-0.198	0.313	0.349
5	0.374	-0.148	0.304	0.319	0.406	-0.169	0.316	0.357
6	0.389	-0.131	0.307	0.323	0.420	-0.148	0.318	0.363
7	0.401	-0.118	0.310	0.327	0.432	-0.134	0.319	0.367
8	0.412	-0.107	0.311	0.330	0.443	-0.122	0.320	0.370
9	0.422	-0.099	0.313	0.332	0.452	-0.113	0.321	0.373
10	0.430	-0.092	0.314	0.334	0.460	-0.105	0.321	0.375
Inf	1.000	0.000	0.000	0.370	1.000	0.000	0.000	0.413

Table 3
Test Quality

n	No Partial Knowledge	Partial Knowledge
2	66.7%	27.3%
3	67.5%	24.9%
4	68.6%	24.1%
5	69.4%	23.9%
6	70.1%	23.9%
7	70.7%	24.0%
8	71.1%	24.1%
9	71.5%	24.2%
10	71.8%	24.3%
Inf	75.0%	50.0%

Table 4
Test Score

Assuming that there is no partial knowledge (using Table 1)

	80% Correct / 20% Incorrect					80% Correct / 10% Incorrect / 10% Omitted				
n	Score	CI (N = 30)		CI (N = 100)		Score	CI (N = 30)		CI (N = 100)	
2	70.0	61.4	78.6	65.3	74.7	75.9	68.3	83.5	71.8	80.1
4	76.3	68.6	84.0	72.1	80.5	80.8	74.3	87.2	77.2	84.3
Inf	85.0	80.9	89.1	82.8	87.2	87.1	84.2	89.9	85.5	88.6

Assuming that there is partial knowledge (using Table 2)

	80% Correct / 20% Incorrect					80% Correct / 10% Incorrect / 10% Omitted				
n	Score	CI (N = 30)		CI (N = 100)		Score	CI (N = 30)		CI (N = 100)	
2	58.2	48.9	67.5	53.1	63.3	60.6	51.4	69.7	55.6	65.6
4	61.6	51.8	71.3	56.2	66.9	63.4	53.9	72.9	58.2	68.6
Inf	90.0	85.3	94.7	87.4	92.6	90.0	86.5	93.5	88.1	91.9

Figure 1: Estimate of test score. No partial knowledge, no omits.

Estimate of test score as a function of percent of correct responses. Assumes that there is no partial knowledge, and does not account for omits. Given for two, four, and infinite number of answers per question.

Figure 2: Confidence interval of test score. No partial knowledge, no omits.

Estimate of test score with four answers per question, along with two confidence intervals. The inner confidence interval is for a test with one hundred questions; the outer confidence interval is for a test with thirty questions.

Figure 3: Accounting for omits. No partial knowledge.

Estimate of test score with four answers per question. Given if zero percent, ten percent, and twenty percent of the questions are omitted. In the overwhelming majority of cases, omits increase test score.

Figure 4: Probability of correct response, accounting for partial knowledge.

Probability of correct response conditional on amount of knowledge. Given for two, four, and five answers per question.

Figure 5: PDF of knowledge given correct response.

Probability density of amount of knowledge conditional on correct response. Given for two, four, and five answers per question.

Figure 6: PDF of knowledge given incorrect response.

Probability density of amount of knowledge conditional on incorrect response. Given for two, four, and five answers per question.

Figure 7: Estimate of test score. Partial knowledge, no omits.

Estimate of test score. Assumes that there is partial knowledge, but does not account for omits. Given for two, four, ten, and infinite number of answers per question.

Figure 8: Confidence interval of test score. Partial knowledge, no omits.

Estimate of test score with four answers per question, along with two confidence intervals. The inner confidence interval is for a test with one hundred questions; the outer confidence interval is for a test with thirty questions.

Figure 9: PDF of knowledge given correct response, accounting for omits.

Probability density of amount of knowledge conditional on correct response with four answers per question. Given if zero percent, ten percent, and twenty percent of questions are omitted.

Figure 10: PDF of knowledge given incorrect response, accounting for omits.

Probability density of amount of knowledge conditional on incorrect response with four answers per question. Given if zero percent, ten percent, and twenty percent of questions are omitted.

Figure 11: Accounting for omits. Partial knowledge.

Estimate of test score with four answers per question. Given if zero percent, ten percent, and twenty percent of the questions are omitted.

Figure 12: Expected difference in score between high and low risk takers.

Expected difference in score for four answers per question. Given for ten percent and twenty percent of questions on which students have zero knowledge. Low risk takers omit these questions while high risk takers guess on them.

Figure 1: Score: No partial knowledge, no omits

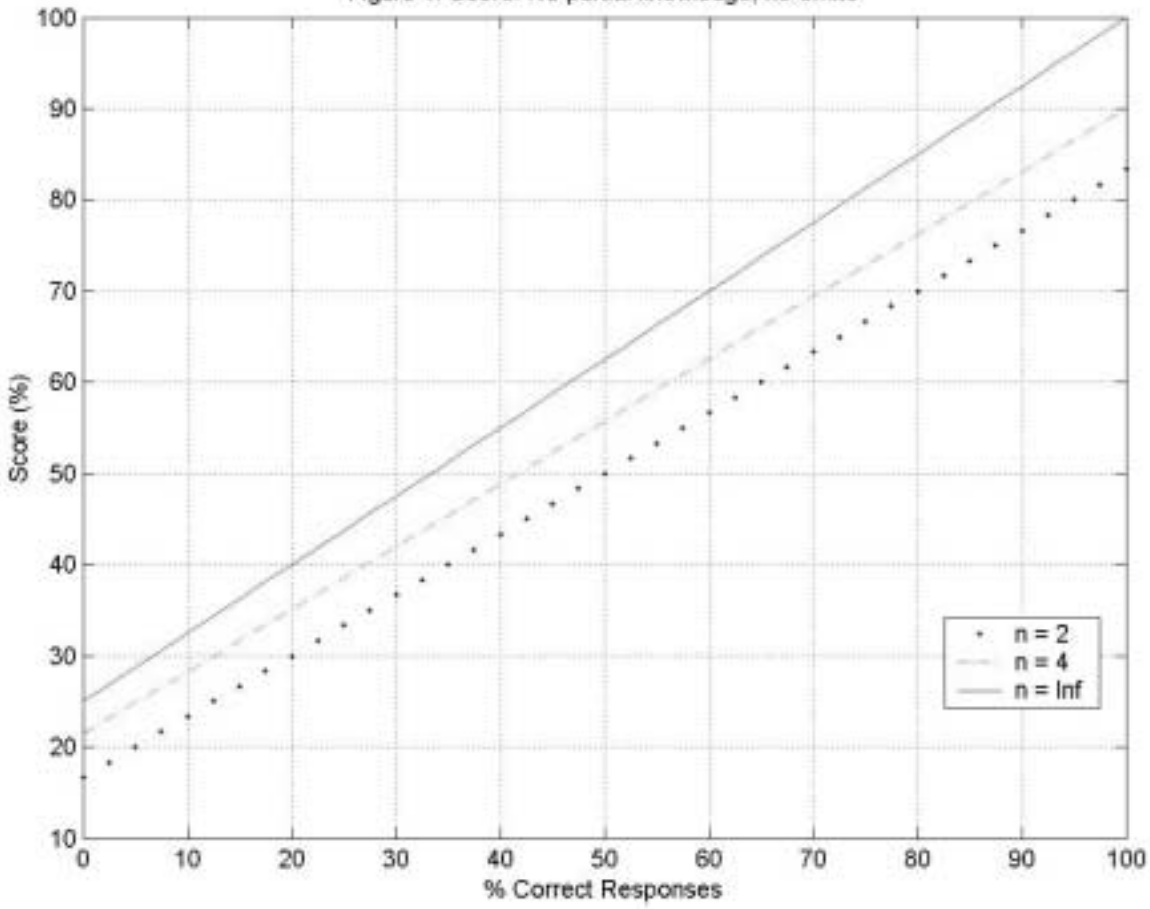


Figure 2: Score with Confidence Intervals: No partial knowledge, no omits, n = 4

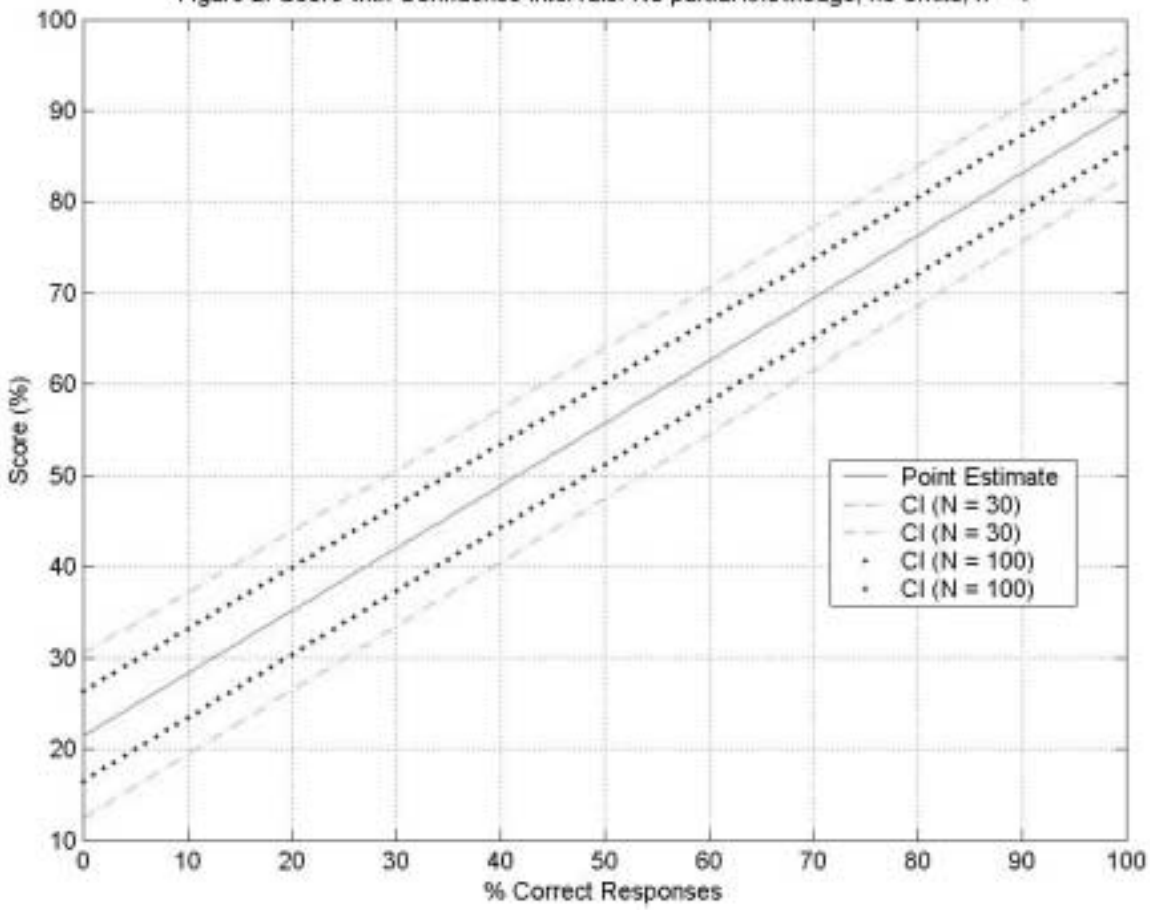


Figure 3: Effect of omits: No partial knowledge, n = 4

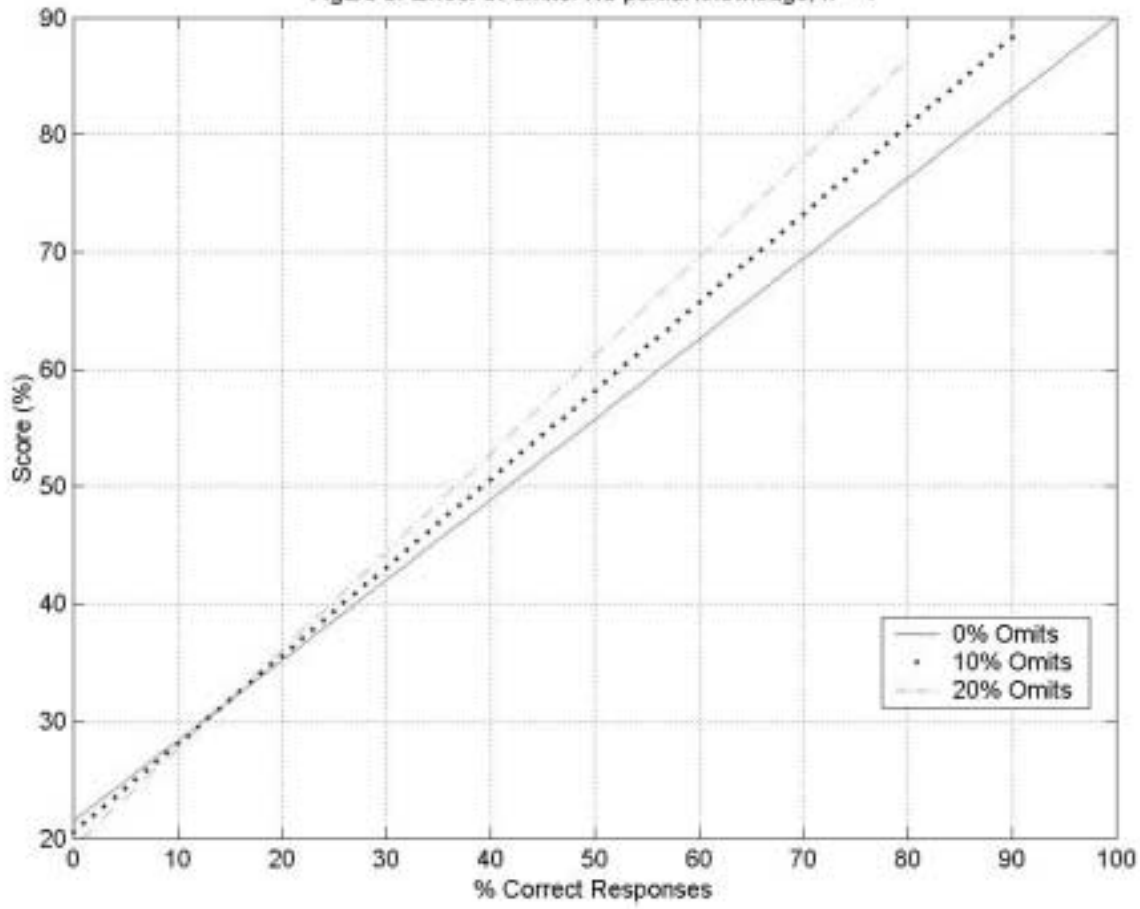


Figure 4: Probability of correct response conditional on amount of knowledge

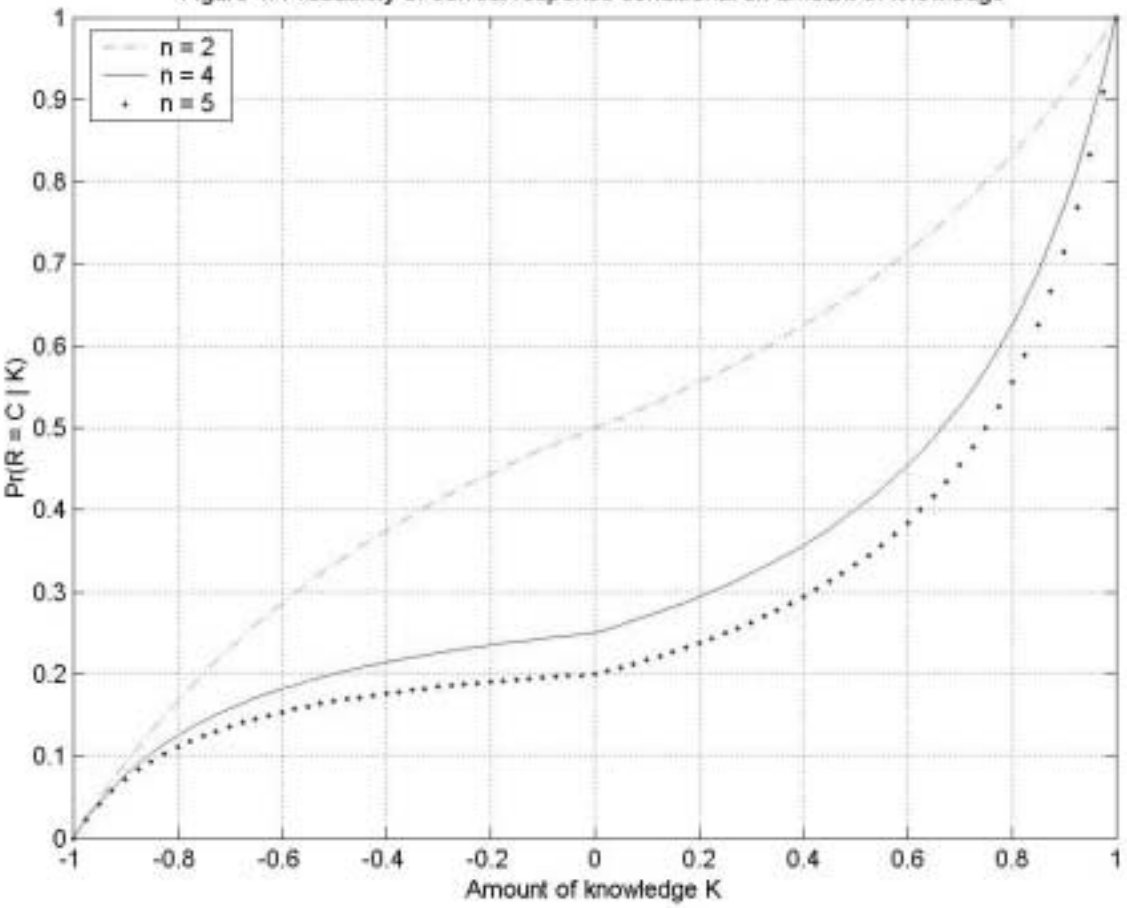


Figure 5: PDF of amount of knowledge conditional on correct response

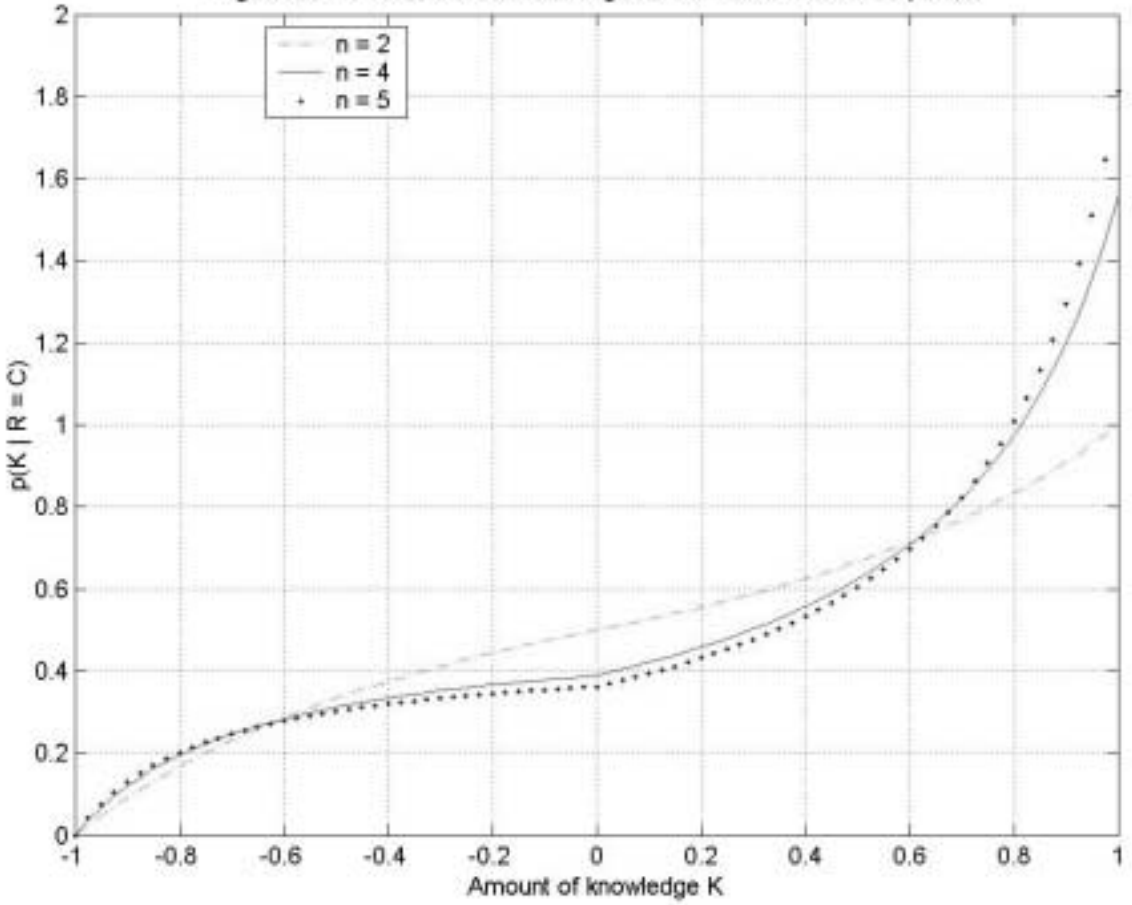


Figure 6: PDF of amount of knowledge conditional on incorrect response

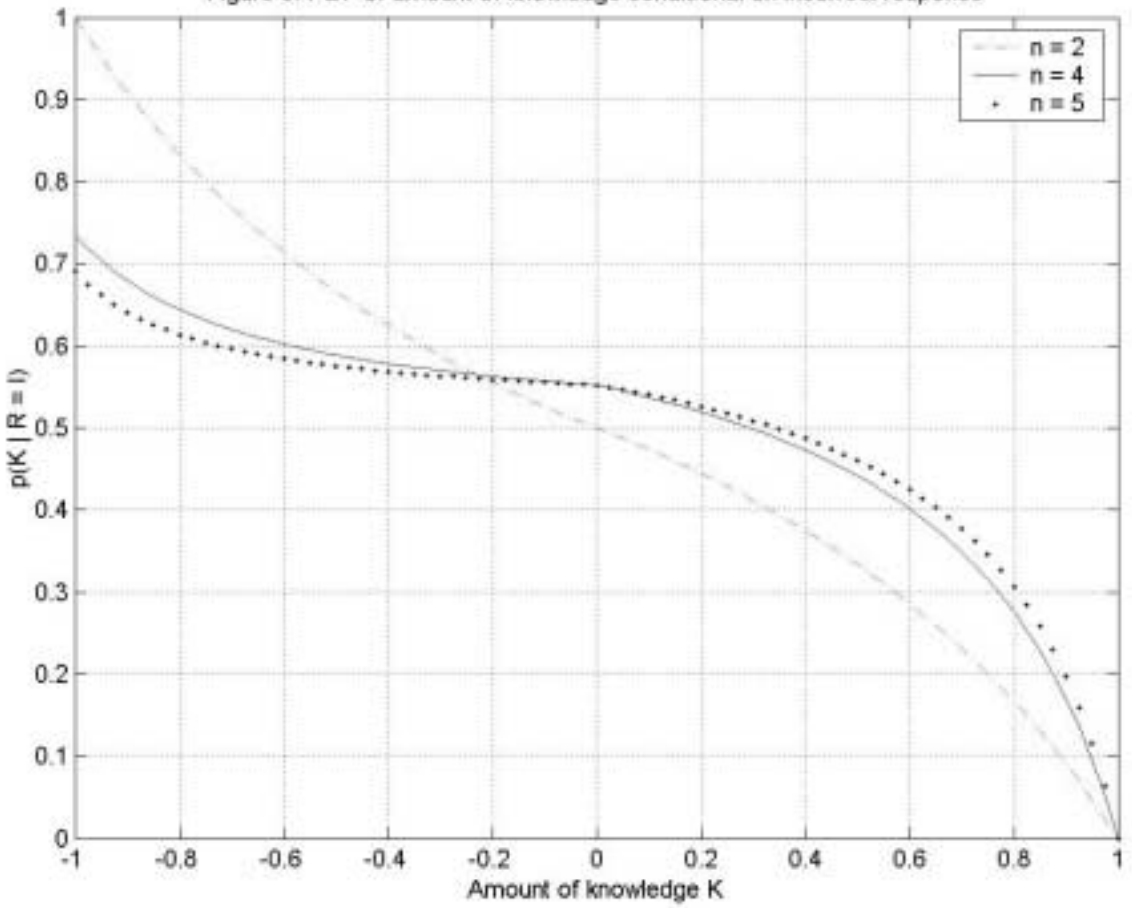


Figure 7: Score: Partial knowledge, no omits

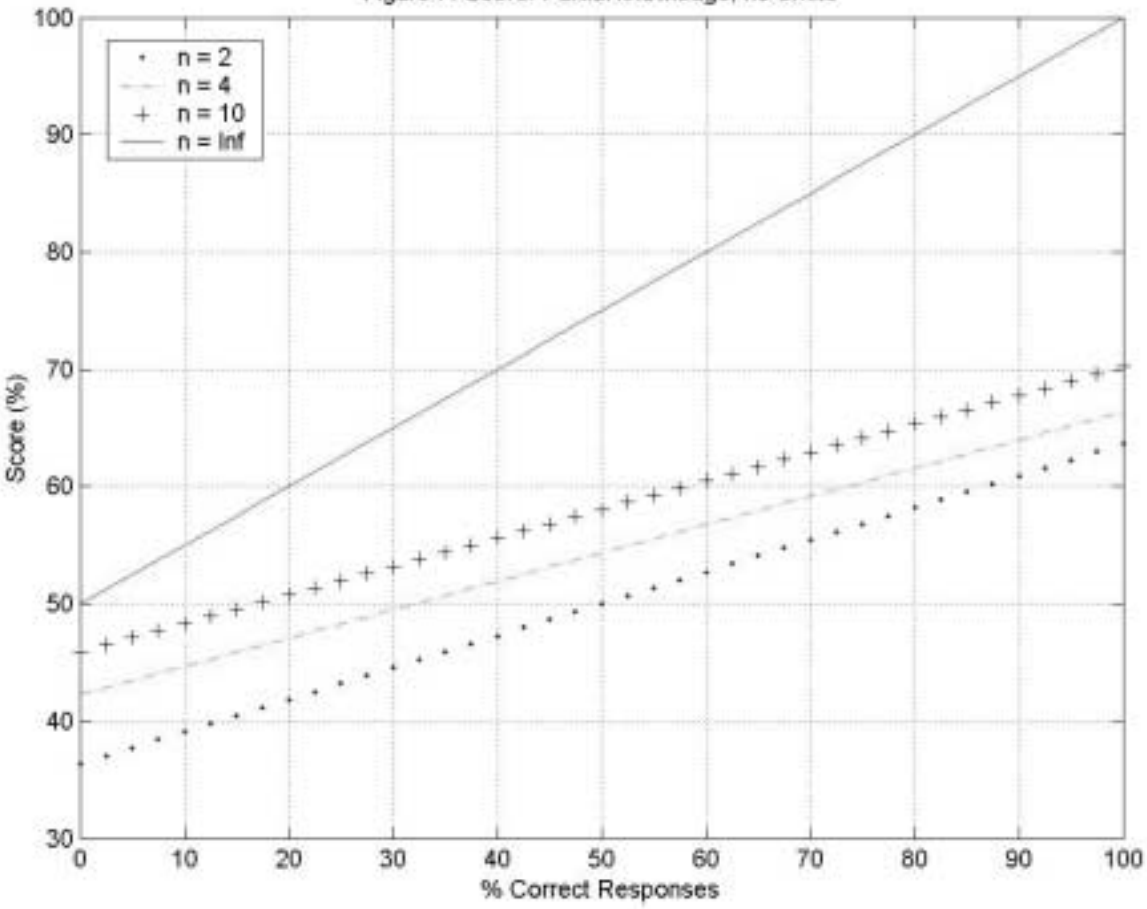


Figure 8: Score with Confidence Intervals: Partial knowledge, no omits, n = 4

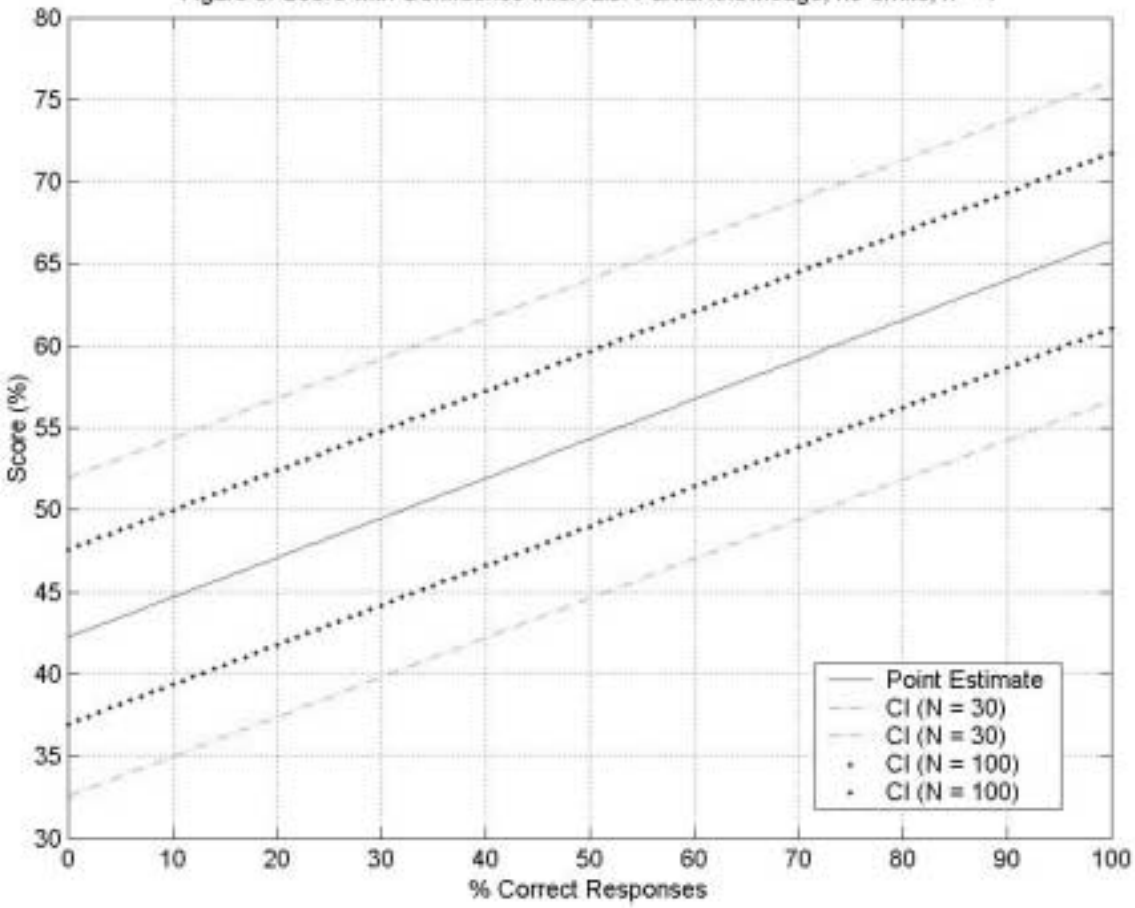


Figure 9: PDF of knowledge conditional on correct response for n = 4

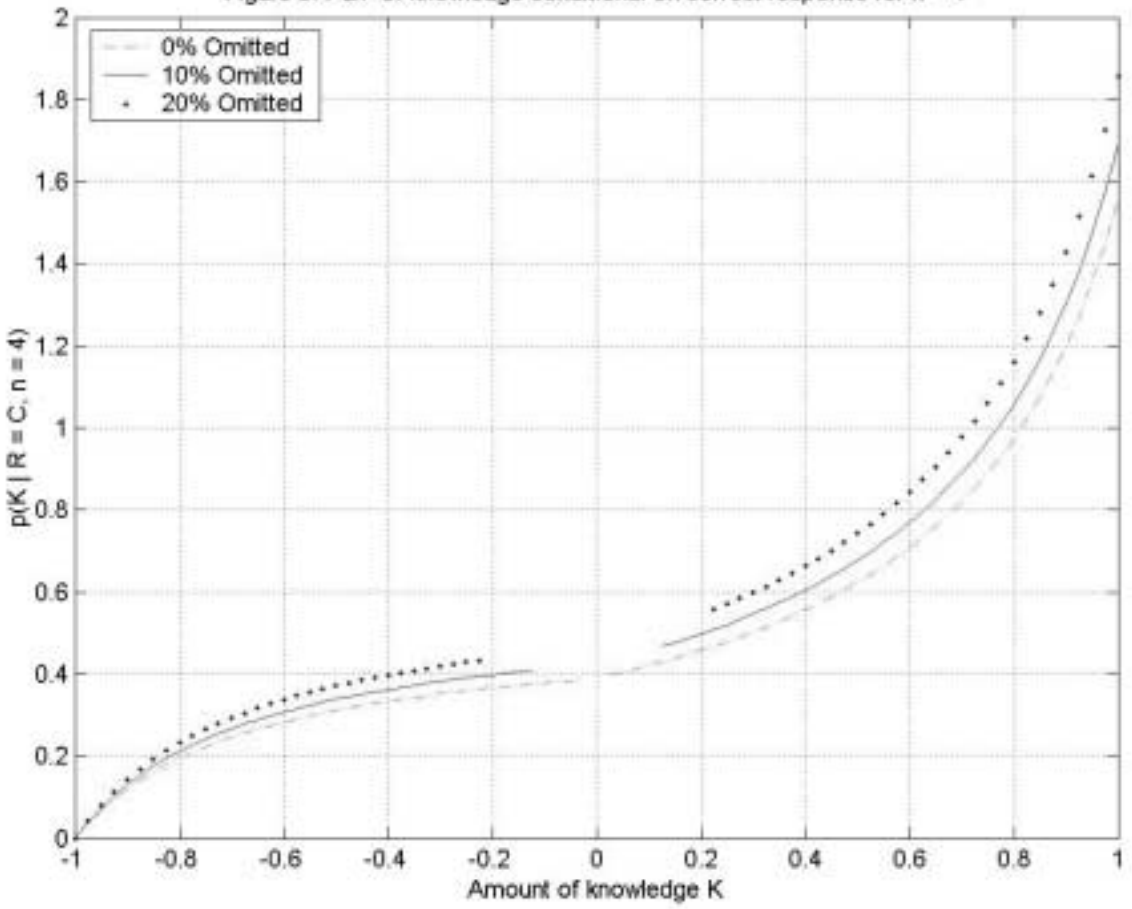


Figure 10: PDF of knowledge conditional on correct response for n = 4

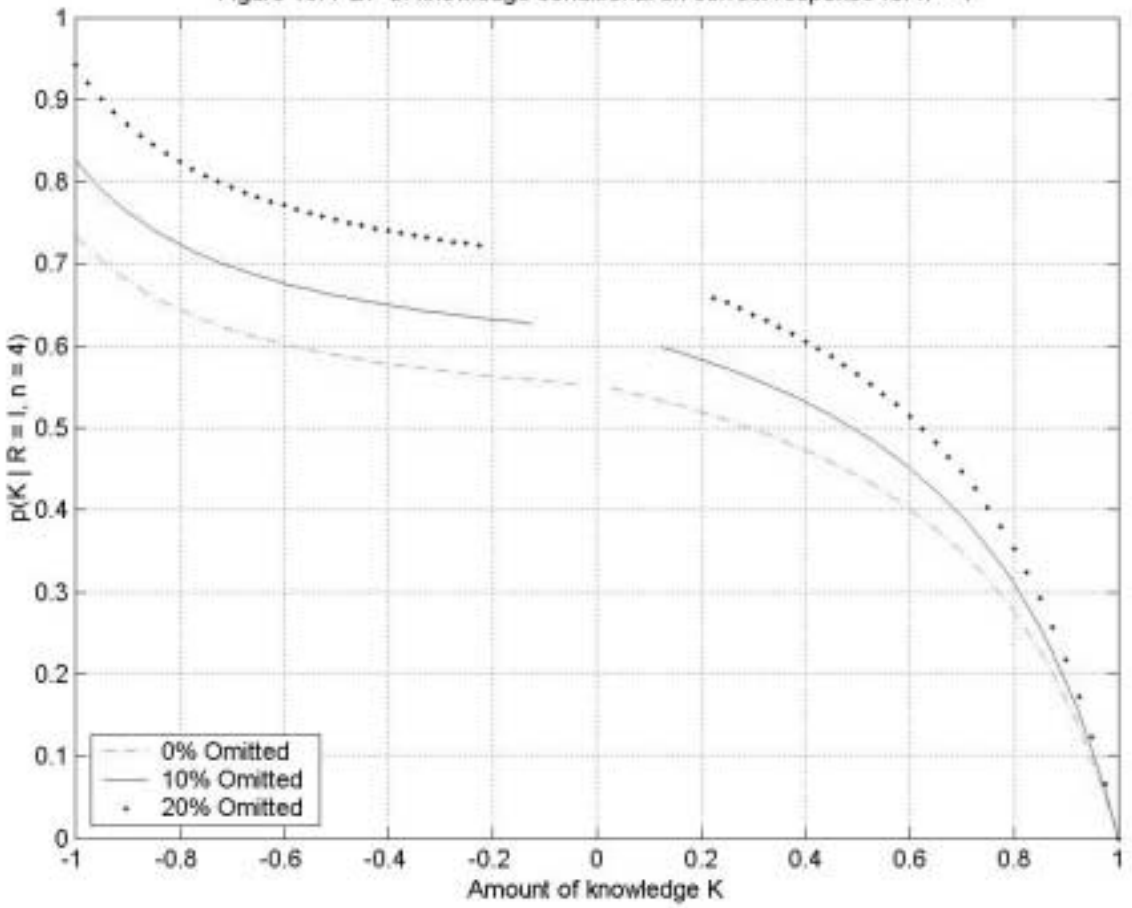


Figure 11: Effect of omits: Partial knowledge, n = 4

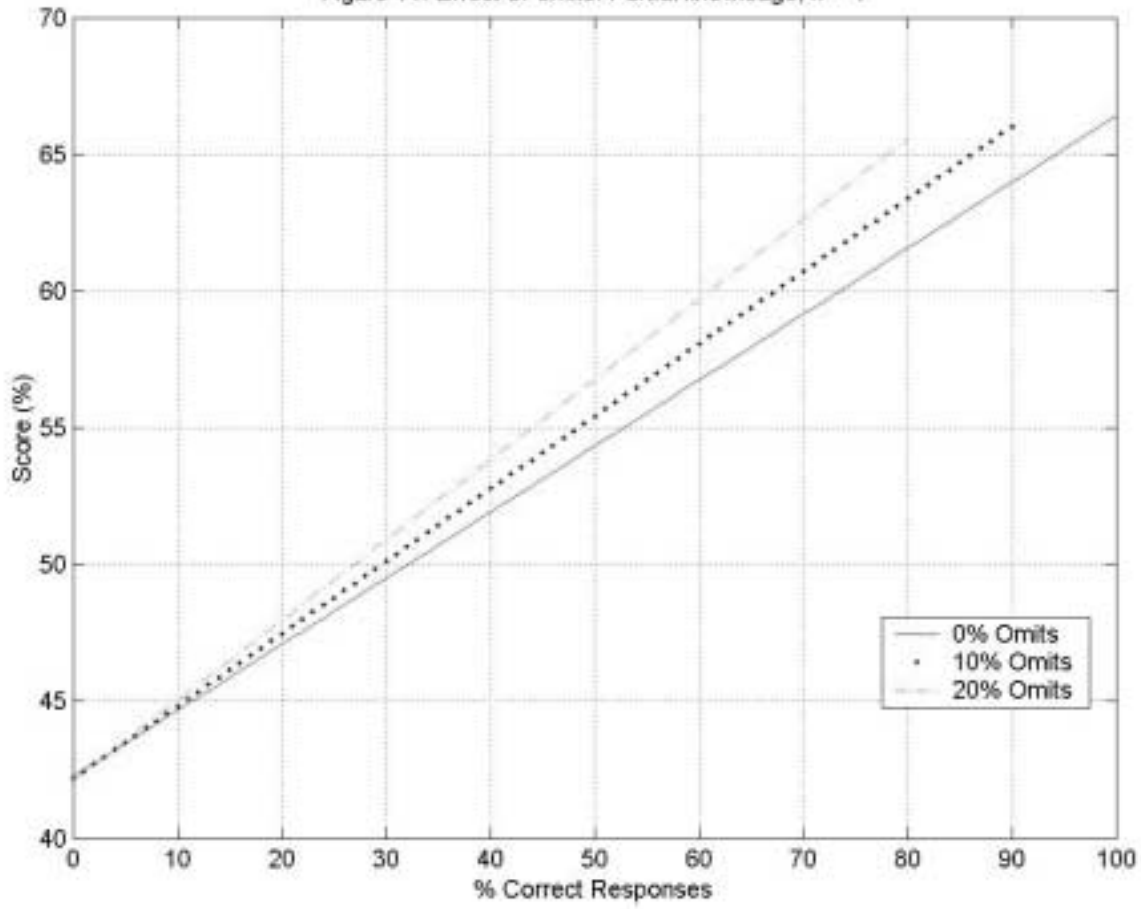
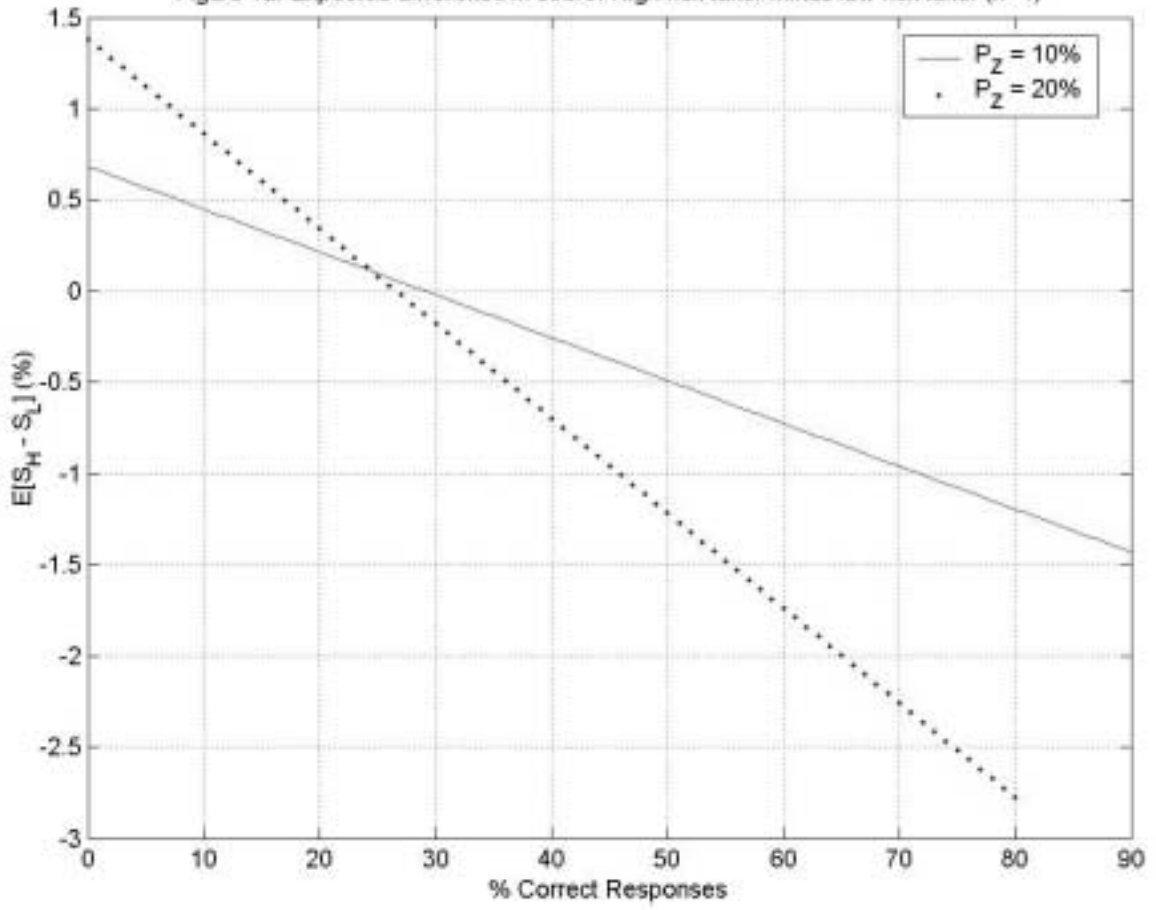


Figure 12: Expected difference in score: High risk taker minus low risk taker (n=4)



Appendix I

To just estimate a student's score under the CKM correction, use equation (3) along with Tables 1 or 2. For the assumption of partial knowledge, use Table 2.

Suppose a test has four answers per question. A student gets 80% of responses correct and 20% incorrect. Using equation (3) and Table 2,

$$E[S] = 0.5 + 0.5 * \left[(0.8)(0.328) + (0.2)(-0.155)\right] = 0.616 = 61.6\% .$$

To estimate the confidence interval for this score, also use equations (4) and (5). Suppose there are thirty questions on the test. Then the score's variance is

$$Var(S) = \frac{(0.8)(0.284) + (0.2)(0.282)}{4 * 30} = 0.0024 .$$

From this, the confidence interval of the student's score is

$$CI(S) = 0.616 \pm 2\sqrt{0.0024} = 0.518 - 0.713 .$$

That is, the score's confidence interval is between 51.8% and 71.3%.

Now, suppose that the student gets 80% correct, 10% incorrect, and omits 10% of the questions. Then the estimate of his score is

$$E[S] = 0.5 + 0.5 * \left[(0.8)(0.356) + (0.1)(-0.174) + (0.1)(0)\right] = 0.634 = 63.4\% .$$

For calculating the variance, also see equation (22):

$$Var(S) = \frac{(0.8)(0.299) + (0.1)(0.313) + (0.1)(0.1)^2(0.333)}{4 * 30} = 0.0023 .$$

Finally, the confidence interval is

$$CI(S) = 0.634 \pm 2\sqrt{0.0023} = 0.539 - 0.729 .$$

So in this case, the confidence interval is between 53.9% and 72.9%.

[** Table 4 **]

Table 4 summarizes results of calculations for this student for different values of answers per question, number of questions on the test, and under the assumptions of no partial knowledge (using Table 1) and partial knowledge (using Table 2).

Appendix II

This appendix lists lengthy formulas that were not given in the body of the paper.

Partial Knowledge without Omits

Variance of knowledge given correct response:

$$Var(K \mid R = C) = \frac{(34\ln n - 41)n^3 + (93 - 48(\ln n)^2 + 26\ln n)n^2 + (-76\ln n - 63)n + 11 + 16\ln n}{12\left[(n-1)\left((\ln n + 1)n - 2\ln n - 1\right)^2\right]}$$

Variance of knowledge given incorrect response:

$$T = 16n^5 + (-60 - 32\ln n)n^4 + (27 + 146\ln n)n^3 + (65 - 48(\ln n)^2 - 94\ln n)n^2 + (-51 - 20\ln n)n + 3$$

$$Var(K \mid R = I) = \frac{T}{12\left[\left(2n^2 + (-5 - \ln n)n + 2\ln n + 3\right)^2 (n-1)\right]}$$

Partial Knowledge with Omits

For all of the following, let
$v = \frac{1}{a} - 1 \approx \frac{1}{P_O} - 1.$

Expected value of knowledge given correct response:

Numerator:

$T_2 = 4n - 1 + \left(-3 - 2\ln(v+1) + 2\ln(1+nv)\right)n^2$
$T_1 = \left((-4 + 4\ln(1+nv) - 4\log(v+1)\right)n^2 - 2 + 6n$
$T_0 = \left(-2\ln(v+1) + 2\ln(1+nv)\right)n^2$
$N = T_2 v^2 + T_1 v + T_0$

Denominator:

$T_1 = -2\ln(1+nv) + 2\ln(v+1) + \left(\ln(1+nv) - \ln(v+1) + 1\right)n - 1$
$T_0 = \left(\ln(1+nv) - \ln(v+1)\right)n + 2\ln(v+1) - 2\ln(1+nv)$
$D = 2\left[\left(T_1 v + T_0\right)(n-1)(v+1)\right]$

$$E[K \mid R = C] = \frac{N}{D}.$$

Expected knowledge given incorrect response:

Numerator:
N is the same as for expected knowledge given correct response.

Denominator:

$$T_1 = 2n^2 + 2\ln(1+nv) - 2\ln(v+1) + \big(-5 - \ln(1+nv) + \ln(v+1)\big)n + 3$$
$$T_0 = \big(\ln(v+1) - \ln(1+nv)\big)n - 2\ln(v+1) + 2\ln(1+nv)$$
$$D = 2\big[(T_1 v + T_0)(n-1)(v+1)\big]$$

$$E[K \mid R = I] = \frac{-N}{D}.$$

Variance of knowledge:
The expressions for variance of knowledge in the presence of omits are so lengthy that they cannot even reasonably fit in the appendix.

Appendix III: Comparison of CKM to Three Parameter Item Response Model

A popular scoring algorithm that is currently used is the Three Parameter Item Response Model (3IRM). There are several major differences between CKM and 3IRM.

Applying 3IRM requires running a computer program. There is nothing wrong with this in principle; however, a lot of teachers might prefer a simple formula. CKM provides such a simple formula.

Because 3IRM has to estimate various parameters associated with each question, a lot of students have to take the test for 3IRM to work well. Under CKM, it doesn't matter how many students take the test. Thus, CKM can be applied in small class settings.

Finally, 3IRT does not formally account for partial knowledge or misinformation. 3IRT assumes that a correct response means that the student

- guessed correctly with a certain probability; or
- answered correctly with a probability determined by the parameters of the question and his own ability parameter.

3IRT does not account for the fact that the student could eliminate some of the answers and guess correctly from among the remaining answers; it also does not account for the possibility that the student might think that a wrong answer is actually the right answer. Both of these are accounted for by the CKM.