Validity of CTBS from an Analysis of Speededness and Item Response Patterns

G. Philip Nagy[1]
the ontario institute for studies in education

This study examines the validity of the Canadian Tests of Basic Skills from the perspective that several of the subtests of this battery are highly speeded. One tool in the investigation is a modification of the Sato caution index, an index which measures the congruence of individual response patterns with group patterns. Results show that there is very little relationship between the number of items attempted and the number correct. Students who attempt the most questions have answer patterns least congruent with the group pattern, strongly suggesting that those who attempt more items do so by guessing. Slow- and fast-working students are not distributed randomly across schools, thus allowing identification of *slow-* and *fast-working schools*. This suggests that guessing behaviour may be at least partly controlled by teacher direction rather than individual preference. Implications for the validity of item difficulties and comparisons across groups and individuals and across subtests are discussed.

Cet article examine a validité des tests canadiens d'aptitudes élémentaires (Canadian Tests of Basic Skills) en tenant compte du fait que les élèves répondent très rapidement à un grand nombre de ces tests. Un des outils de recherche est une version modifiée du *Sato caution index*, instrument qui sert à mesurer la conformité des répouses de l'individu par rapport à celles du groupe. Les résultats indiquent qu'il y a très peu de relations entre le nombre de répouses données et le nombre de répouses correctes. Les éléves qui tentent de répondre au plus grand nombre de questions sont ceux dont les répouses diffèrent le plus de celles du groupe, ce qui semble indiquer qu'ils essaient plutôt de deviner la répouse. Les éléves qui répondent lentement ou rapidement au questionnaire peuvent être identifiés. C'est donc dire que l'enseignant peut, au moins partiellement, contrôler la tendance des élèves à deviner plutôt que de laisser ceux-ci répondre à leur propre rythme. Ou discute de ce que cela implique en terme de validité des niveaux de difficulté des items, de comparaison entre les groupes et les individus et entre les tests secondaires.

INTRODUCTION

This study serves a dual purpose. First, it examines the validity of the Canadian Tests of Basic Skills (CTBS) from the perspective that several of the subtests of this battery are quite highly speeded. Second, it investigates the behaviour of Sato's caution index (see Harnisch & Linn, 1981), a measure of individual congruence with group item difficulties, in the light of these speededness problems.

The study proceeds by examining, first, the nature and extent of speededness problems at the individual level, second, the relationship between the caution index and speededness, and, third, the nature of the phenomenon when data are aggregated at the school level. At the heart of this study is the issue of test-taking behaviour and the effect of this behaviour on the validity of test result interpretation. That is, under conditions of speededness, can we infer the extent of guessing behaviour on the part of subjects writing the test and the effects of both their guessing and their failure to reach later test items?

The validity questions are of three types and concern the interpretation of individual and group scores, the comparison of individuals or groups, and the interpretation of individual item difficulties. One perspective which will be introduced springs from the literature on test bias. Evidence of bias either for or against those who exhibit a slow or fast test-taking style will be investigated. Since we have no information on how slowly or quickly students have actually worked during the testing, we infer speed from the relative number of items attempted.

BACKGROUND

The literature relevant to this report can be organized around two major themes and several minor ones. The first and oldest deals with the measurement of *speededness* and the relationship of *speed of working* to achievement or ability. The second and more recent concerns the development of indices to assess what might be called, in a general sense, the believability of a test score as an accurate reflection of an individual's achievement or ability. As well, where appropriate, research on item bias, guessing behaviour, item skipping, and the interpretation of scores at the chance level will be mentioned.

Maruhnich (1980) and Rindler (1979) both review the attempts to quantify test speededness. Although he did not initiate the area of study, Gulliksen (1950) gave the basic definitions of speed- and power-tests and a method of measuring the degree of speededness in a single test administration based on items right, wrong, and not reached. While Gulliksen assumed that no items were skipped over, Stafford's (1971) one-administration index assumed that skipped and not reached items were the same. Incidentally, although little research is available on item skipping behaviour, Rindler (1980) found that only better students benefited from skipping over more difficult items. Despite the fact that CTBS directions are silent on item skipping, the distinction between items skipped over and those not reached must be dealt with in this report.

Rindler (1979) criticized attempts to quantify speededness based on a single administration because of the difficulty in distinguishing between items skipped and those not reached. Instead, she suggested the measure proposed by Cronbach and Warrington (1951) involving two administrations, timed and untimed. Despite Rindler's comments, the data set for this study comes from a single administration, and we are thus limited to larger inferences about effects of time limits than Rindler would accept. However, this is not a serious limitation. Maruhnich (1980) compared four single-administration indices, concluding that they were so similar that the simplest, Swineford's (1956) proportion of students who finished 75% of the items, should be used. This will be one method adopted for this study.

There is a long history of research establishing the weak or nonexistent relationship between speed of working and level of knowledge, achievement, or intelligence. Beck (1933) reviewed 34 earlier studies, reporting the following median correlation coefficients with intelligence: for 14 investigations involving reaction time, 0.16; for 5 studies of reading speed, 0.30; for 6 studies involving speed of sentence completion, 0.12; and for 6 studies of response latency, 0.06.

Baxter (1941) was one of the first to point out a basic flaw in some earlier research which had found stronger relationships between speed and achievement than those reviewed by Beck. (Note that more difficult items have been shown to take longer according to Bennett & Doppelt, 1956.) In the work reviewed by Baxter, a typical methodology involved administering a test under strict time limits and then allowing continuation until completion using different coloured pencils. The result was a pair of scores for timed and untimed performances yielding spuriously high coefficients due to their part-whole nature. Baxter found that such a relationship existed along with a correlation between strict time limit scores and time needed to finish. However, when he correlated time needed to finish with total score under untimed conditions, the result was only 0.06 and he concluded:

> whatever is determining the scores in unlimited time is not of major importance in determining the scores in limited time, but it is comparatively important in determining that "aspect" of the [time] limited scores that is useful in predicting college grades. (1941, p. 290).

Other reports of weak relationships include Mollenkopf (1960), Terranova (1972) and Burack (1967).

Early factor analytic studies (Davidson & Carroll, 1945; Lord, 1956;

Myers, 1952) showed similar results, although the small sample sizes must leave questions. Joreskog (1979) reanalysed Lord's data using analysis of covariance structures and concluded that speed itself might be better thought of as more than one factor. In keeping with Joreskog, Bridgeman (1979) also found that speed of work was better explained by more than one factor. Finally, in this vein, Himmelweit (1946) found that personality factors were important in the preference for speed over accuracy. However, as his subjects were institutionalized depressives and hysterics, the results might not be relevant to public education.

Of more relevance, however, are the results of Boag and Neild (1962) who found that, on a reading comprehension test, middle ability students benefited more from extra time than those of either high or low ability. Also, and very pertinent to the issue of test score interpretation, Daly and Stahmann (1968) examined the placement exam results of college freshmen who were candidates for remedial English programs. Two scores were obtained on the pretest, timed and untimed, students being nominated for remediation on the timed version. Results showed that students did no better on the timed posttest than on the untimed pretest. Also, had the untimed results been used for placement, 41% of the group would not have been nominated for remedial classes. Daly and Stahmann questioned whether the students really needed remedial work at all.

While many of the individual studies cited, especially the early ones, can be criticized on the basis of the methods used, they at least point out that we should not take for granted the implicit assumption of a speeded test that a student who did not reach an item probably would have answered incorrectly anyway. One obvious consequence of an inadequate time limit for a test is that students pressed for time might begin to respond by assuming no penalty and guessing (Khan, 1968). However, we should not expect a strong, simple relationship between the press of time and tendency to guess. For example, Traub and Hambleton (1972) found evidence of the influence of test directions and personality factors on the tendency to guess. CTBS directions are silent on guessing.

There has been considerable recent work on methods of determining whether student responses to tests reflect achievement or some other factor such as cheating, guessing, or improper aligning of question numbers and answer numbers. While this work has apparently assumed that tests were unspeeded, it seems appropriate to investigate the effect on student responses when the press of time as well as the lack of knowledge might force some into random or near random responses.

Methods of examining item response patterns for evidence of response on the basis of something other than knowledge are based on Guttman's

(1950) scalogram analysis. Guttman argued that, in a perfectly homogeneous test, if items were arranged from easiest to hardest and coded 0 or 1, then all students' answer patterns would show only 1 up to the level of their ability and 0 beyond that point. Further, if students were arranged from highest achiever to lowest, then the resulting data matrix (students by items) would be characterized by 1 in the top left and 0 in the bottom right, with a clear border between the two regions. An index of correspondence to such a pattern could be interpreted as a measure of test homogeneity.

Extending this principle, Sato (see Harnisch & Linn, 1981) developed a caution index to measure the conformity of both individuals and items to this pattern. High indices suggest that caution should be exercised in interpreting the score of the individual. As the original Sato index was unbounded and thus difficult to interpret, Harnisch and Linn modified it to be bounded by 0 and 1. Extending the applicability of the technique, they have shown that differences in group average caution indices can be traced back to different curriculum practices.

A large number of other indices have been proposed to accomplish the same purpose as Sato's caution index. Harnisch and Linn review and compare many, as do Rudner (1983), Petersen (1980), and Rudner, Getson, and Knight (1980a; 1980b). These indices fall into two groups, those based on item response patterns, such as Sato's, and those based on item characteristic curve (ICC) theory. While preference was expressed for one or the other indices by many of the above authors, as well as by Divgi (1983) and Drasgow (1982), much of this discussion may have been relegated to history by recent work of Tatsuoka and Linn (1983) and Tatsuoka (1984) showing the theoretical relationships between the item response and ICC approaches and suggesting several extensions to the basic Sato index which incorporate the two approaches.

While the extensions of the Sato index by Tatsuoka and Linn are too recent for application in this analysis, one difficulty with the index as modified by Harnisch and Linn has been corrected. First, to clarify terminology, the original Sato index is often referred to in the literature as CI (caution index), while the bounded version due to Harnisch and Linn is MCI (modified caution index). Divgi (1983), among others, has pointed out that MCI is correlated with achievement level, a factor leading to difficulties in comparing MCIs among individuals with different total scores. This problem has been corrected in this study by normalizing the MCI within a group at a given total score, in effect partialling out the correlation with total score. Thus, these modified and normalized indices (MCIZ) are bounded by 0 and 1 and comparable across different achievement levels. Although the utility of an index of caution, interpretability, or believability is obvious enough in general, one particular use is pertinent to this data set in which median scores fall somewhat below the national norms: namely, the use of an index to help determine whether a response near the chance level represents an honest effort or some degree of random response.

CTBS, like the parent Iowa Tests of Basic Skills and many multiple choice norm-referenced tests, offer information in the manual for the assigning of grade equivalent scores to raw scores well below the "guessing level." Although both random guessing and grade equivalents are dubious concepts, for the version used here (Level 10, Form 4) the scores at the chance level fall only between 1.5 and 0.7 "years behind" the norm. Despite difficulties with the notion of truly random guessing, we cannot ignore the fact that, depending on the subtest, between 16% and 32% of subjects score at or below what might be expected on the basis of chance. A caution index provides a way of examining whether a student who scores near the chance level was guessing nearly randomly or actually knew about one-quarter of the information required for the test. The former case gives a high caution index and the latter a low one (Donlon, 1981).

The concept of item bias is, put simply, that different groups perform differently on different items despite controls for overall achievement. That is, group-by-item interactions can be found. Most studies have investigated bias outside the context of speededness. Donlon (1981) reports an earlier unpublished study in which group differences in consistency indices were found when groups were defined on the basis of sex, race, and ethnicity. McArthur (1982) found more inconsistent responses from Spanish-speaking students taking a Spanish translation of an achievement battery than for English-speaking students taking the original. Linn and Harnisch (1981) found that, when blacks and whites from the same performance quintile were compared, item-by-group interactions could be found.

There is a small body of work investigating problems of bias in the context of speededness problems. As an aside important for speeded norm-referenced test, Cronbach and Warrington (1951) have shown that the KR-20 measure of reliability, almost universal in test manuals, is not accurate to estimate the reliability of a speeded test and is at best only an upper bound. Dubin, Osburn, and Winick (1969) and Evans and Reilly (1972) have both shown that minority and low SES groups benefited equally with majority and high SES groups from the provision of extra time. The latter study also reported that the test was more speeded for the minority students, although the degree of speededness seems relatively small by comparison with the present data set. Moreton and Butcher (1963) addressed the issue of whether speeded tests were more penalizing for rural children than for urban children. They found small differences, and, although the issue of rurality is pertinent to this study, the British context of their study may limit its generalizability. Finally, Douglass (1981) found in a speeded test that the degree of item bias between blacks and whites and item position were related, indicating that speededness was having differential effects on the two groups.

SAMPLE

The sample for this study consists of item scores for more than 12,000 Grade 4 students who wrote the CTBS, Form 4, Level 10, in October, 1978 as part of the testing program of the Newfoundland Department of Education. CTBS consists of 425 items in 11 subtests, and is administered over three or four days. Exact sample sizes vary from subtest to subtest due to absences. Virtually all Grade 4 students in the province wrote the test.

The Newfoundland population is, by urban Canadian standards, rather thinly spread. The 12,000 students attended Grade 4 in 445 schools, an average of fewer than 30 per school. Of these schools, more than 100 had fewer than 10 students in Grade 4 at the time of testing.

DATA ANALYSIS

Each subtest was first analyzed separately with no comparison being made among subtests until later. Due to the costs of manipulating such a large data file, not all analyses were completed on all subtests. Since the object of investigation was the phenomenon of speededness, once this had been assessed for each subtest, representative subtests were chosen for complete analysis. The analysis proceeded as follows:

1. Initial student responses were coded as right, wrong, skipped over, or not attempted. Four total scores were developed for each student.
2. Columns and rows of the data matrix were rearranged, and the MCI (Harnisch & Linn, 1981) calculated for each student. Students with perfect or zero scores were assigned an MCI of 0.0.
3. Students were arranged by achievement level, and the MCI was converted to a z-score, MCIZ, within each achievement level. In those cases where only one student had a given raw score (at either extreme, for example) this calculation was not possible and the MCIZ became

missing data for the student. This happened in fewer than 50 of the more than 130,000 cases.

4. Average values for each school were calculated for each of the item scores (number right, wrong, skipped, and not reached) as well as for MCI and MCIZ.
5. Analysis proceeded by cross tabulations, frequency counts, and correlations. Details are given below in the reporting of the results.

RESULTS

Degree of Speededness

Table 1 contains basic data on the degree of speededness and on the difficulty of the subtests. It also lists the short forms of the subtests which will be used for the remainder of this report. The second column contains the Swineford measure of speededness, the percentage of students reaching 75% of the items. This count assumes that items skipped were duly considered and not answered. The third and fourth columns contain two other measures of test speededness, the percentage of students not reaching the last three items, and the percentage of items not reached by the average student. On two of the subtests, W1 and W3, barely half the students get to the end of the test. These two subtests have the most serious speededness problems. While there are no concrete criteria by which to judge satisfactory speededness, a test in which the average student cannot get to the last 10% of the items seems heavily speeded. By this criterion, W1 and W3 would be judged seriously speeded; V, R, L1, and L2 marginal; and the remainder satisfactory. Further comment on this issue will appear with more detailed results.

Table 1 also contains, in column 5, data on the mean number of items skipped. As can be seen, item skipping is not widespread. It is difficult to argue, in the light of these data, that skipping is more prevalent in speeded tests than in nonspeeded ones. The V results are out of line with the others, but that seems unrelated to any other measures.

The final column reports the percentage correct for the average student. In comparison with the grades students in elementary schools are typically assigned for classroom tests, these are quite low from a student's perspective. Students are not given the raw scores, of course, but one must wonder about the psychological effect of writing a test of much greater difficulty than classroom norms. As mentioned before, the directions for administration suggest no warning to students that they will

TABLE 1

Magnitude of Speededness, Skipping, and Difficulty

Subtest	*(1) Number of items in subtest*	*(2) Percentage of students reaching 75% of items*	*(3) Percentage of students reaching last three items*	*(4) Mean (across students) percentage of items not reached*	*(5) Mean (across students) percentage of items skipped*	*(6) Mean (across students) percentage items correct*
Vocabulary (V)	38	81	62	10.9	2.6	36.6
Reading (R)	68	82	63	9.9	1.0	42.5
Spelling (L1)	38	83	63	10.3	0.9	42.9
Capitalization (L2)	39	83	63	10.4	1.1	38.8
Punctuation (L3)	39	90	77	6.3	1.1	35.4
Usage (L4)	32	93	86	3.9	0.7	38.9
Map reading (W1)	32	71	55	15.0	1.8	36.6
Graphs and tables (W2)	24	91	78	5.7	0.9	47.2
Reference materials (W3)	52	70	51	16.6	1.5	35.2
Math concepts (M1)	36	93	81	4.3	1.5	41.1
Math problems (M2)	27	80	74	7.1	1.1	38.9

find the test difficult nor that they are not expected to finish. For clarity, it should be noted that these students scored somewhat below the national average, clustering roughly around the 30th percentile.

Item Skipping

Since one of the difficulties with this type of analysis is the inference needed to interpret an unanswered item, a closer look at item skipping is required. At issue is whether a skipped item has been duly considered so that the response should be counted as wrong, or whether it was skipped only because later items looked, at first glance, to be easier.

In order to distinguish between students who skip a few items as part of a serious attempt to deal with the test and those who in effect choose to answer only a few items, students were divided into three groups, those skipping up to two items, skipping between three items and 25% of the items, and skipping more than 25% of the items. These students will be referred to as low, medium, and high skippers. Students were also divided into quintiles by achievement so that the relationship between achievement and item skipping could be examined for each subtest. Since scores come in whole numbers, the division by quintiles could only be approximate.

As might be expected, since skipping was a relatively uncommon phenomenon, the low skippers were distributed quite equally across achievement levels. The high skippers, who typically comprised between 1 and 2% of the entire sample, were heavily over-represented in the first quintile. About 60% of these students were in the lowest achievement group, while the remaining quintiles had approximately 20, 10, 5 and 1% of the high skippers. This pattern shows that not all students who skip a lot of items are the lowest achievers. Further insight can be gained by considering the results for the middle skipping group by subtest, as reported in Table 2.

As can be seen from Table 2, the behaviour of the moderate skippers is quite complex over the 11 subtests. From the final column, it can be seen that more students skip items in the speeded tests W1 and W3, but this is also true of V and R, two moderately speeded tests, and M1, one of the least speeded. There is more of a tendency within some tests than others for higher achievers to skip a moderate number of items. This is true of the two highly speeded subtests, W1 and W3, but even more so of M1, one of the least speeded. In fact, the distribution of item skippers in M1 is essentially independent of achievement.

There are other interesting aspects to the data in this table. Within W1, there are more moderate skippers in the third and fourth quintiles than in

TABLE 2
Achievement Distribution of Students Skipping from 3 Items to 25% of Items

	Quintile					
Subtest	*1*	*2*	*3*	*4*	*5*	*Number*
V	31.5	24.1	17.3	15.0	12.0	971
R	31.5	25.2	16.7	19.5	7.0	682
L1	37.5	29.9	16.9	12.3	3.4	261
L2	33.5	22.6	21.2	14.5	8.1	433
L3	30.3	28.2	17.0	11.9	12.6	429
L4	36.1	29.7	18.3	9.4	6.4	202
W1	25.6	16.7	25.9	21.3	10.6	700
W2	30.3	17.5	29.4	17.5	5.3	228
W3	24.7	27.9	19.2	16.3	11.9	806
M1	21.6	20.7	19.8	18.9	19.0	652
M2	27.9	18.4	24.7	16.2	12.9	365

the second. Similar patterns are evident elsewhere, but some might be written off to small sample sizes. It is interesting to note two quite different interpretations of the skipping pattern for W1: either more of the better students are forced into skipping with the press of time or those who benefit from skipping tend to be the better students. Although these data are interesting, little more can be said without access to further information, particularly on the match of curriculum to tests.

Speed and Achievement

The analysis of speed and achievement compares items not reached with items correct. That is, items skipped are treated as incorrect. Since, however, relatively few items were skipped, treating them as not reached would not change the following analysis, summarized in Table 3, very much.

Pearson product-moment correlations between speed and achievement range from 0.18 to 0.45. Although there are several exceptions, notably the two mathematics tests with the weakest relationships, the trend is for the more speeded tests to have the weaker relationships between speed and achievement. In other words, although the relationship between speed and achievement is moderate at best, it weakens under the press of time. Also, as the data in Table 3 suggest (compare, for example, V and R), the relationship is not a simple linear one.

TABLE 3
Percentage of Second and Fourth Quintiles Attempting less than One-half, Two-thirds, and Four-fifths of Items

	Less than one-half		*Less than two-thirds*		*Less than four-fifths*	
Subtest	*Q2*	*Q4*	*Q2*	*Q4*	*Q2*	*Q4*
V	5.5	0.1	16.0	3.0	29.5	10.4
R	3.9	0.0	12.0	1.5	24.2	17.4
L1	3.7	0.1	12.7	1.3	25.5	8.0
L2	4.6	0.1	15.3	2.2	24.1	7.5
L3	2.5	0.1	6.6	0.7	12.1	3.4
L4	1.2	0.1	4.5	0.5	10.8	3.1
W1	10.2	1.0	27.8	14.9	38.0	34.8
W2	1.4	0.0	9.5	1.4	16.5	9.7
W3	17.2	0.7	32.3	9.3	44.1	28.2
M1	1.0	0.0	5.3	2.0	8.1	9.3
M2	4.0	0.2	10.7	5.9	19.5	20.2

In Table 3, the overall trend is for students in the fourth quintile to have answered more items than those in the second. From the first pair of columns, it can be seen that few students, even in the second quintile, attempted fewer than one-half the items. However, these numbers are largest in the two highly speeded tests, W1 and W3. This pattern tends to hold up for the other two fractions of the tests examined. However some nonlinear trends bear examining. If we compare R and L1, which are quite similar in overall speededness, the values are much the same except for the right-hand column which reports the proportion of the fourth quintile which attempted fewer than 80% of the items. Although the proportions of the second quintile in this speed category are close, more than twice as many higher achieving students ran out of time in R than in L1. Thus, more high-ability students were affected by time limits in R than in L1. The same sort of conclusion can be drawn for M2 and W2, two of the less speeded tests.

Two other noteworthy observations come out of this table. First, a comparison of W1 and W3, the speeded tests, confirms that speed of working is least related to achievement under heavy time pressures. Second, in M1 and M2, more of the fourth quintile than the second failed to attempt more than 80% of the items. In this case, we have evidence that the better students were actually working more slowly than the weaker ones.

The Caution Index

The modified caution index (MCI) of Harnisch and Linn (1981) gives a value of 0 for a personal order of difficulty perfectly in line with that of the group, and 1 for a personal order exactly contradictory to that of the group. Thus, if a student tends to get items correct which the group finds easy and tends to miss those more difficult for the group, the inference is that the student's responses represent an honest effort at the test and the total score is a reasonable estimate of achievement. On the other hand, if the student tends to get easy items wrong and harder items right or has successes which seem random with respect to group difficulty, then this can point to some reason for caution, such as misalignment of question and answer numbers, cheating, or guessing.

Across individuals, the correlation of the MCI (not normalized) with achievement ranged from −0.07 to −0.36 over the eight subtests for which this calculation was performed. One scatterplot, done on the R results, suggested that the relationship was not linear and that the variance of the MCI was greater at low achievement levels than at higher. Whether this result is an artifact of the calculation or really reflects the fact that more caution is required in the interpretation of lower scores is not answerable without looking at the MCIZ results.

First, the correlation between MCIZ and achievement is, by definition, 0. For eight subtests, the correlation between MCIZ and number of items attempted ranged from 0.30 to 0.65. That is, students who answered the most questions had the highest caution indices. This suggests strongly that, on the average, those who answered the most questions did so at least partly by guessing. The correlation reported is highest for W1 and W3, 0.65 and 0.62 respectively. To add further weight to this interpretation, when the group is split by achievement, above and below the mean, the relationship between items attempted and MCIZ is consistently stronger for the lower achievement group than for the higher, reaching 0.73 for the lower achievement half of the W3 group. In light of these data, no other explanation for such a strong relationship is more compelling than that of guessing in the light of time pressures.

Aggregated School Data

With strong evidence that students are guessing at items when time is short, it is natural to ask next why this is the case for some students and not for others. We are limited by the absence of any personality trait information on the students, but we can look at group effects with the data aggregated by school. If guessing behaviour is an individual character trait, then we would expect that it would be distributed throughout the sample independently of school. Thus, with aggregated data, the variance of both MCIZ and items attempted would be reduced as would the correlation between them. However, at the school level, two facts emerge.

First, there is already evidence that achievement is not distributed randomly across schools in this data set (Newfoundland Department of Education, 1979). Over eight subtests, the correlation between school average achievement and items attempted ranged from 0.10 to 0.50, with a median value of 0.25. No pattern within the values can be discerned. However, when school average MCIZ is correlated with items attempted, the values range from 0.36 (M1) to 0.77 (W1 and W3). In every case, the number of items attempted is more closely related to MCIZ than to achievement. To guard against the danger that these results were an artifact of extremely small school sizes, the calculations were done with all schools (N =445) and only with those with ten or more students in the grade (N = 329). Results were substantially the same. Those reported here are without the very small schools.

This suggests that, despite the absence of direction on the part of the CTBS manual, students tend to behave as a group, either guessing a lot (high MCIZ) or a little (low MCIZ). This suggests that students follow school level (or perhaps even board level) directions about what to do in the event of not knowing the answer. Such an hypothesis can be investigated by an examination of schools' speed across different subtests.

Fast and Slow Schools

There were a number of schools with an average number of items missed of 0.0. That is, in some schools, all students reached the last item. These schools were isolated for each subtest, and the listed schools compared across subtests. There were 742 cases out of a possible 4,895 (11 subtests times 445 schools). However, these 742 cases were confined to 172 schools. In fact, there were 12 schools which appeared on the list 11 times, once for each subtest. Forty-five of the 445 schools account for 55% of these situations, strongly suggesting that the students were acting under direction. Further, 57% of these cases can be attributed to only eight school boards. These school boards are responsible for only 33% of the schools in the province, but 59% of their schools appear on the list of fast-working schools at least once.

Such a result might not be surprising since common sense suggests this would be the case when there is no penalty for guessing. However, it is distressing from two points of view: first, to be discussed later, for the validity of item interpretation; and, second, for the validity of school comparisons when not all schools direct their students in this manner. To this second point, we now turn our attention.

It is natural to ask, since we found consistently fast-working schools, whether we could also find consistently slow-working schools. But, since some of the slow-working schools would be those where most students were simply not up to the educational expectations of the CTBS, we should confine interest to slow-working but high-achieving schools. To this end, for each subtest, the 100 slowest working schools were identified. Then, from this group, the highest achieving one-third of the schools were selected. Then, with this selection of the highest achieving third of the slowest quarter (roughly) of the schools, comparisons were made across subtests. In this case, only 36 (8%) of the schools accounted for 54% of the cases of the slow, high-achieving schools.

A check on achievement on four of the subtests showed that the fast group was about 5% below the provincial average and the slow group about 5% above. The method of selection of the slow group guarantees moderate performance, so not too much should be read into this. However, the point stands that neither group differed remarkably from the provincial mean except in speed of working. Finally, it is important to note that the average MCIZ values for the slow group were quite low and those for the fast group quite high, indicating the guessing by more students (schools) in the fast-working groups.

Validity and Bias

In a trivial manner, any item a student does not get a chance to answer is biased. However, apart from that and in the absence of information on either instructional practices or match between test items and curriculum, we can approach the problem of validity from two angles: first, from an examination of rank order of item difficulties for different groups and, second, from an examination of which portions of the test produce the correct answers for different groups.

For this analysis, three subtests were selected, L4 and W1 because they were close to the two extremes of speededness and M1 because of the almost complete lack of relationship between speed of working and achievement. Within each subtest, four groups were identified, *slow* and *fast* students within the second and fourth achievement quintiles. Slow students were those who attempted only between 50% and 80% of the items, and fast were those who failed to reach at most two. Next, item difficulty rank orders were obtained for each group and for the entire sample. Rank order correlations were calculated for each group with the entire sample and for each combination within either the same quintile or the same speed range. Calculations were done only on items for which the slowest group had a minimum number of responses (200 in most cases, fewer in two).

Rank order correlations with the total sample item difficulties were all over 0.80 with the exception of the fast second quartile group writing W1 (0.61). This suggests that this group is among the most prone to guessing. On the comparisons, the subgroups were above 0.80 with the exception of the two second quartile groups in L1 (0.76) and W1 (0.53) and the two fast groups in W1 (0.70). The second of these correlations, 0.53 in W1, in particular demonstrates that many items behave quite differently for the slow test-takers than for the fast. Note that these coefficients are calculated only on the items reached by both groups.

Those who answer many items obtain their correct answers from the entire test, while those who work more slowly must get their marks only from the items they reach. This obvious statement masks the phenomenon that slow-working students must do considerably better than fast-working students from the same overall achievement level on the items they do reach.

Using the four subgroups from each subtest just described, within the second quintile the fast-working students obtained the following percentages of the scores of the slow-working students on the items both groups reached: W1, 74; L4, 71; and M1, 72. This means that the fast group earned 26%, 29%, and 28% of their correct answers on items that the slower working students did not have a chance to attempt. Within the fourth quartile, the results show less of a difference, with the fast students earning, respectively, 21%, 19% and 26% of their scores on items the slow group did not reach.

DISCUSSION

These results provide evidence that, first, at least some of the subtests of CTBS with a group somewhat below the national norm have a serious problem with too restrictive time limits. There is also evidence that the time limits affect performance in complex ways, nonlinearly with respect to achievement in some cases and quite independently of achievement in others. The worse the speededness problem in a subtest, the weaker the relationship between speed of working and achievement.

Use of the modified and normalized caution index suggests that, on average, those who answer a lot of items guess considerably. When the data are aggregated at the group level, there is evidence that some teachers are probably encouraging students to guess while others are not. Given the evidence above about the relationship between speed and achievement and the behaviour of the caution index, the index seems to serve a useful role in identifying test takers whose behaviour must be questioned in the face of a speeded test.

If groups which have attempted any or few items are identified, there are noticeable differences across them in the relative difficulty of items. Of course, it is possible that two random samples from the entire data set would produce quite different rank orders of item difficulty, but, in light of the other evidence, this is not likely.

The main attacks on the validity of CTBS coming out of this study are:

1. Drawing curriculum implications from individual items when unknown but substantial numbers of students are guessing should not be attempted.
2. Comparing students' total scores when there are large and systematic differences in the number of items attempted is difficult if not impossible unless information such as the MCIZ is used as well.
3. Comparing group averages when there is evidence that one subgroup of schools operates under those in instructions quite different from other schools is difficult if not impossible unless directions are made uniform. This, of course, would invalidate the national norms established in the absence of specific instructions on guessing.
4. Given at best a weak relationship between speed of working and achievement, it is not recommended that the first three points be ignored on the supposition that students failing to reach an item would have got it wrong.
5. Given that the effects of speed are different and nonlinear across subtests, comparisons of the relative standing of students or groups across subtests should not be undertaken.

CONCLUSION

This study had the purpose of investigating both the behaviour of a modification of the Sato caution index and the validity of the CTBS. It is limited in that the sample is far enough below the national norm that results might not be as extreme for a group closer to the norm. Also, in the absence of any concrete information on directions given to students, we are limited to inference about those directions and about whether students actually guessed when they were pressed for time and produced a high caution index. We also have no information about the match of curriculum and test either within a subtest across schools or across subtests. Since there are many unexplained divergences from what would be expected in the data, a study of test match to curriculum or instruction is needed to investigate more fully the validity of CTBS. Given what can be seen from these results, however, perhaps further work on the validity of this particular test in this context is not worth the effort.

One untapped source of information exists within the present data set, that of response choices. This analysis proceeded on a right-wrong basis, ignoring the potentially fruitful area of which wrong answers were chosen by which students. Again, evidence of substantial differences in patterns of wrong answers would further invalidate the test as a measure of the same achievement trait across students.

This study suggests caution in interpreting results from this test, especially for students below the test norms. At the same time, more long-term benefit will ensue from work aimed more at test design. Considerable information can be gleaned by looking at which wrong answers are chosen rather than by treating scores as merely right or wrong (Alderman, Swinton, & Braswell, 1979; Levine & Drasgow, 1983). The group centred at Illinois is developing a line of research in which carefully developed distractors are used with caution indices to pinpoint student misconceptions (Birenbaum & Tatsuoka, 1982; Harnisch, 1983; Harnisch, 1981; Tatsuoka, 1983; Tatsuoka & Tatsuoka, 1983). Their efforts currently focus on analyzing student misconceptions in the area of addition of integers and dealing with the difficulty that, in a great many cases, incorrect algorithms can lead to correct answers. The end target of such work is to develop true diagnostic tests useful for more than comparing the total scores of individuals or groups.

NOTE

[1] This research was supported by funds from the Ontario Institute for Studies in Education. I wish to thank the Newfoundland Department of Education for access to the data.

REFERENCES

Alderman, D. L., Swinton, S. S., & Braswell, J. S. (1979). Assessing basic arithmetic skills and understanding across curricula: Computer-assisted instruction and compensatory education. *The Journal of Children's Mathematical Behavior, 2*, 3–28.

Baxter, B. (1941). An experimental analysis of speed and level in an intelligence test. *Journal of Educational Psychology, 32*, 285–296.

Beck, L. F. (1933). The role of speed and intelligence. *Psychological Bulletin, 30*, 169–178.

Bennett, G. K., & Doppelt, J. E. (1956). Item difficulty and speed of response. *Educational and Psychological Measurement, 16*, 494–496.

Birenbaum, M., & Tatsuoka, K. K. (1982). On the dimensionality of achievement test data. *Journal of Educational Measurement, 19*, 259–266.

Boag, A. K., & Neild, M. (1962). The influence of the time factor on the scores of the Triggs Diagnostic Reading Test as reflected in the performance of secondary school pupils. *Journal of Educational Research, 55*, 181–183.

Bridgeman, B. (1979). *Generality of a "fast" or "slow" test-taking style across a variety of cognitive tasks*. Paper presented at a joint meeting of the American Educational Research Association and the National Council on Measurement in Education, San Francisco.

Burack, B. (1967). Relationship between course examination scores and time taken to finish the examination, revisited. *Psychological Reports, 20*, 164.

Cronbach, L. J., & Warrington, W. G. (1951). Time limit tests: Estimating their reliability and degree of speededness. *Psychometrika, 16*, 167–188.

Daly, J. L., & Stahmann, R. F. (1968). The effects of time limits on a university placement test. *Journal of Educational Research, 62*, 103–104.

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time limit scores: A factor analysis. *Educational and Psychological Measurement, 5*, 411–427.

Divgi, D. R. (1983). *Some reservations concerning indices for unusual response patterns*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Donlon, T. F. (1981). Uninterpretable scores: Their implication for testing practice. *Journal of Educational Measurement, 18*, 213–220.

Douglass, J. B. (1981). *Item bias, test speededness, and Rasch tests of fit*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6*, 297–308.

Dubin, J. A., Osburn, H., & Winick, D. A. (1969). Speed and practice: Effects on Negro and White test performances. *Journal of Applied Psychology, 53*, 19–23.

Evans, F. R., & Reilly, R. P. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9*, 123–131.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley and Sons.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stauffer (Ed.), *Measurement and prediction*. Princeton, NJ: Princeton University Press.

Harnisch, D. L. (1981). *Analysis of item response patterns: Consistency indices and their application to criterion referenced tests*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*, 191–206.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133–146.

Himmelweit, H. T. (1946). Speed and accuracy of work as related to temperament. *British Journal of Psychology, 36*, 132–144.

Joreskog, K. G. (1979). Analyzing psychological data by structural analysis of covariance matrices. In J. Magidson (Ed.), *Advances in Factor Analysis and Structural Equation Models*. Cambridge, MA: Abt Books.

Khan, S. B. (1968). The relative magnitude of speed and power in SCAT. *Journal of Educational Measurement, 5*, 327–329.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675–686.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109–118.

Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21*, 31–50.

Maruhnich, N. (1980). *An empirical comparison of four indicators of test speededness*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

McArthur, D. L. (1982). *Detection of item bias using analyses of response patterns*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. *Educational and Psychological Measurement, 20*, 223–230.

Moreton, C. A., & Butcher, H. J. (1963). Are rural children handicapped by the use of speeded tests in selection procedures? *British Journal of Educational Psychology, 33*, 22–30.

Myers, C. T. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika, 17*, 347–352.

Newfoundland Department of Education. (1979). *Newfoundland standards testing program, Grade Four, 1978: Final report*. St. John's.

Petersen, N. S. (1980). Bias in the selection rule, bias in the test. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruiter (Eds.), *Psychometrics for educational debates* (pp. 103–122). Chichester, U.K.: John Wiley and Sons.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*, 261–270.

Rindler, S. E. (1980). The effects of skipping over more difficult items on time-limited tests: Implications for test validity. *Educational and Psychological Measurement, 40*, 989–998.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20*, 207–220.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213–234.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980b). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17*, 1–10.

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. *Journal of Educational Measurement, 8*, 275–277.

Swineford, P. (1956). *Technical manual for users of test analyses* (Statistical Report 56-42). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K. (1983). Rule space: An approach to dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95–110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7*, 81–96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20*, 221–230.

Terranova, G. (1972). The relationship between test scores and test time. *Journal of Experimental Education, 40*, 81–83.

Traub, R. E., & Hambleton, R. K. (1972). The effect of scoring instructions and degree of speededness on the validity and reliability of multiple-choice tests. *Educational and Psychological Measurement, 32*, 737–758.

Diese Abhandlung untersucht die Gültigkeit der kanadischen Tests der grundlegenden Fähigkeiten und kommt zum Schluß, daß mehrere der Teiltests der Gruppe zu schnell ausgefürt werden müssen. Die Untersuchung stützt sich auf eine Abänderung des Sato Caution Indexes welcher die Übereinstimmung zwischen den Antworten von Einzelpersonen und denen von Gruppen mißt. Die Ergebnisse zeigen, daß die übereinstimmung zwischen der Zahl der Fragen die probiert werden und denen die richtig beantwortet werden, gering ist. Das Antwortschema von Schülern die die meisten Fragen probieren, stimmt am wenigsten mit dem Antwortschema der Gruppe überein welches den Schluß zuläßt, daß die Ersten die Antworten raten. Schulen, obwohl sie gleiche Leistungen aufweisen, können durch schnelle oder langsame Arbeitsweisen unterschieden werden. Dieses deutet darauf hin, daß das Raten, wenigstens zum Teil, von Lehrern bestimmt wird und nicht von den einzelnen Schüler. Die Auswirkung dieser Folgen auf dir Gültigkeit der Schwierigkeit der Fragen und Vergleiche zwischen Gruppen und Einzelnen und zwischen Teiltests werden diskutiert.

Este estudio examina la validez del Canadian Tests of Basic Skills (Tests canadienses de habilidades básicas) desde la perspectiva de que varios de los sub-tests de esta bateria se haven demasiado rápido. Una de las herramientas de la investigación es una modificación del Sato caution index (indice de caución Sato), un indice que mide la congruencia de las respuestas individuales con las del grupo. Los resultados muestran que hay muy poca relación entre le número de items intentados y el número que resultan correctos. Los estudiantes que intentan contestar el mayor número de preguntas tienen las contestaciones menos congruentes con las del grupo como un todo, lo que indicaría que los que intentan más items están adivinando. Se pueden identificar escuelas que trabajan lenta y rápidamente al mismo nivel de rendimiento, indicando que adivinar puede estar parcialmente controlado por las instrucciones del docente, más que por la preferencia individual. Se discuten las implicaciones para la validez de la dificultad de los artículos y las comparaciones entre grupos e individuos, y entre los sub-tests.

G. Philip Nagy is an associate professor in the Department of Measurement, Evaluation, and Computer Applications and acting head of the Midwestern Centre of the Ontario Institute for Studies in Education, 25 Joseph Street, Kitchener, Ontario N2G 1H9.