Individual Score Validity in a Modest-Stakes Adaptive Educational Testing Setting

Carl Hauser and G. Gage Kingsbury, Northwest Evaluation Association

April 16, 2009

Paper presented at the Annual Meeting of the National Council on Measurement in Education,

San Diego, CA

Individual Score Validity in a Modest-Stakes Adaptive Educational Testing Setting

Carl Hauser and G. Gage Kingsbury, Northwest Evaluation Association

(April, 2009)

Perhaps no concept in modern educational tests and testing has been the topic of more philosophical discussion and study than that of *validity*. Over the past 100 years, validity has been examined from a number of perspectives (see Kane, 2001 for a concise history). For example, a cursory review of the two chapters devoted to validity and its measurement in Anastasi (1976) reveals an accumulation of 10 different forms of validity, most with several aspects to them. These forms of validity were extended in the 1980's and 1990's to include the identification of the intended uses and consequences of observed scores (e.g., Mehrens, 1997; Messick, 1981, 1988, 1989; Moss, Girard & Haniford, 2006; Shepard, 1993, 1997).

In virtually all the work undertaken to define and understand what should constitute validity and its bounds, the construct has been treated as a fundamental characteristic of a test. Given the depth and breadth of this century of work, practitioners have an abundance of tools and methods available for assembling comprehensive and compelling bodies of evidence to support claims of a test's validity *in the general case*. However, as we have argued previously (Kingsbury & Hauser, 2007; Hauser, Kingsbury, & Wise, 2008) evidence of validity at the population level does not imply validity for a specific, individual test taker. This argument is simple and straight-forward.

In general, validity arguments require test takers to interact with the test. Such interactions usually result in summaries of performance, most commonly in the form of scores. While the scores are used in combination with other information to build validity evidence, the quality of test-test taker interactions that gave rise to them are commonly assumed to be consistent, if they are considered at all. The exceptions might be in the extreme cases where, for example, a test taker becomes ill, or refuses to take the test seriously or otherwise obviously provides answers that sabotage the test purpose. Other than cases such as these, test-test taker interactions are ignored and tolerated as a nuisance factor contributing to measurement error. From the perspective of treating validity as a characteristic of the test, this practice may be reasonable, at least on statistical grounds. Ignoring test-test taker interactions at this stage is likely to have minimal impact on the type of validity evidence being assembled. Group performances are being summarized and presented; test validity coefficients are essentially estimates of covarying group performances. Therefore, validity arguments whether they are made to demonstrate the relationship of a test's performance to a relevant standard, to support a test score use, or to support its consequences, are made from the general case.

However, validity claims formed in the general case may not apply to an individual's score when test-test taker interactions that resulted in the score are ignored. If the test taker's performance has not resulted

from engaged interaction with challenging test content, the end score will be a less precise and perhaps less accurate estimate of the test taker's status on the trait of interest. It is in this sense that we consider *individual score validity* to be a necessary first step in attributing any validity argument, formed in the general case, to the specific case of the individual test taker.

Recent investigations of examinee test-taking behaviors have suggested that several readily available variables in a computerized adaptive testing environment may be leveraged in the study of individual score validity. Common to all these investigations is the use of item response latency – the elapsed time from the instance a test item is displayed on the computer monitor until the test taker has submitted their response. Response latency has been used to examine the effects of speeded tests (e.g., Schnipke & Scrams, 1997, 2002), cheating (van der Linden & van Krimpen-Stoop, 2003), item bank compromise and appropriation (Wise & Kingsbury, 2006), and test taker effort (e.g., Wise, Bhola, & Yang, 2006; Wise & Demars, 2006; Wise & Kong, 2005). While most all this work holds implications for examining individual score validity, the work focused on using response latency in studying test taker effort is most directly related. For example, Wise and Demars (2006) showed that by modifying the 3PL model to include a condition for response latency as an indicator of effort, better model fit, more accurate parameter estimates, and more accurate test information estimates were attained relative to the standard 3PL. Kingsbury and Hauser (2007) and Hauser, Kingsbury and Wise (2008) using test events from computerized adaptive tests looked at the use of mean adjusted response latency in combination with response correctness, item response residuals, the quality of item targeting in the test and standard error of change in test scores (growth) in investigating indicators of individual score validity.

One common feature of these studies is that they focus on variables or combinations of variables that can serve as plausible indicators of test score invalidity. Conspicuously absent, however, is an assessment of the consequences of ignoring the effects of the indicator variables on the quality of the final proficiency estimates, though Wise and Demars (2006) come closest to addressing this issue. In most cases this omission is appropriate; the field is only in the early stages of identifying or creating indicator variables and understanding how these variables relate to the quality of proficiency estimates. Nevertheless, the presence of indicator information suggesting that odd or unexpected aspects of the test-test taker interaction cannot be ruled out as an explanation of the test taker's performance is, at best, a qualified qualitative finding. The consequences of ignoring such findings have not been made explicit.

The purpose of this study is to examine the immediate consequences of ignoring odd or unexpected test-test taker interactions that could adversely affect proficiency estimates in operational, low to moderate stakes computerized adaptive tests. Immediate consequences, here, are considered to be the effects on resulting proficiency estimates and their standard errors. The study focuses on test events that are "suspicious" from a test-test taker interaction perspective. Short (< 3 seconds) response latencies and low proportion correct are used to define suspicious test-test taker interactions. Test events that are clearly free of odd or unexpected test-test taker interactions are not included.

Method

Sample data.

The test records of over 69,000 students were retrieved from the *Growth Research Database* (GRD, NWEA, 2007). All test records were from computerized adaptive reading and mathematics tests administered in the spring of 2007 to students in a single state as part of their district-sponsored testing programs. A total of 1044 schools located in 313 school districts were represented. Although only limited information is available about test use in these specific schools and districts, tests from this testing system commonly range from low to moderate stakes for individual students. The reading test records were from students in grades 3 and 9. Mathematics tests were from students in grades 4 and 10. All the available test records from the state and testing term were included; no attempt was made to sample from this collection. Approximately 82% of the students represented had an ethnic code of European American.

Content in all tests was aligned with the state content standards at the strand level. All tests were fixed length with 40 operational items in reading tests and 50 operational items in mathematics tests. While these tests were designed as power tests, the possibility that they became speeded tests in practice cannot be ruled out.

For each test record included, the overall test score on the NWEA RIT scale and its standard error were retrieved. In addition, the item response record from each test was retrieved. Item response records included each item presented to the student, its difficulty using the one parameter logistic item response model, the student's response, whether the response was correct, the number of seconds that the student took to give a response and the maximum likelihood estimate of theta after responding to the item.

Overall performance at each grade level is presented for reading and mathematics in Table 1. Performance in reading (grades 3 and 9) was slightly above the NWEA 2008 norm levels by .11 and .13 SD, respectively. Mathematics performance for grade 4 was .26 SD above the NWEA norms; for grade 10, mean performance was slightly below the NWEA norms (-0.08 SD).

Table 1.
Performance on target tests by subject and grade level.

	RIT scores					SEM		
Grade	Mean	SD	Med	p05	p95	Mean	SD	N
				Reading				
3	200.5	13.63	202	175	220	3.36	0.108	16209
9	224.3	14.86	227	197	243	3.36	0.141	18705
				Mathematics				
4	215.0	12.76	215	194	234	3.00	0.101	15532
10	235.6	19.67	238	201	264	3.03	0.161	18718

Unexpected Test-Test Taker Interactions.

To identify odd and unexpected test-test taker interactions, a straight-forward procedure developed by Wise, Kingsbury and Hauser (2009) was used. This procedure used a set of rules which, if violated, would trigger one or more flags related to several aspects of each test event. Flags were triggered by applying criteria tied to response latency and proportion of items answered correctly. The criterion for each flag is summarized in Table 2, below. A more complete description and explication of them is provided in Wise, et.al. (2009). In its intended use, triggering any of the five flags would be interpreted as an indication that the proficiency estimate from the test event should be treated as an untrustworthy estimate of the test takers true proficiency.

Table 2.
Flags used to identify unexpected test-test taker interaction.

Flag	Criterion
1	Response latency < 3 seconds to at least 15% of all items
2	Less than 30% of all items answered correctly.
3	No more than 20% of items answered correctly AND response latency < 3 seconds to at least 3 items in any of the 10-item rolling subsets.
4	No more than 20% of items answered correctly in at least 20% of the 10-item rolling subsets.
5	Response latency < 3 on at least three items in 20% of the 10-item rolling subsets.

The criteria given in Table 2 were applied to all test events, recording the status of each flag for each test event.

Follow-up procedures.

As a method of estimating the impact of ignoring odd or unexpected test-test taker interactions from a test event, a sequence of test event manipulations was used. These were intended to first obtain more valid estimates of the test takers' proficiency levels and then to use those levels to estimate the loss of information from the original test event which would commonly be unknown without attending to test-test taker interactions. These procedures were limited to those test events which were flagged for one of the response latency related criteria - Flags 1, 3, and 5 in Table 2.

Rescore 1 (Rs1). Test events were prepared for rescoring by treating all item responses given in less than 3 seconds as missing items. The shortened test events were rescored with the standard ML scoring routine used for all tests.

Item response simulation. The ML estimates from the first rescoring were taken as the best estimates of proficiency of the flagged test events. These estimates were used as $\hat{\theta}$ to determine P, the probability of a correct response using $P = \frac{1}{1+\exp^{-(\hat{\theta}-\delta)}}$, where δ was the item difficulty calibration. The simulation of each response was completed by generating a random number $k = \{0,1\}$. If $k \leq P$, the item response was coded as correct, otherwise it was coded as incorrect.

Rescore 2 (Rs2). All targeted test events were rescored using the original responses ($\geq$ 3 seconds) and the simulated responses that replaced the original (< 3 second) responses to the remaining items. The focus for this procedure was on standard errors of the resulting scores.

Results

Test Events with Flags Triggered.

For the elementary level tests, approximately 5% of the test events across content areas triggered at least one flag. At the high school level, approximately 7% of the test events across content areas triggered at least one flag. Between content areas, reading test events from elementary students triggered about 2.5 times the percentage of flags as mathematics tests did. For high school students, the percentages of test events with flags triggered was more consistent (8.8% in reading; 7.6 % in mathematics). The patterns of these differences were repeated when considering the total number of flags triggered across tests. These differences are shown in Table 3.

Table 3.
Flags Triggered in Test Events by Subject and Grade

	Test Events		
Grade	Total	Tests with at Least One Flag	Total Flags Across All Tests
		Reading	
3	16209	1117	1391
		6.9%	8.6%
9	18705	1437	2737
		7.7%	14.6%
		Mathematics	
4	15532	375	549
		2.4%	3.5%
10	18718	1188	2491
		6.3%	13.3%

Individual flags triggered as a percentage of all test events with at least one flag triggered are provided in Table 4. This table also provides these percentages by the total number of flags triggered. The least triggered flag across all set was Flag 2, (Less than 30% of all items answered correctly). This is consistent with the expectations of an adaptive test and only occurred in conjunction with other flags being triggered. The most commonly triggered flag in the elementary test events was Flag 4, (No more than 20% of items correct in at least 20% of the subsets). These Flag 4 triggers seemed to have little relationship to Flag 2, the other “proportion-correct” flag. This suggests that the Flag 4 tended to be triggered in rolling subsets that, if not contiguous, were at least in close proximity within the test event. The remaining flags related to response latency (Flags 1, 3 and 5) for elementary test events formed a pattern only with respect to the order of magnitude of their percentages in each set of test events.

Table 4.
Percentages of Flagged Test Events by Content Area by Flag and by Total Flags Triggered

Reading

		Grade 3 (n = 1391)						Grade 9 (n = 2737)					
		Total Flags Triggered						Total Flags Triggered					
Flag	Description	1	2	3	4	5	Total	1	2	3	4	5	Total
1	Latency < 3 seconds to at least 15% of all items	0.4	2.4	2.1	2.5	0.1	7.5	0.5	6.5	6.8	6.7	0.6	21.1
2	Less than 30% of all items answered correctly.	0.0	2.3	0.1	0.1	0.1	2.6	0.0	0.4	0.1	0.1	0.6	1.2
3	No more than 20% of items correct AND response latency < 3 seconds to at least 3 items in any subset.	0.7	1.2	2.7	2.4	0.1	7.2	1.1	2.3	7.5	6.8	0.6	18.3
4	No more than 20% of items correct in at least 20% of the subsets.	65.6	2.9	1.2	2.5	0.1	72.2	23.4	1.2	1.2	6.8	0.6	33.2
5	Response latency < 3 on at least three items in 20% of the rolling subsets.	2.2	3.1	2.8	2.4	0.1	10.6	3.3	7.8	7.6	6.8	0.6	26.2
	Totals	68.8	11.9	8.8	10.1	0.4		28.2	18.2	23.1	27.3	3.1	

Mathematics

		Grade 4 (n = 549)						Grade 10 (n = 2491)					
1	Latency < 3 seconds to at least 15% of all items	0.2	1.0	3.1	5.0	0.2	9.5	0.2	3.3	7.4	6.9	1.2	18.9
2	Less than 30% of all items answered correctly.	0.0	0.3	0.0	0.0	0.2	0.5	0.0	0.4	0.0	0.0	1.2	1.7
3	No more than 20% of items correct AND response latency < 3 seconds to at least 3 items in any subset.	3.3	4.8	4.1	6.9	0.2	19.2	3.1	5.1	8.9	6.9	1.2	25.3
4	No more than 20% of items correct in at least 20% of the subsets.	42.3	1.0	1.0	6.9	0.2	51.4	15.3	1.8	1.7	6.8	1.2	26.9
5	Response latency < 3 on at least three items in 20% of the rolling subsets.	3.4	4.8	4.1	6.9	0.2	19.4	3.1	7.1	8.8	6.9	1.2	27.2
	Totals	49.1	12.0	12.4	25.6	0.9		21.8	17.7	26.9	27.5	6.2	

Unlike the elementary test events, the test events from high school showed no clearly dominant flag, overall. However, similar to the elementary test events, of those with a single flag triggered, Flag 4 (No more than 20% of items correct in at least 20% of the subsets) was the most common. As a percentage of flagged test events, those from the high school set had all five flags triggered about seven times as often as elementary test event in both reading and mathematics.

Table 5 provides the frequencies and percentages of pairs of triggered flags in each set of test events. This table is restricted to the same flagged test events that were presented in Table 4. In each test event set in Table 5, shaded cells along the diagonal contain the frequencies of test events that triggered the particular flag. Each cell above the diagonal contains the number of test events that triggered the pair of flags defining the cell. Similarly, each cell below the diagonal contains the percentage of all test events that triggered the pair of flags defining the cell. The shaded cells below the diagonal contain the

percentages the percentages of test events that triggered related flags (e.g., both criteria included response latency or both included proportion correct). It should be noted that all cells below the diagonal in each table will not sum to 100%; only *pairs* of triggered flags are being reported, but any test event could have triggered between one and five flags.

Table 5.
Frequencies and Percentages of Pairs of Flags Triggered in Test Events by Content Area and by Grade

Reading

		Grade 3						Grade 9					
		Flag						Flag					
Flag	**Description**	**1**	**2**	**3**	**4**	**5**	**Total**	**1**	**2**	**3**	**4**	**5**	**Total**
1	Latency < 3 seconds to at least 15% of all items	104	3	62	40	96		578	19	384	210	557	
2	Less than 30% of all items answered correctly.	0.2	36	3	36	2		0.7	33	20	31	22	
3	No more than 20% of items correct AND response latency < 3 seconds to at least 3 items in any subset.	4.5	0.2	100	55	79		14.0	0.7	500	251	444	
4	No more than 20% of items correct in at least 20% of the subsets.	2.9	2.6	4.0	1004	50		7.7	1.1	9.2	910	236	
5	Response latency < 3 on at least three items in 20% of the rolling subsets.	6.9	0.1	5.7	3.6	147		20.4	0.8	16.2	8.6	716	
							1391						2737

Mathematics

		Grade 4						Grade 10					
		Flag						Flag					
Flag	**Description**	**1**	**2**	**3**	**4**	**5**	**Total**	**1**	**2**	**3**	**4**	**5**	**Total**
1	Latency < 3 seconds to at least 15% of all items	55	1	49	30	53		472	33	387	204	462	
2	Less than 30% of all items answered correctly.	0.2	3	1	3	1		1.3	43	32	41	33	
3	No more than 20% of items correct AND response latency < 3 seconds to at least 3 items in any subset.	8.9	0.2	101	40	77		15.5	1.3	630	274	516	
4	No more than 20% of items correct in at least 20% of the subsets.	5.5	0.5	7.3	288	36		8.2	1.6	11.0	669	243	
5	Response latency < 3 on at least three items in 20% of the rolling subsets.	9.7	0.2	14.0	6.6	102		18.5	1.3	20.7	9.8	677	
							549						2491

A cursory inspection of Table 5 reveals that flag pairs with response latency criteria in both flags were triggered at a much higher rate than those with proportion correct criteria in both flags of the pair. Moreover, the flag pairs with proportion correct criteria appeared in similar percentage magnitudes as flag pairs with only a response latency criterion in one flag and a proportion correct criterion in the other. This would seem to run counter to the common assumption that rapid responses are more likely to result in incorrect answers.

Follow-up procedures.

As stated earlier, the follow-up procedures were used to estimate the impact of ignoring odd or unexpected test-test takers interactions that are present in a test event. They were limited to test events in which Flags 1, 3 and/or 5 had been triggered in any combination but neither Flag 2 nor 4 (the proportion-correct flags) were triggered. A first rescoring (Rs1) was carried out after eliminating the items that had been responded to in less than 3 seconds. Theta estimates from the first rescoring were used to simulate responses to excluded items which were subsequently incorporated into the second rescoring (Rs2).

Score and standard error differences from Rs1 are provided in Table 6 for reading by grade and by orthogonal flag set. The rescoring procedure generally resulted in somewhat higher scores. There was a weak tendency for greater differences to be associated with the number of flags included in the flag set. As expected, standard error differences followed a pattern similar to the differences in scores, with the standard errors from the rescore being somewhat higher – roughly 4% to 27%. The corresponding results for mathematics are comparable in magnitude and pattern (see Table 7).

Table 6.
Differences in Scores and Standard Errors Following Rescore 1 for Reading -- (Rescore 1 estimate - Original estimate)

	Orthogonal		Score Differences			SE Differences		
Grade	Flag Sets	N	Mean	SD	Med	Mean	SD	Med
3	1	5	1.9	0.97	2	0.31	0.05	0
	3	10	2.6	1.23	3	0.13	0.04	0
	5	30	1.3	1.47	1	0.20	0.06	0
	1, 3	1	4.1	--	4	0.27	--	0
	1, 5	33	2.6	2.25	2	0.47	0.24	0
	3, 5	9	2.6	1.00	2	0.20	0.05	0
	1, 3 & 5	25	3.5	3.26	3	0.64	0.39	1
9	1	13	1.9	1.61	2	0.34	0.09	0
	3	30	2.2	1.37	2	0.15	0.06	0
	5	90	0.9	1.53	1	0.20	0.07	0
	1, 3	4	2.8	1.40	3	0.30	0.04	0
	1, 5	173	3.2	3.46	3	0.70	0.80	0
	3, 5	39	2.1	1.20	2	0.20	0.06	0
	1, 3 & 5	175	5.1	4.63	5	0.87	1.06	1

Table 7.
Differences in Scores and Standard Errors Following Rescore 1 for Mathematics -- (Rescore 1 estimate - Original estimate)

Grade	Orthogonal Flag Sets	N	Score Differences Mean	Score Differences SD	Score Differences Med	SE Differences Mean	SE Differences SD	SE Differences Med
4	1	1	-0.4	--	0	0.27	--	0
	3	19	2.0	0.87	2	0.10	0.04	0
	5	20	1.5	1.13	2	0.16	0.05	0
	1, 3	1	2.3	--	2	0.26	--	0
	1, 5	5	4.2	2.37	4	0.39	0.08	0
	3, 5	23	2.5	1.14	2	0.16	0.06	0
	1, 3 & 5	18	5.3	2.85	6	0.54	0.25	0
10	1	5	1.0	1.35	1	0.28	0.01	0
	3	78	2.1	1.25	2	0.13	0.06	0
	5	78	1.2	1.29	1	0.19	0.06	0
	1, 3	2	3.7	1.00	4	0.27	0.02	0
	1, 5	79	3.9	3.93	3	0.52	0.35	0
	3, 5	95	2.1	1.54	2	0.19	0.07	0
	1, 3 & 5	180	5.3	3.98	5	0.61	0.57	0

The score and standard error differences reported in Tables 6 and 7 are depicted graphically in Figure 1 as a function of the number of responses given in less than 3 seconds (rapid responses). These panels make clear the relationships between the score and standard error differences provided by the Rs1 procedure and the original scoring; except in rare cases, score differences are positive and are accompanied by positive SEM differences. In general, SEM differences were under .5 RITs (.05 logits) when the number rapid responses was below 8 for reading test events and was below 11 for mathematics test events.

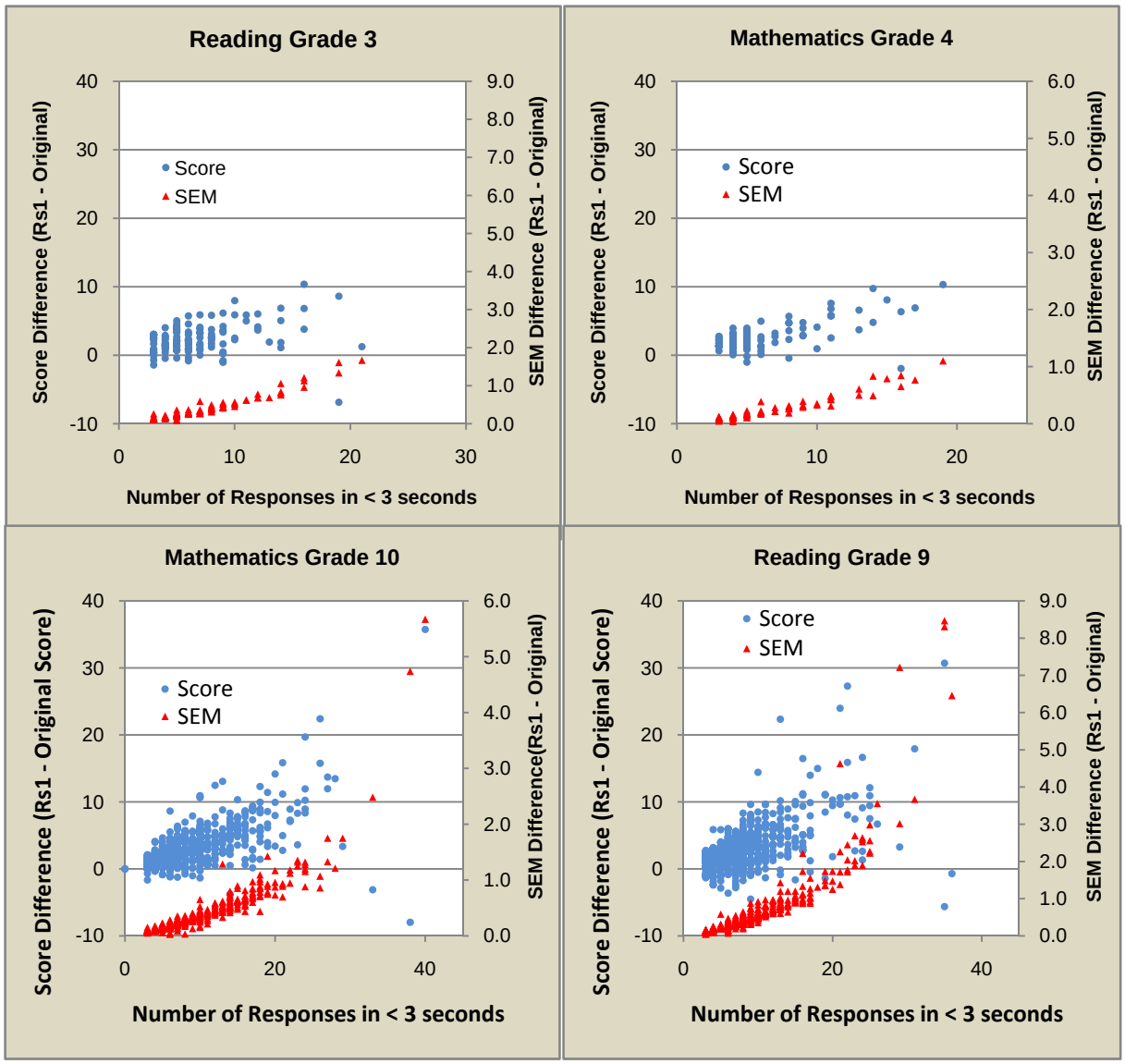

Figure 1. Score and SEM differences between Rs1 and original scoring.

The standard error differences from Rs2 are shown in Tables 8 and 9 for reading and mathematics, respectively. In these tables the standard error differences between the original scoring and Rs1 are repeated for reference. It can be seen that in both tables that the mean differences between the standard errors estimated by Rs2 and the original standard errors were usually quite small. However, the 1,5 and 1,3,5 flag sets were notable across all comparisons in that they were always larger and were accompanied by relatively sizable variance.

Table 8.
Differences in Standard Errors Following Rescore 1 and Rescore 2 for Reading -- (Rescore estimate - Original estimate)

Grade	Orthogonal Flag Sets	N	SE Differences (Rescore 1 - Original) Mean	SD	Med	SE Differences 2 (Rescore 2 - Original) Mean	SD	Med
3	1	5	0.31	0.05	0	-0.02	0.07	0
	3	10	0.13	0.04	0	-0.01	0.04	0
	5	30	0.20	0.06	0	0.01	0.03	0
	1, 3	1	0.27	--	0	-0.04	--	0
	1, 5	33	0.47	0.24	0	0.02	0.07	0
	3, 5	9	0.20	0.05	0	-0.01	0.02	0
	1, 3 & 5	25	0.64	0.39	1	0.04	0.13	0
9	1	13	0.34	0.09	0	0.01	0.02	0
	3	30	0.15	0.06	0	-0.02	0.04	0
	5	90	0.20	0.07	0	0.01	0.04	0
	1, 3	4	0.30	0.04	0	0.00	0.02	0
	1, 5	173	0.70	0.80	0	0.07	0.14	0
	3, 5	39	0.20	0.06	0	0.00	0.03	0
	1, 3 & 5	175	0.87	1.06	1	0.08	0.26	0

Table 9.
Differences in Standard Errors Following Rescore 1 and Rescore 2 for Mathematics -- (Rescore estimate - Original estimate)

Grade	Orthogonal Flag Sets	N	SE Differences (Rescore 1 - Original) Mean	SD	Med	SE Differences 2 (Rescore 2 - Original) Mean	SD	Med
4	1	1	0.27	--	0	0.00	--	0
	3	19	0.10	0.04	0	-0.01	0.02	0
	5	20	0.16	0.05	0	0.01	0.01	0
	1, 3	1	0.26	--	0	-0.01	--	
	1, 5	5	0.39	0.08	0	0.04	0.08	0
	3, 5	23	0.16	0.06	0	0.01	0.05	0
	1, 3 & 5	18	0.54	0.25	0	0.05	0.09	0
10	1	5	0.28	0.01	0	0.01	0.02	0
	3	78	0.13	0.06	0	-0.01	0.03	0
	5	78	0.19	0.06	0	0.01	0.02	0
	1, 3	2	0.27	0.02	0	0.02	0.03	0
	1, 5	79	0.52	0.35	0	0.04	0.10	0
	3, 5	95	0.19	0.07	0	0.01	0.04	0
	1, 3 & 5	180	0.61	0.57	0	0.05	0.20	0

Test information recovery. Another perspective on the standard error differences from rescoring is shown in Figures 2 through 5 in the terms of the rescore procedures' effects on test information. Test events were included for these figures if Flags 1, 3 and/or 5 had been triggered. Unlike the data sets used for Tables 6 through 9, however, those used for these charts did include test events that had also triggered one or both of the proportion-correct flags; i.e., Flags 2 and 4. Figures 2 and 3 are for reading in grades 3 and 9, respectively. These two figures illustrate the decay that would result if the standard errors from Rs1 were to be used to estimate test information. Since tests are increasingly shorter moving from left to right in the chart, the effects on test information are expected. However, given a shorter rescored test, it is still reasonable to ask what the level of test information was likely to have been had the test taker responded to each test item in a manner that was consistent with the measurement model. This was the purpose of the simulated responses to previously eliminated rapid response items. In both figures it can be seen that there is a point at which the test information estimated from Rs2 clearly diverges from the estimates from the original test event. These points are after items 14 and 12 for grades 3 and 9 (Figures 2 and 3), respectively. For grade 3 (Figure 2), this results in test information estimates comparable to estimates from the original tests for 94% of the test events. For grade 9 (Figure 3), only about 73% of Rs2 procedure yielded test information estimates that were comparable to the original test event. It should be noted that these test information estimates, even on the original test events, were always somewhat below where test information is commonly observed for these tests (the horizontal bar at about 9).

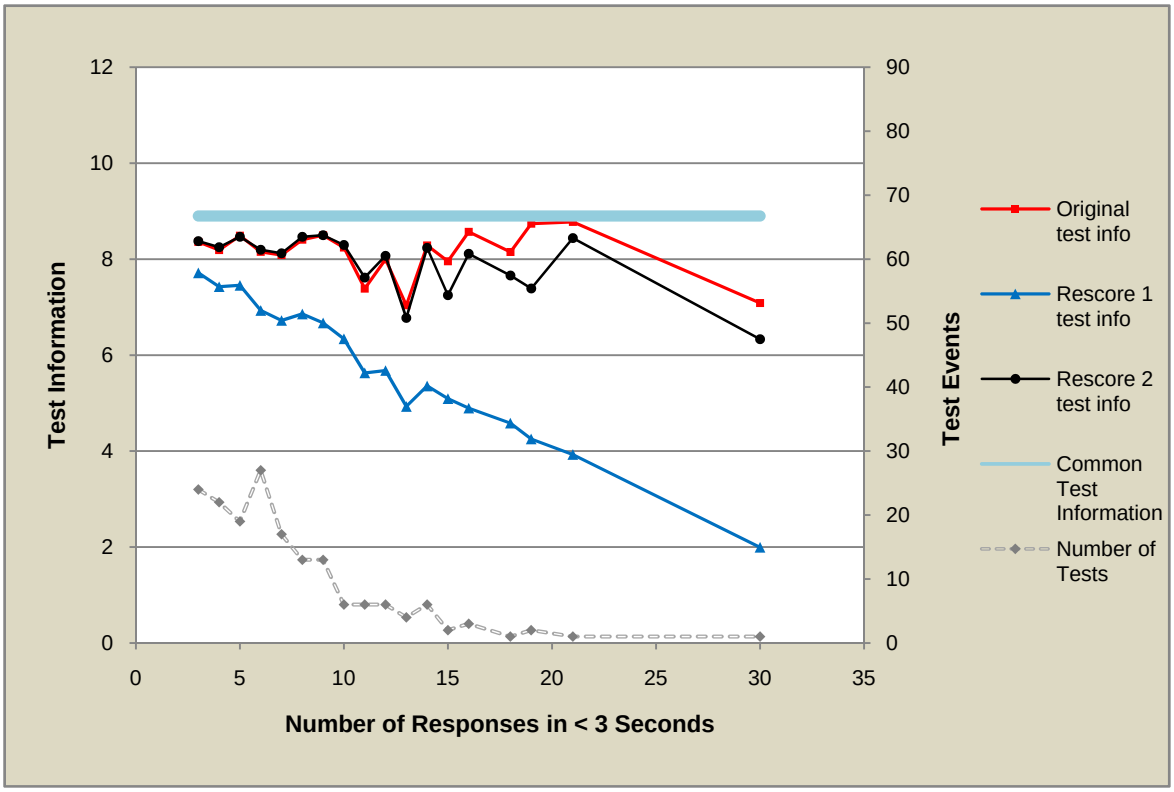

Figure 2. Test information before and after rescoring for grade 3 reading.

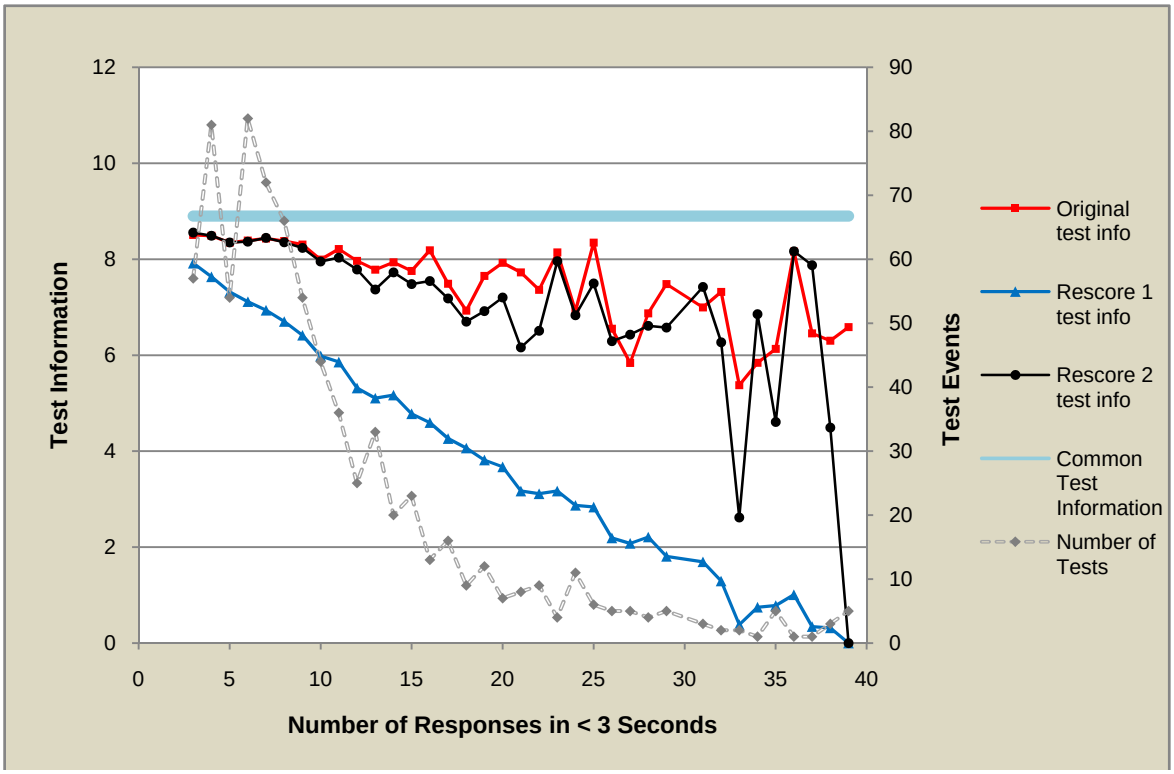

Figure 3. Test information before and after rescoring for grade 9 reading.

It is interesting to note that grade 3 test takers were much less likely than grade 9 test takers to give rapid responses to more than one-fourth of the items. Only 19% of the grade 3 test events in this data set (2% of all grade 3 flagged test events) contained more than 10 rapid responses. In contrast, 35% of the grade 9 test events in this data set (10% of all grade 9 test flagged test events) contained more than 10 rapid responses.

The comparable results from mathematics are displayed in Figures 4 and 5 for grades 4 and 10, respectively. Once again there is a very similar pattern across content areas for the test events coming from elementary test takers. The same is true for high school level test takers. Similar to the results for reading, we see for both the grade 4 and grade 10 data, a point at around 12 or 13 rapid responses where the test information from the original test and that from Rs2 begin to diverge. The percentages of test events with more than 12 rapid responses were 19% for grade 4 and 33% for grade 10.

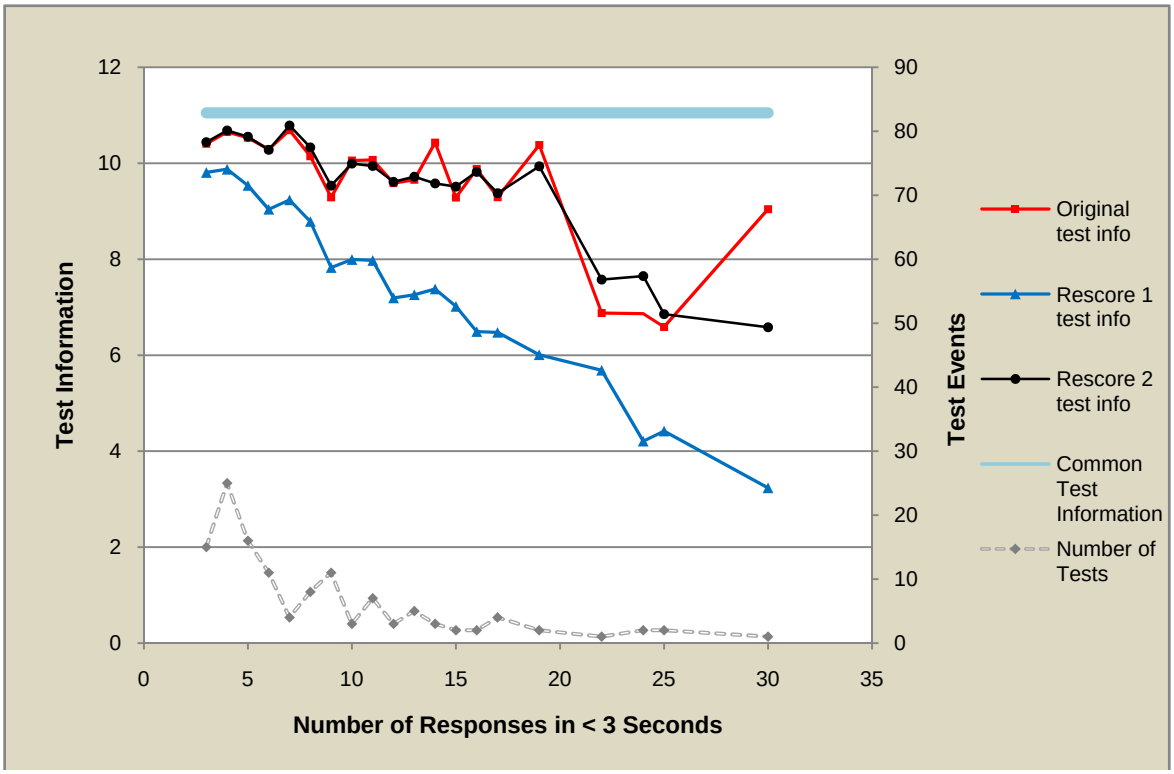

Figure 4. Test information before and after rescoring for grade 4 mathematics.

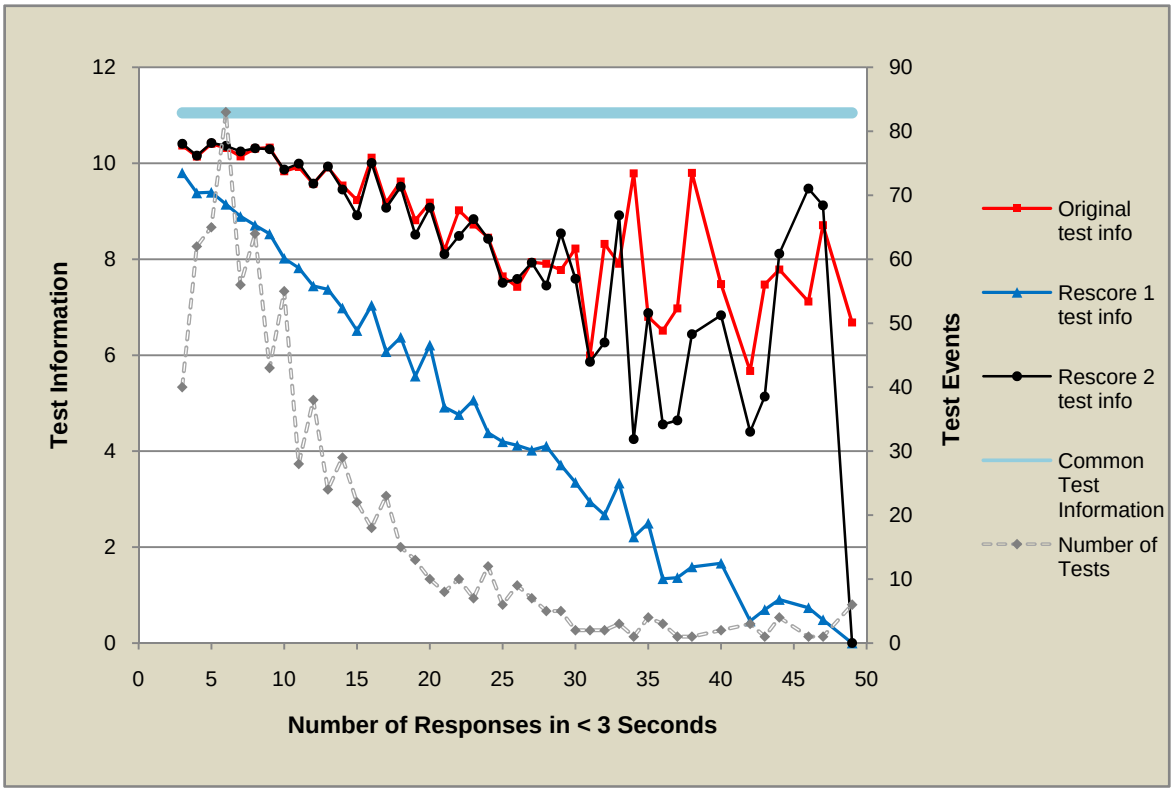

Figure 5. Test information before and after rescoring for grade 10 mathematics.

Test information recovery and score distributions. Nearly all the test events flagged for the rescoring procedure (Flags 1, 3 and/or 5 triggered) were from the first three quartiles in the 2008 NWEA norms. When the various test information estimates from the original and the two rescores procedures are viewed in the context of the original proficiency estimates, a clear distinction can be seen between elementary and high school test takers. Figures 6 through 9 contain displays of the same information presented in Figures 2 through 5 but presented as a function of the original proficiency estimates rather than as a function of the number of rapid response items. For purposes for the displays, scores well below the first percentile were omitted. This amounted to 10 and 5 test events being eliminated from the data sets for grades 3 and 4, respectively. For grades 9 and 10, 66 and 80 test events, respectively, were eliminated.

As a group, Figures 6 through 9 reveal that the vast majority of test events from all data sets were, on average, yielding less test information than those test events that had no flags set. The smallest differences appear at the upper end of these score distributions for high school test events, which are commonly in quartile 3. Across all data sets, there are few potions of the score distributions where the estimates of test information from Rs2 were consistently aligned with those from the original scoring. Where the Rs2 and original test information estimates did coincide, the scores tended to fall in quartile 2 and to some extent in quartile 3 in mathematics. Beyond these tendencies, there was little consistency between the test information estimates from the original test events and those from Rs2.

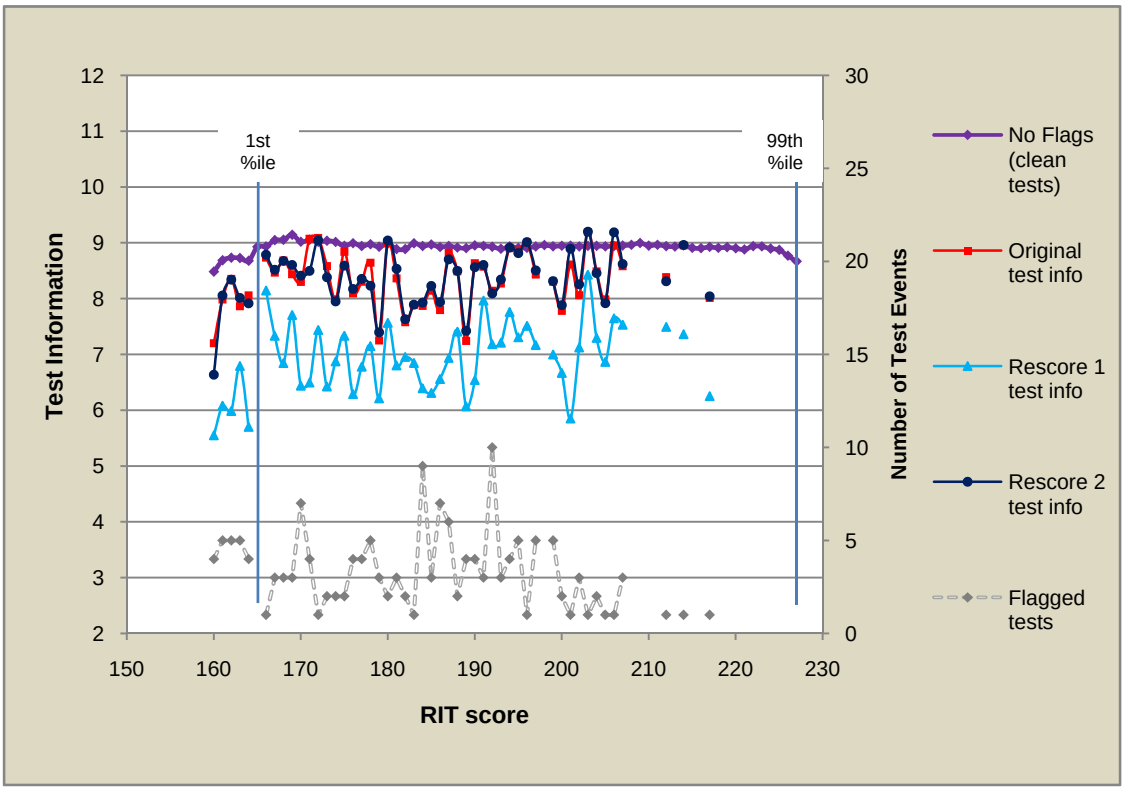

Figure 6. Test information by original proficiency score before and after rescoring for grade 3 reading.

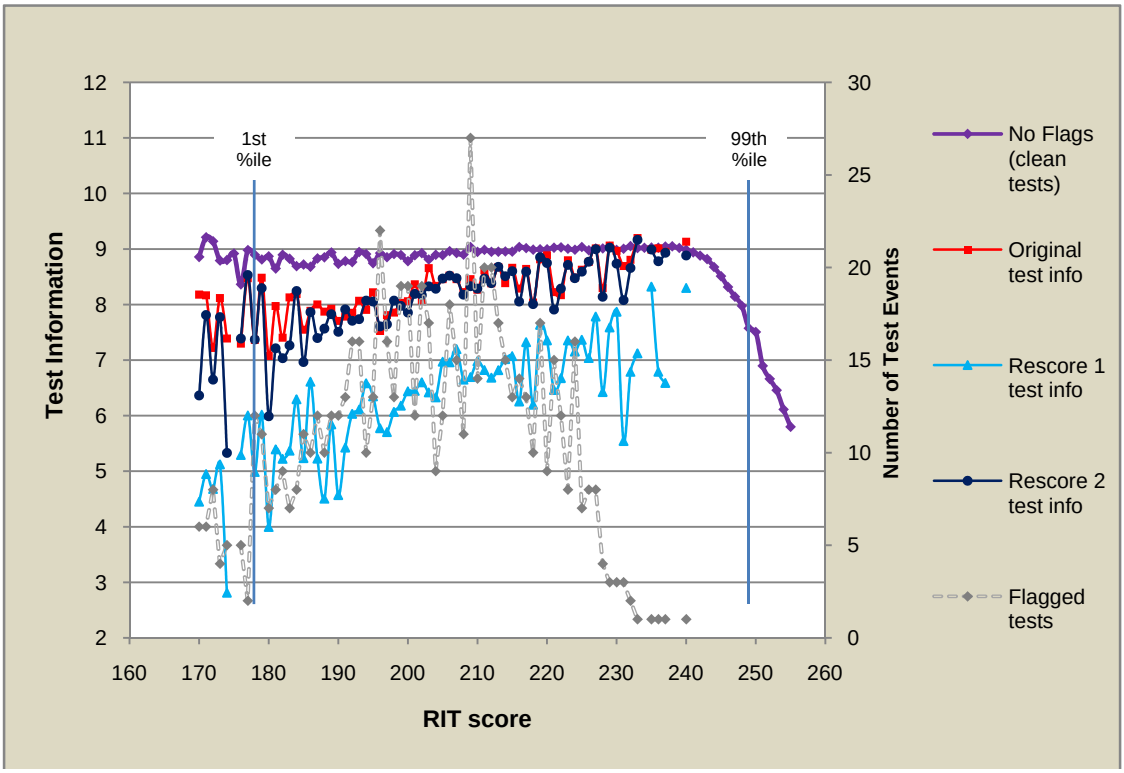

Figure 7. Test information by original proficiency score before and after rescoring for grade 9 reading.

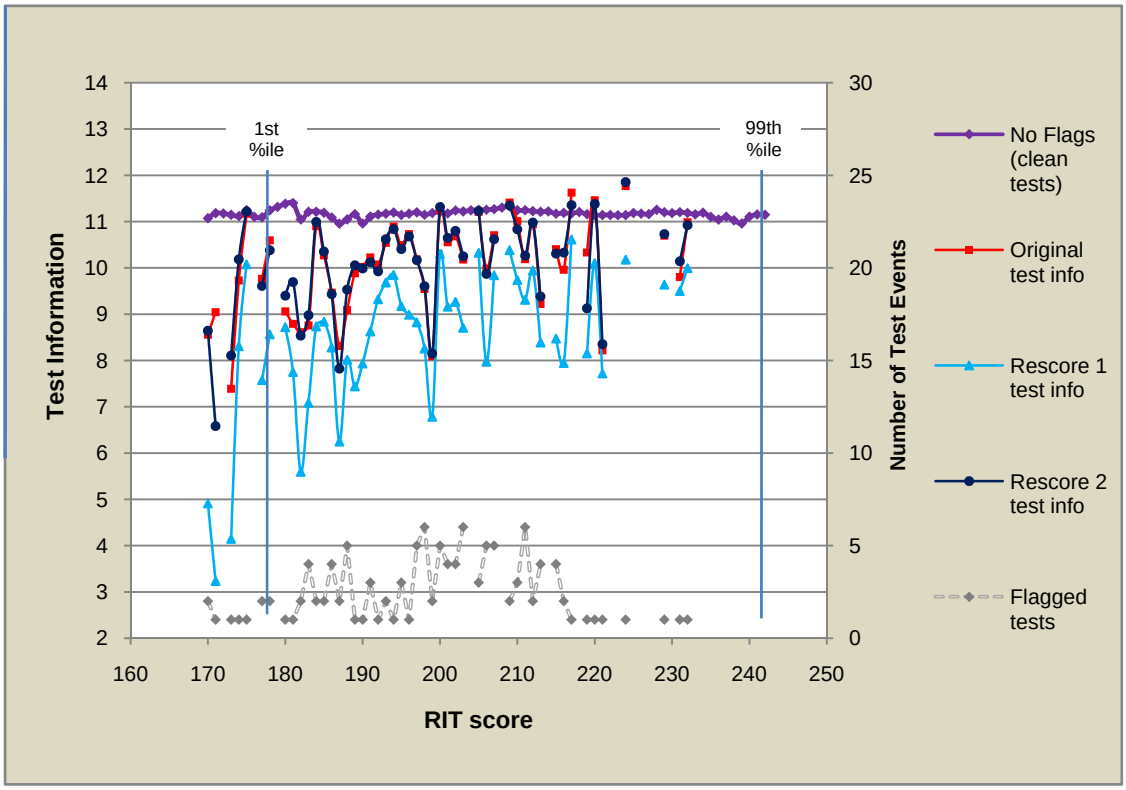

Figure 8. Test information by original proficiency score before and after rescoring for grade 4 mathematics.

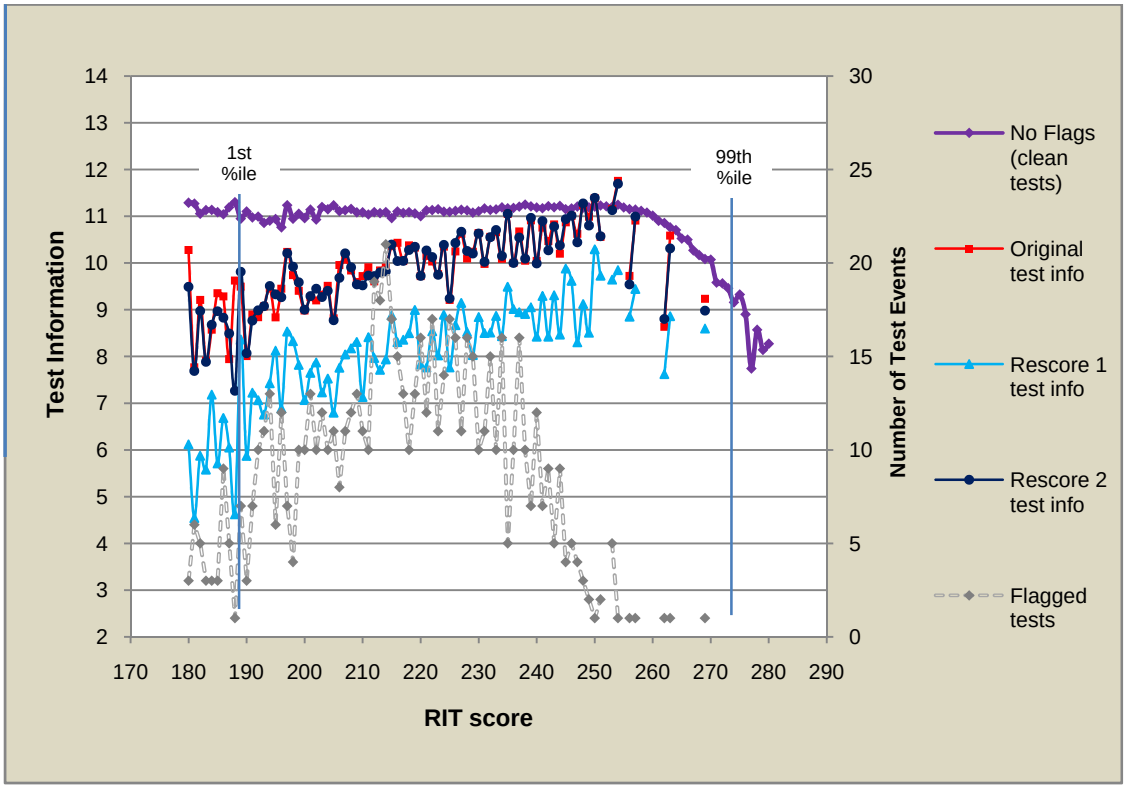

Figure 9. Test information by original proficiency score before and after rescoring for grade 10 mathematics.

Discussion

This study examined the immediate consequences of ignoring odd or unexpected test-test taker interactions evidenced within test events. A system of rule-based indicators focusing on response latency and proportion of items answered correctly was used to identify odd or unexpected test-test taker interactions. Such interactions are considered important in that they hold implications for the validity of the final proficiency estimates and the use of those estimates in making decisions at the individual test taker level. Applying the system of indicators resulted in familiar patterns of test-test taker interactions reported or suggested in other studies (e.g., Cronin, 2006; Hauser, Kingsbury & Wise, 2008; Kingsbury & Hauser, 2007). This included observing more incidences of short latency (less effort) among upper grade or high school students than among younger elementary students.

These observations include students who answered test questions too quickly to apply what they know of the content area. Therefore, the standard errors of these scores are smaller than they should be, and the scores themselves tend to be lower than they should be, due to the inappropriate interaction between the students and the test questions.

To examine the consequences of ignoring these interactions, a sequence of test manipulations was used to rescore test events after removing responses given in less than 3 seconds. Proficiency estimates from

this step were compared to the original proficiency estimates. These estimates were also used to simulate test takers' responses to the omitted items under the assumption that if the test taker had been engaged with these items, their responses would have been consistent with the measurement model. Using these responses to complete the response vectors, the tests were rescored to estimate the test information that would have obtained had the test taker been fully involved during the test event.

Proficiency estimates from rescored tests were, on average, one to five RITs higher than the original estimates. The size of the difference was positively related to the number of rapid (less than 3 second) responses in the test event. As expected, however, there was a similar pattern of increases in the differences in standard errors as the number of rapid (removed) responses increased. These trends had an obvious deteriorating effect on the test information yielded from the second rescoring. It is important to note that a great preponderance of original test events that were used in the rescoring procedures was yielding less information than non-flagged events. The rescoring procedures did not change this. In the best case, Rs2 merely returned the test information estimate to the level of the original test.

It would appear from these results that the effect of ignoring rapid responses by test takers can be tolerated without adversely affecting proficiency estimates or their accompanying standard errors as long as the number of rapid responses does not exceed 20% of the test length. Test events falling into this category comprised between 65% and 85% of all test events with more than two rapid responses. When the percentage of rapid responses does not exceed 20% of the total test length, there may be benefit to rescoring the test event without the rapid response items. This should provide more valid proficiency estimates; standard error and therefore test information will be minimally affected. This is the more obvious conclusion from this study. However, there are several issues related to the study that merit attention. Each of these represents a limitation to the current study as well as an area that would benefit from additional study.

Rescoring after dropping rapid responses was used here in a post hoc fashion. While this procedure appears to be, up to a point, a reasonable method of obtaining more valid proficiency estimates, it may be more efficient and useful if incorporated into the testing procedure itself. This might be done by monitoring flags during the test event and modifying the event in situ or at the back end of the test. Such a procedure may also benefit from the addition of feedback to the test taker about the level of effort they are expending (Wise, Bhola & Yang, 2006).

Low proportion correct, in and of itself, may or may not be a consequence of low effort – it may simply be an artifact of a mismatch between test taker proficiency and item difficulty. For this reason low proportion correct should be treated cautiously when used as an indicator for test score validity. We know, for example, that it is extremely unusual for a test taker to incorrectly answer less than 40% of the items on these adaptive tests (or to correctly answer more the 60% of the items). The test score may be invalid but not for reasons of effort; it may be that the test was not well targeted to the test taker. This possibility can be ruled out by examining previous test events in the same content domain from the same test taker.

The use of response latency in studying individual score validity would benefit from further consideration of how latency can be used to define an inappropriate (low effort, disengaged) response. The less than 3 second definition used here was a best approximation based on the work of Wise and Demars (2006). A response in less than 3 seconds to almost any item certainly seems defensible as a low effort or disengaged response. Nevertheless, 3 seconds as a constant threshold may be too limiting. It also makes an implicit assumption that non-rapid responses (more than 3 seconds) represent engagement or full effort with a test item. This may not be warranted. For example, it seems that if the mean response time for an item is 96 seconds, a response made after 10 seconds would be considered as too rapid and thus, low effort. Alternatively, for "identification" type items where the average response latency is, say, 7 seconds, a response given in 2 seconds may not be reflecting low effort. Response latency considered conjointly with an estimate of the test taker's speed and the time demand of the item estimated from similar test takers may be more revealing. Van der Linden (2006) demonstrated how this could be addressed using a lognormal model to estimate these parameters. Even with such an approach, however, the problem of identifying appropriate thresholds to make decisions about individual score validity will still be with us.

It is clear that in this moderate-stakes setting, a small percentage of students display low levels of effort that can be identified through a series of consistently applied rules. It is also clear that these students are receiving achievement level estimates that are inappropriate and standard errors that are inaccurate. Two rescoring approaches have been demonstrated that identify the magnitude of the errors that we are making by using the initial scores as if they were valid. While a variety of approaches might be used to address issues of low effort, it is clear that not addressing the issues can result in inaccurate information for individual students. If we plan to use our tests for making instructional decisions about individual students, issues of low effort must be considered and addressed.

References

Anastasi, A. (1976). *Psychological testing*. (4th ed.). New York: McMillan.

Cronin, J. (April, 2006). The affect of test stakes on item response time, response accuracy, and growth as measured on a computerized-adaptive test. Paper presented to the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Hauser, C., Kingsbury, G. G., & Wise, S. L. (March, 2008). Individual validity: Adding a Missing Link. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kingsbury, G. G. & Hauser, C. (April, 2007). Individual validity in the context of an adaptive test. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of validity. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Meaning and values in test validation. *Educational Researcher, 18*(2), 5-11.

Moss, P. A., Girard, B. J. & Haniford, L. C. (2006). Validity in educational assessment. In J. Green & A. Luke (Eds.), *Review of Research in Education, 30*, 109-162.

Schnipke, D. L. & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213-232.

Schnipke, D. L. & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analysis. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 19*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics 31*(2), 181-204.

Van der Linden, W. J. & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika, 68,* 251-265.

Wise, S. L., Bhola, D. S. & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice* 25(2), 21-30.

Wise, S. L. & DeMars, C. E. (2005). Low test taker effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment 10*(1), 1-17.

Wise, S. L. & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement 43*(1), 19-38.

Wise S. L. & Kingsbury, G. G. (2006). An investigation of item response time distributions as indicators of compromised NCLEX item pools. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wise, S. L., Kingsbury, G. G. & Hauser, C. (April 2009). A generalized framework for identifying individual score validity (ISV) in a variety of testing settings. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.