4-2009

Individual score validity and student effort in higher education assessment

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Steven L. Wise

Lisa F. Smith

Running head: ISV AND STUDENT EFFORT

Individual Score Validity and Student Effort in Higher Education Assessment

Christine DeMars

James Madison University

Steven L. Wise

Northwest Evaluation Association

Lisa F. Smith

University of Otago

Paper presented at the annual meeting of the National Council on Measurement in Education,

San Diego, California, April 2009.

Abstract

This study explored the use of the five invalidity flags plus a new sixth flag based on self-reported effort. Participants were 155 entering first-year university students who were measured during an orientation week and again 18 months later. The instruments were a faculty-developed test of oral communications skills with 40 four-option multiple-choice items and a self-reported measure of test-taking motivation (Student Opinion Survey; Sundre, 1999 adapted from Wolf and Smith, 1995). Results indicated that the Flags explored in this study generalized well to university students. There was a moderate correlation between Response Time Effort and Effort as measured by the Student Opinion Scale, suggesting there was a relationship not captured by the dichotomized flags.

ISV and Student Effort in Higher Education Assessment

As described in the other papers in this symposium, some examinees exhibit very low effort in low-stakes testing situations. Test scores for these examinees thus are not valid indicators of what they know and can do. Two methods of identifying very-low-effort examinees have been described in the introductory paper in this symposium (Wise, Kingsbury, & Houser, 2009): response time effort (RTE) and accuracy. These methods are applied in the current paper as well. Additionally, self-reported effort was explored as a third basis for flagging low effort.

The context of this study differs from the other papers in this session in terms of the examinee population and the circumstances surrounding the test administration. The examinees were university students, not K-12 students or licensure applicants. Further, the context in which these tests were administered and in which the scores were used differed from the other studies in the session. The test scores were used only for program assessment, not for any individual accountability. Students were asked to cooperate so that the university could have accurate data for statewide reports and for planning changes to curricula, but students knew that the scores would not affect their grades or appear on their transcripts. Examinees first took the test as incoming first-year students, as part of orientation the weekend before classes began. They were then re-tested 18 months later on a day when classes were cancelled for assessment activities. Incoming students are generally much more positive and cooperative about the assessment process, and in a previous analysis of this data we found far less rapid-guessing by incoming students (Wise & DeMars, 2008). Thus, it was anticipated that any severely low effort would occur at the second time point, leading to unexpectedly low growth. This is in contrast to the Houser and Kingsbury study in this symposium where severely low effort could occur at either time point and lead to either unexpectedly high growth (low effort at time 1) or unexpectedly low growth.

The invalidity flags were described earlier in this session (Wise, Kingsbury, & Houser, 2009). In brief:

Flag 1: If the student gave rapid responses to at least 15% of the items (overall RTE ≤ .85). RTE, as described earlier in the symposium, is 1 - the proportion of items to which the examinee responded so quickly that there was no reasonable possibility that the item was read and processed. For the test items in this study, the threshold for determining whether was response was rapid-guessing or solution behavior was determined by visual examination of the response time plot for each item (Wise, 2006). A small but conspicuous number of students responded within a few seconds, then the response frequencies dropped off and gradually increased over time. The time immediately after the small peak was chosen as the threshold. For each of the items in this study, the threshold was either 3 or 4 seconds (Wise & DeMars, 2008).

Flag 2: If the student passed less than 30% of the items (overall accuracy ≤ .30).

Flag 3: If the student passed no more than two items (local accuracy ≤ .20) AND the student gave three or more rapid responses (local RTE ≤ .70) on any 10-item subset.

Flag 4: If the student exhibited low effort (local accuracy ≤ .20) on at least 20% of the rolling subsets.

Flag 5: If the student exhibited low RTE (local RTE ≤ .70) on at least 20% of the rolling subsets.

An additional flag was added for this study, based on self-reported effort on the test:

Flag 6: If the student reported a mean score ≤ 2 on a 5-point Effort scale.

A very low score (below the neutral point) was chosen as the criteria because, to be consistent with the other five flags, the purpose was to tag only the examinees giving extremely low effort.

The purpose of this study was to explore the use of the five invalidity flags, plus the new sixth flag based on self-reported effort. This last flag was added to this study because it would

have greater generalization to testing contexts where response time is not available, including paper and pencil tests.

Method

Participants

At James Madison University (JMU), entering first-year students participate in a three-hour assessment period during orientation (Time 1). Approximately 18 months later, these same students participate in another assessment session on a day when all classes are cancelled for assessment purposes (Time 2). Students are randomly assigned to various tests of general education topics. The participants in this study were 155 students who were given an Oral Communications (OC) test at Time 1 and Time 2. Each student completed the university's oral communication requirement by enrolling in one of three communications courses between Time 1 and Time 2. Of the 155 students, 150 were matched to self-reported motivation at time 1 and 139 were matched to self-reports at time 2. Students could not be matched if they provided an incorrect ID on the motivation scale or if they left before the motivation scale was distributed after the tests (not common because these students would have to return another day for a make-up testing session).

Instruments

The Oral Communications (OC) test is a faculty-developed test of oral communications skills. It has 40 four-option multiple-choice items. It is not designed to be narrowly tied to one particular course curriculum but is instead broadly aimed at concepts typically covered in university introductory communications courses. Time 1 students would also be expected to have had some exposure to these concepts in high school English and communications courses, though not at the depth of university courses.

The Student Opinion Survey (SOS) is a self-reported measure of test-taking motivation. Sundre (1999) based this instrument on Wolf and Smith's 8-item survey (1995); one item was deleted, another item was re-worded, and three additional items were added. It was administered after examinees completed the OC test and other tests. Examinees responded to these 10 items on a 5-point scale (Strongly Disagree to Strongly Agree). Five items asked how hard the examinee tried on the tests (Effort) and five items asked about how important the examinee thought the tests were (Importance). The Effort score used in this study was the average of the five Effort items. Averaging the items rather than summing them puts them on the same 5-point scale as the individual items, making interpretation easier. In this sample, the coefficient alpha reliability was .81 for the Effort scale and .73 for the Importance scale at Time 1 and .85 and .84 at Time 2.

Conceptually, self-reported Effort might be expected to tap into the same construct as RTE. Ratings of test Importance would be expected to have an indirect effect through decreasing effort. Thus, Flag 6, described in the introduction, was based on self-reported Effort rather than Importance. The self-reported Effort scale is very different from response time and accuracy because it would be expected to increase monotonically, if not linearly, with effort. Response time and accuracy are related to effort only at very low levels; beyond a threshold, response time and accuracy have little relationship to effort. Thus, a dichotomous indicator is a reasonable approach for response time and accuracy but is a waste of some of the information available in the self-reported Effort scale. In other contexts, where there is no need to create an index comparable to Flags 1 – 5, use of scores on the full scale is recommended.

Results

Identification of Low-Effort Students

Mean RTE was .999 (*SD* = .009) at Time 1 and .950 (*SD* = .163) at Time 2. In other words, there was virtually no rapid-guessing at Time 1, but 5% of the responses at Time 2 were

rapid-guesses. Table 1 shows the number (and percentage) of students flagged by each indicator. At Time 1, only two students were identified by any of the flags: one by Flag 4 (local accuracy) and one by Flag 5 (local RTE). At Time 2, 15 students were identified by one or more of the flags, with much overlap among the groups marked by different flags. Most (13) of these students were identified by both Flag 1 (overall RTE) and Flag 5 (local RTE), sometimes in combination with the accuracy flags. One student was tagged by Flags 1 and 2 (but not Flag 5), and another was selected only by Flag 5. Of the two students flagged at Time 1, one was not identified as severely low effort at Time 2, but the other was marked by Flags 1, 3, and 5 at Time 2.

On the self-reported motivation scale examinees reported similar effort at Time 1 (mean = 3.30, *SD* = 0.65) and Time 2 (mean = 3.31, *SD* = 0.70)[1]. These means are just above the middle point, leaning toward *Agree*. Correlations among self-reported effort, RTE, and test performance are displayed in Table 2. The test was scored using maximum likelihood estimation and a one-parameter logistic (1PL) model, discussed in more depth in the next section. At Time 1, correlations were negligible, likely due to almost no variance in RTE. At Time 2, self-reported scores on the Effort scale had a moderate correlation with RTE and test performance, but RTE had a stronger correlation with test performance. Given the RTE and self-report correlation, Flag 6 might be expected to overlap with the other flags. However, using the Effort scale to create Flag 6 was more problematic. As noted, mean Effort scores were virtually the same at Time 1 and Time 2. On the surface, this differs from Flags 1 – 5, which tagged far more Time 2

[1] In contrast to Effort ratings, students rated the importance of the test much higher at time 1 (mean = 3.28, sd = 0.59) than at Time 2 (mean = 2.97, sd = 0.71), a difference of 0.48 standard deviation units. However, test performance at Time 2 was more highly correlated with Effort ratings ($r = .42$, $p < .001$) than with Importance ratings ($r = .27$, $p < .001$). At Time 1, neither Effort nor Importance ratings were associated with test scores ($r = .12$, $p = .133$ for Effort and $r = .03$, $p = .744$ for Importance). We have seen a similar pattern in many other data sets: Compared to Time 1 students, Time 2 students have lower Importance scores but approximately equal Effort scores, and Effort is a stronger predictor of test performance.

examinees for low effort. Possibly there could still be more Time 2 examinees in the extreme tail of the distribution of Effort scores. This would yield a greater number of flagged examinees at Time 2 than Time 1, even though the means on the total continuum did not differ very much. However, this was **not** the case. At both Time 1 and Time 2, four different examinees were flagged for very low effort using Flag 6 (mean self-reported effort ≤ 2). The four examinees flagged at Time 1 did not include the two examinees targeted by Flags 4 and 5 (though two of the four were later tagged by other flags at Time 2). Of the four examinees flagged at Time 2, two were not identified by any other flags, one was tagged by all five of the other flags, and one was marked by Flags 1 and 5. Increasing the cutoff to ≤ 2.5 increased the number of flagged examinees at Time 2 to 20, 8 of whom were marked by at least one of the other flags. However, it also increased the number of flagged examinees at Time 1 to 21, of whom only 1 was targeted by the other flags. Overall, Flag 6 had little overlap with Flags 1 – 5. Because of these results, the effects of removing those tagged by Flag 6 were not examined in the next section.

Test Scores

Item parameters were calibrated with a 1PL model, with the mean item difficulty centered at zero and the discrimination fixed to one, using marginal maximum likelihood estimation in Conquest (1998). After the items were calibrated, IRT test scores (θs) were estimated through maximum likelihood. Test score reliability was relatively low in both groups. The marginal reliability (reliability of person separation) based on the θs was .67 at Time 1 and .77 at Time 2. The mean standard error of measurement was lower at Time 1 than at Time 2, but the scores were more homogeneous at Time 1 as well, leading to lower reliability of person separation. Some of the increased variance at Time 2 may have been due to a few very low scores related to very low effort, artificially increasing the reliability, as discussed in Wise (2006) and Wise and DeMars (2006).

The mean and standard deviation of the IRT scores are reported in Table 3. At both time points, the average person measure was above the average item difficulty of zero. Mean growth was about .37 standard deviation units using the pooled standard deviation or .45 using only the standard deviation from Time 1.

Figures 1 to 5 show scatterplots of Time 1 and Time 2 scores. In Figure 1, the examinees tagged by Flag 1 at Time 2 are indicated with a plus sign. The regression line, estimated from the scores of examinees not tagged by Flag 1, is also plotted. Similar data is plotted in Figures 2 – 4 for the other flags. For Flags 2 (overall accuracy), 3 (accuracy and RTE), and 4 (local accuracy), all of the flagged examinees clearly had the largest negative residuals. For Flags 1 (overall RTE) and 5 (local RTE), most of the tagged examinees had negative residuals, though two or three examinees fell on or near the regression line. All of the most negative residuals were flagged, though not all of the examinees flagged were among those with the most negative residuals. If large negative residuals were regarded as a criterion, one could say that Flags 1 and 5 had high sensitivity but moderate specificity. Flags 2, 3, and 4 tagged fewer examinees, but those flagged clearly had the most negative residuals.

The two examinees tagged at Time 1 are not labeled in the Figures. One of them was also tagged at Time 2; this examinee, like others tagged at Time 2, had a negative residual. The other examinee flagged at Time 1 was flagged for low local accuracy. This examinee's growth was somewhat higher than predicted but his Time 2 score remained lower than average. The low accuracy of this student at Time 1 may have been due to low proficiency in the tested concepts rather than extremely low effort.

Based on these figures, the correlation between Time 1 and Time 2 scores was expected to increase if the flagged examinees were removed. Table 4 provides the correlations based on the examinees who were **not** flagged. These correlations are based mainly on the same

examinees, with slight variations on which examinees were removed. Use of any of the flags increased the correlation by a meaningful amount. Flag 4 is most remarkable in that the removal of only 6 students yielded an increase of .10 in the correlation.

Also based on these figures, the flagged examinees had lower than average growth from Time 1 to Time 2. Therefore, removing the flagged examinees was expected to increase the mean growth. Table 5 shows the mean scores after removing flagged examinees. At Time 2, the standard deviation decreased and the mean score increased with the removal of flagged examinees. With this change in standard deviation, Cohen's *d* (difference in means divided by pooled standard deviation) would not be comparable across the different contrasts. Instead, standardized growth was computed as the mean growth divided by the standard deviation at Time 1 using all students. The mean estimated growth increased by 33% -50% when the flagged students were removed.

Discussion

Each of the five Flags selected 3 – 15 examinees (2%-9%) at Time 2, but no more than 1 examinee at Time 1. Removing the flagged examinees increased the correlation between scores at Time 1 and Time 2, and increased the estimated mean growth. By removing very low effort examinees, faculty could get a better estimate of the change in student knowledge.

The flags were intended only to identify examinees with extremely low effort. As such, the resulting growth estimates were likely still conservative. No adjustments were made for examinees who put forth enough effort to avoid the flags but did not give enough effort to demonstrate what they really knew.

Fewer students were flagged for accuracy (Flags 2, 3, and 4) than for RTE. To meet the low RTE criteria, an examinee only had to respond before the time threshold on 15% of the total items, or 30% of the items in a substring (usually both). But to meet the low accuracy criteria, an

examinee of moderate or high ability would have to answer almost all items (overall or in a local subset) randomly to obtain a score as low as 30% overall or 20% in a subset. Conversely, if there had been any very low ability examinees in the sample, they could have met this criteria without responding randomly. If the flagging criteria were raised, more students who were genuinely low in oral communications knowledge could have been mistakenly flagged. The other tests examined for this symposium were computer adaptive, so this problem was not encountered. On adaptive tests, examinees who exerted reasonable effort would be expected to answer around 50% correctly. Regardless of ability, the flag is activated if the percent correct drops from the expected 50% to 20% for a local subset. In the Oral Communications test, in contrast, for a student at the mean of the ability distribution the percent correct must drop from a mean of 75%, and the necessary decrease is higher for high ability examinees and lower for low ability examinees.

The accuracy flags raise another concern in fixed item tests: by removing the least accurate examinees, are we simply removing the examinees with the least knowledge? Even if this were the case, the mean growth would not be expected to increase—both Time 1 and Time 2 scores were removed for each examinee whenever the examinee met the Time 2 flagging criterion. If low-knowledge examinees were especially likely to be targeted by the accuracy flags, the mean Time 1 score would drop proportionally with the mean Time 2 score, leaving the growth rate constant. An examination of Table 5 reveals that the mean Time 1 scores were essentially unchanged regardless of which examinees were dropped from the analysis.

The self-reported Effort scale was included in this study because it can be used where response time is not available. Unfortunately, it did not flag the same examinees. There was a moderate correlation between RTE and Effort, suggesting there was a relationship not captured by the dichotomized flags. Another alternative would be to use Effort as a covariate. The

problem with this is the heterogeneity of slopes; Effort was less correlated with test scores at Time 1 than Time 2. Even if this were ignored, there was little difference in self-reported Effort at Time 1 and Time 2, so the adjusted mean test scores would be the same as the unadjusted means. The lack of difference in self-reported Effort across the two time points, despite the difference in RTE, also calls into question whether the respondent interpretation of the Effort scale changed between Time 1 and Time 2. Respondent interpretations of their own effort or the items on the scale may also explain why the correlation between RTE and Effort was not higher. For the purposes of individual score validity, what we want to know is whether examinees exerted enough effort to validly demonstrate what they know or can do. But examinees do not necessarily answer that question. A cooperative but knowledgeable examinee may think "This test was pretty easy, so I did not have to give much effort". This examinee would have high RTE but a somewhat low Effort score. Other examinees may downplay their self-reports because they fear they did not score very well and wish to attribute their performance to lack of effort rather than lack of ability, an ego-defensive attribution (Pintrich & Schunk, 2002, chap. 3). Still others may try to give socially-desirable responses. All of these things may combine to explain why RTE and Effort did not have a higher correlation and why RTE had a higher correlation with test performance than Effort did.

Conclusion

The Flags explored in this study generalized well to university students, in that: (1) they identified only a small proportion of students; (2) the identified examinees did not have disproportionately low knowledge at Time 1; (3) correlations between Time 1 and Time 2 scores increased when the flagged examinees were not included; and (4) estimated mean growth increased when the flagged examinees were not included. While it is impossible to know with real data whether the "right" examinees were flagged, and Figures 1 and 5 suggest several

examinees with nearly normal growth were tagged by Flags 1 and 5, this general pattern is encouraging. Taken as a whole, this pattern suggests that the growth rates estimated without the flagged examinees are likely more accurate.

References

Houser, C., & Kingsbury, G. G. (2009, April). ISV in a Modest-Stakes Adaptive Educational Testing Setting. In G. G. Kingsburg (Chair), *Individual Score Validity: How unexpected interaction between a test and a test taker influences the usefulness of a test score*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego.

Pintrich, P. R., & Schunk, D. H. (2002). Motivation in education: Theory, research, and applications (2nd ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.

Sundre, D.L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19,* 95-114.

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

Wise, S. L., & DeMars, C. E. (2008, March). Examinee non-effort and the validity of program assessment results. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Wise, S. L., Kingsbury, G. G., & Houser, C. (2009, April). A generalized framework for identifying Individual Score Validity (ISV) in a variety of testing settings. In G. G. Kingsburg (Chair), *Individual Score Validity: How unexpected interaction between a test and a test taker influences the usefulness of a test score*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego.

Wolf, L. F., and Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.

Table 1

Number of Examinees Flagged by Each Indicator (N = 155)

Type of Flag	Time 1	Time 2
Flag 1: Overall RTE ≤ .85	0	14
Flag 2: Overall Accuracy ≤ .30	0	3
Flag 3: Local accuracy ≤ .20 AND Local RTE ≤ .70 on any 10-item subset	0	8
Flag 4: Local accuracy ≤ .20 on at least 20% of the 10-item subsets	1	6
Flag 5: Local RTE ≤ .70 on at least 20% of the 10-item subsets	1	14

Table 2

Correlations among RTE, Effort, and Test Performance

	Time 1			Time 2	
	Effort	Test Score		Effort	Test Score
RTE	.057	.164	RTE	.350	.678
Effort		.123	Effort		.420

Table 3

Average IRT Scores (N = 155)

	Mean	SD
Time 1	1.07	0.54
Time 2	1.31	0.76

Table 4

Correlation between Time 1 and Time 2 Test Performance

	N	r
All Examinees	155	0.487
Criteria for Removal at Time 2		
Flag 1: Overall RTE ≤ .85	141	0.567
Flag 2: Overall Accuracy ≤ .30	152	0.567
Flag 3: Local accuracy ≤ .20 AND Local RTE ≤ .70 on any 10-item subset	147	0.565
Flag 4: Local accuracy ≤ .20 on at least 20% of the 10-item subsets	149	0.586
Flag 5: Local RTE ≤ .70 on at least 20% of the 10-item subsets	141	0.553
Any Flag	140	0.555

Table 5

Mean Rasch Scores at Time 1 and Time 2, after Removing Examinees Flagged at Time 2

	Time 1	Time 2		
	Mean (SD)	Mean (SD)	Raw Growth	Standardized Growth
All Examinees(N = 155)	1.07 (0.54)	1.31 (0.76)	0.24	0.45
Criteria for Removal at Time 2				
Flag 1 (N = 141)	1.09 (0.53)	1.45 (0.57)	0.36	0.66
Flag 2 (N = 152)	1.06 (0.54)	1.36 (0.68)	0.29	0.54
Flag 3 (N = 147)	1.08 (0.53)	1.42 (0.60)	0.33	0.62
Flag 4 (N = 149)	1.07 (0.54)	1.39 (0.63)	0.32	0.60
Flag 5 (N = 141)	1.10 (0.52)	1.46 (0.56)	0.36	0.66
Any Flag (N = 140)	1.10 (0.52)	1.46 (0.57)	0.36	0.67

Note: The standardized growth was defined as the mean growth divided by the standard deviation of the scores of all examinees at Time 1.

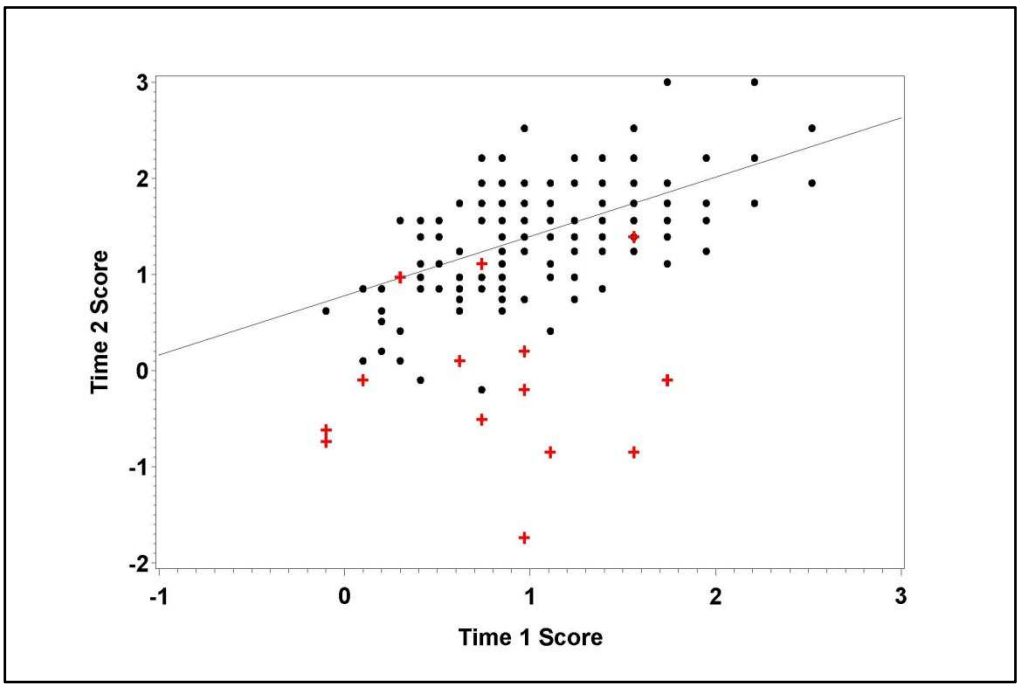

Figure 1: Examinees marked by Flag 1 (overall RTE) are indicated by the + symbol.

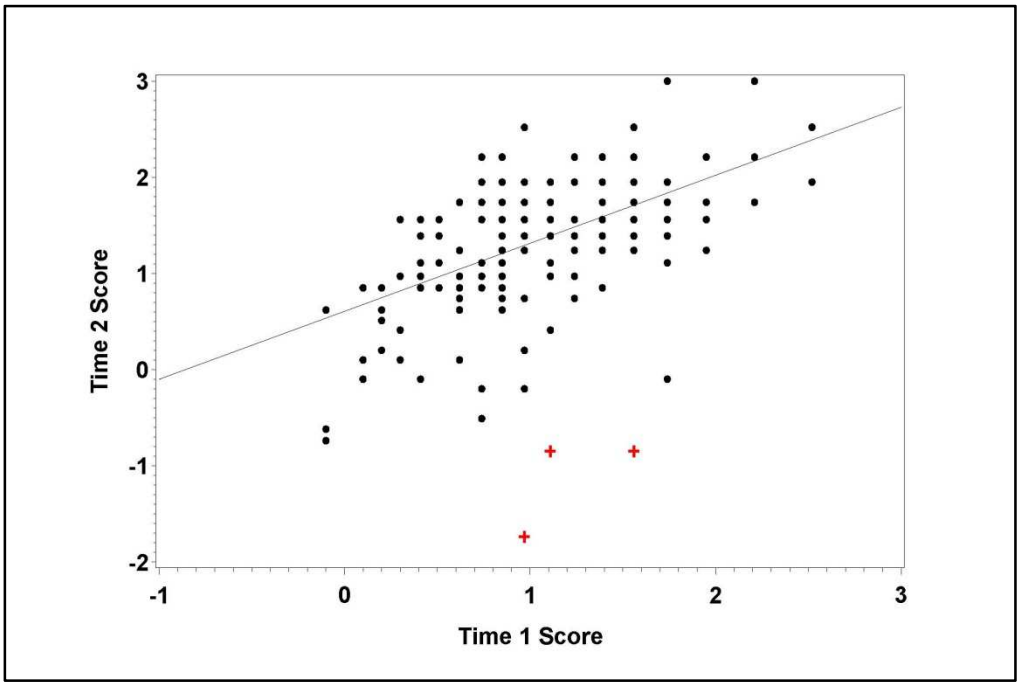

Figure 2: Examinees marked by Flag 2 (overall accuracy) are indicated by the + symbol.

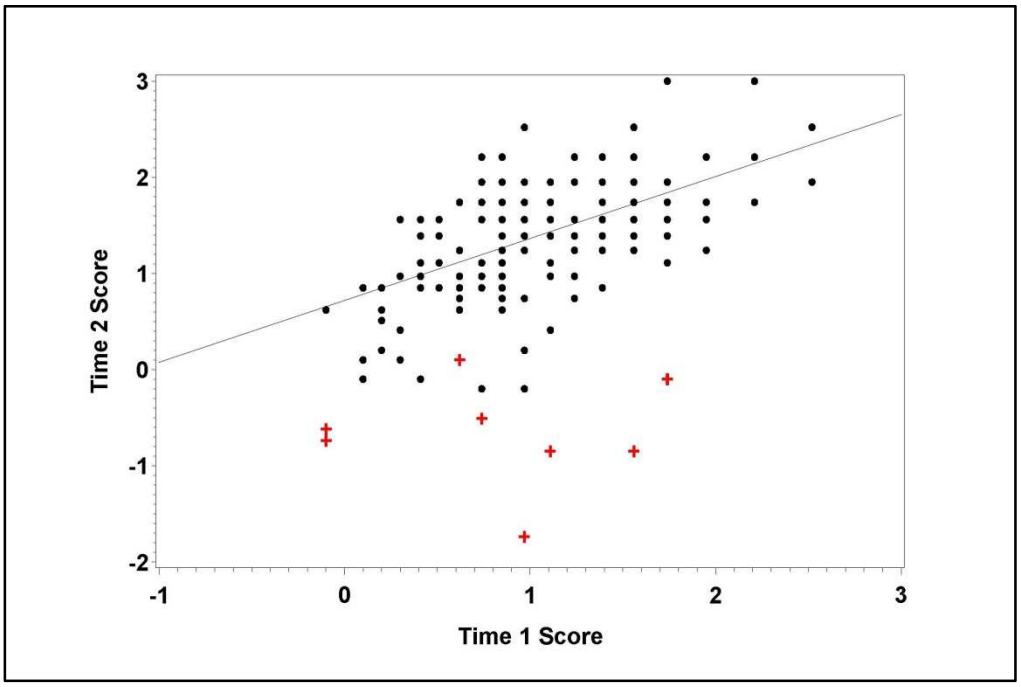

Figure 3: Examinees marked by Flag 3 (local RTE and local accuracy) are indicated by the + symbol.

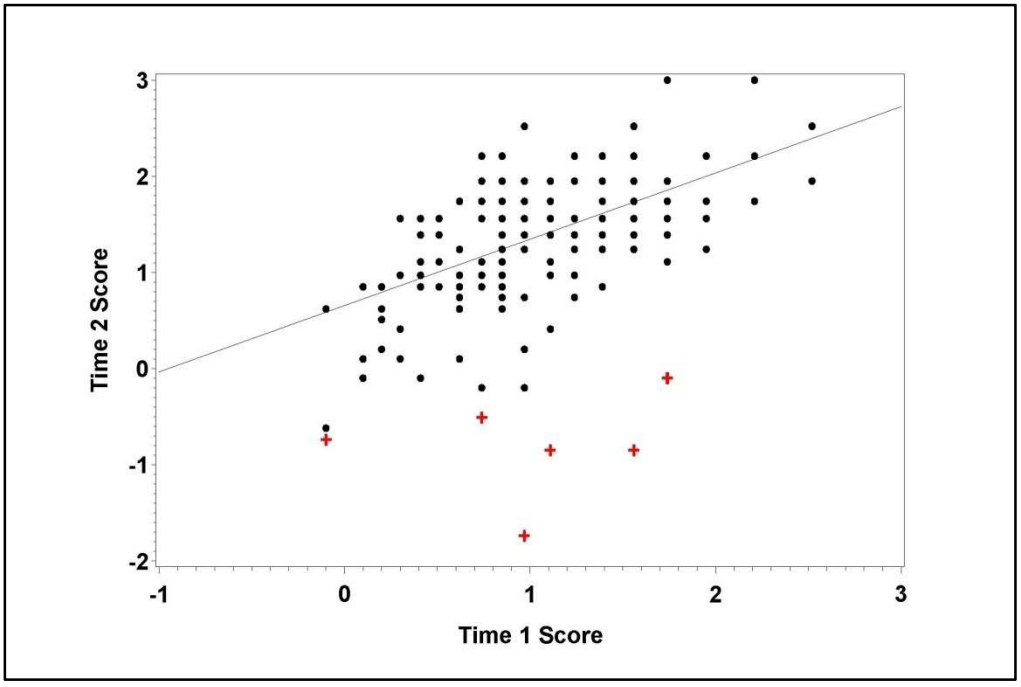

Figure 4: Examinees marked by Flag 4 (local accuracy) are indicated by the + symbol.

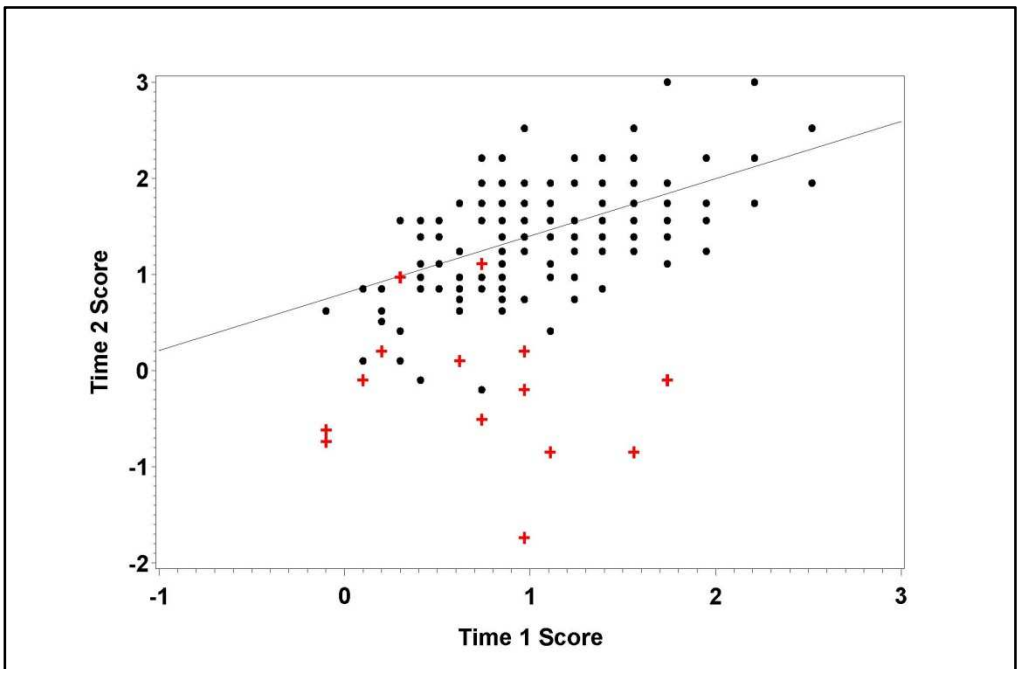

Figure 5: Examinees marked by Flag 5 (local RTE) are indicated by the + symbol.