

**An Investigation of the Changes in Item Parameter Estimates  
for Items Re-field Tested**

Jason L. Meyers

Xiaojing Jadie Kong

Katie Larsen McClarty

*Pearson*

*Paper Presented at the Annual Meeting of the American Educational Research Association*

*March 2008*

## **An Investigation of the Changes in Item Parameter Estimates for Items Re-field Tested**

Large-scale state testing programs typically rely upon a large bank of items to select from when building assessments. The typical item bank includes test items with varying item characteristics that are considered suitable for specific subject areas and instructional objectives. To ensure that the highest quality item statistics are used when building tests, test constructors are typically encouraged to use the most recently developed items. Many state testing programs have either implicit or explicit rules about the age at which an item must be retired from the item bank. In the state testing program studied here, test items are considered “outdated” three years after they were field tested. This three-year policy is primarily based on the assumption that item-level data become less accurate over time rather than upon any empirical research. To date, no known research investigates the age at which item statistics should no longer be trusted. Although item-response-theory (IRT)-based item characteristics, such as Rasch Item Difficulties (RIDs), are theoretically stable over time and across populations, outside factors (Bergstrom, Stahl, & Netzky, 2001) , such as test-taking motivation (Wise 2006; Wise & DeMars, 2005; Wise & DeMars, 2006), changes in curricula (Wells, Subkoviak, & Serlin, 2002), and item exposure or cheating (Stahl, Bergstrom, & Shneyderman, 2002) may influence the stability of item statistics. To investigate the assumption that item-level data become less accurate over time and the extent to which this occurs, therefore, empirical evidence is necessary.

The purpose of this study was to evaluate the decision to consider items outdated three years after field testing by investigating the changes in item parameter estimates for items that were re-field tested. This study builds upon a previously conducted pilot study that shed some light on item parameter changes over time, but raised additional questions.

In the current study, item characteristics were analyzed for a set of research items which had initially been field-tested in 2003 and were re-field tested in 2006. The findings of this study were used to evaluate the stability of the data associated with these items and to inform policies pertaining to item bank maintenance and test construction.

### **Pilot Study**

The initial study aimed at investigating item statistic stability by comparing changes in item statistics over a two year period.

### **Data**

The data used in this pilot study came from an assessment program in a large southern state in which students take standardized tests in the areas of reading/English language arts, mathematics, writing, social studies, and science. These assessments have been administered operationally since the 2003 school year. In grades three through eight, the tests contain between 40 and 50 multiple choice items and in high school the tests contain as many as 60 items. English Language Arts assessments also contain short answer and essay items in addition to the multiple choice items. The assessments are un-timed, and students have up to an entire school day to complete the assessment. The assessments are scaled and equated using the Rasch measurement model. In this pilot study, a large set of items originally field tested in 2002 was systematically re-field tested in 2004. Table 1 displays Rasch Item Difficulty estimates for the items during both administrations. Field test difficulty estimates are calculated using several thousand student responses.

### **Methods**

Differences between the two sets of RIDs (2002 vs. 2004) were calculated for each grade and subject. To further assess the observed differences, the 0.30 logits rule (Miller,

Rotou, & Twing, 2004) used in the testing program's operational equating procedure was applied to assess the practical significance of the differences in item difficulties. The testing program being analyzed utilizes a Rasch-based, common item equating design where the operational tests are equated by re-estimating item parameters and linking them back to their field-test estimates. As part of the post-equating procedure, a screening process is used to eliminate items from the common item set if the new and previous estimates differ by more than 0.30 logits. Items are placed on the appropriate scale by applying an equating constant representing the difference in average Rasch Item Difficulties between field-testing and live testing.

In the pilot study, the RID correlations were also examined. Additionally, the item discriminations (point-biserial correlations) were compared across time for each of the items re-field tested.

## Results

Table 2 displays the average RIDs in 2002 and 2004, the correlation between RIDs, the number and percentage of items flagged by the 0.30 logits rule, and the average equating constant in 2004 and 2005. Average point-biserial correlations are presented in Table 3.

Results indicated that the mean RID differences were considerably large. For instance, the grade 9 reading results indicated that there was an average RID increase of 0.53 logits from 2002 to 2004. The percentages of items flagged by the 0.30 logits criterion were 41%, 53%, 69%, 57% for math, science, social studies, and reading and ELA, respectively. Average mean differences were larger than the average equating constants observed in 2004 and 2005, indicating that the average change in difficulty across the two year period was more than is typically seen between field-testing and live testing.

### Limitations

Although the results showed that there were sizeable changes in the item statistics between the two years, the results from these analyses were confounded for several reasons. First, the year 2002 represented the initial field testing of a new state assessment. These items were administered via a separate field test in which students were asked to complete the exam after already completing a live assessment (this assessment's predecessor). While the state historically had concerns about student motivation on separate field tests, lack of motivation was suspected to be of particular concern at the high school grades. Secondly, this new assessment contained new, more rigorous item types than its predecessor. Students and teachers were not accustomed to seeing these new item types. Hence, not only were students asked to participate in an additional examination, they were asked to complete new, more difficult items that would have no impact on their earned and reported scores. Any observed change in RIDs across field test administrations (2002 and 2004) could either be attributed to lack of student motivation, measurement of new content, elapsed time between administrations, or any combination thereof.

### Current Study

The purposes of the current investigation were to determine how much item statistics had changed over a three-year time period, with the possible confounding factors present in the pilot study being controlled for, and to evaluate the practice of not including these three-year-old items on operational assessments. Little observed change over time would suggest that the items should be eligible for inclusion on operational tests. Significant change, however, would support the current practice and suggest that such items should either be retired from the item bank earlier or re-field tested to gather more updated item

statistics prior to their inclusion on an operational assessment. The results of the current study were compared with those of the pilot study to examine the similarities and differences in the findings.

### **Data**

The current study used research items from the same testing program as the pilot study that were embedded in 2003 live tests and re-field tested (without *any* changes or edits to the item) in 2006. Four subject areas and grade levels were examined: Grade 4 Mathematics and Reading, Grade 7 Mathematics and Reading, Grade 8 Social Studies, and Grade 10 Science.

Unlike the pilot study, however, student motivation was not thought to be an issue in this study. The potential effects of the confounding factors mentioned above were minimized in this study, because (1) field-test items were embedded within the live assessment so students did not know which items were field-test items, (2) students did not know that certain test forms contained re-field tested items, (3) these assessments were high-stakes in nature, and (4) students and teachers were sufficiently familiar with the item types in these assessments.

### **Method**

The two sets of item parameter estimates (2003 vs. 2006) were compared for each grade and subject. The following analyses were conducted:

- The magnitude of differences in RIDs was assessed. The 0.30 logits rule used to remove items from the common item equating set was applied in this context.
- Correlations between the two sets of RIDs were examined.

- Changes in RIDs were regressed on changes in item positions.
- Changes in item discrimination were examined.

## **Results**

Results of the current study are presented in Tables 4-7. The change in RIDs was calculated by subtracting the 2003 field test RID from the 2006 re-field test RID. A positive value of mean change indicates that the RID was higher for the 2006 administration (i.e., an increase in item difficulty) while a negative value indicates that the RID was lower for the 2006 administration (i.e., a decrease in item difficulty). Figures 1-6 graphically depict the differences in RID values for the individual items.

As shown in Tables 4 and 5, RIDs did not change substantially when re-field tested in 2006. This pattern was observed across grades and subjects. For most grades and subjects, the mean RID difference was approximately 0.1, and standard deviations of RIDs ranged from 0.411 to 0.897. There were a few items flagged by the 0.30 criterion across grades and subjects. The percentages of flagged items (ranging from 25% to 36%) were noticeably lower than those of the pilot study (ranging from 24% to 84%). The correlations between the RIDs across grades and subjects were highly positive, indicating a high correspondence between the two sets of indices. The RID correlation for grade 4 math was relatively low ( $r = .68$ ). Given the fact that the RID range for Grade 4 Math was small, with a range of 1.108 and 1.392 logits for the 2003 data and 2006 data, respectively, the strength of the RID correlation may have been affected by the restriction of the range.

Previous research (Meyers, Miller, & Way, 2006) suggests that change in item position between testing occasions can have a significant impact on change in item difficulty. To investigate the potential impact on re-field-tested items, the change in RID was regressed

on the change in item position between field test administrations. Results of this analysis are presented in Table 6. Contrary to the previous research, the results indicated that position change was not a statistically significant predictor of item difficulty change across grade levels and subject areas ( $\alpha = 0.01$ ).

In addition to item difficulty, the change in item discrimination was examined. As shown in Table 7, very minor differences in point-biserials were found across grade and subject. The largest difference in average point biserials was observed in Grade 4 Reading, with an absolute value of 0.02. Not surprisingly, the change in item positions was not a significant predictor of change in point biserials either.

### **Conclusions and Recommendations**

The current analyses indicate that the item statistics in this testing program do not substantially change over a three year period. Unlike the data in the pilot study in which the lack of student motivation and the transition of testing programs could have had major impact on item parameter estimation, the present study controlled for the potential influence of test-taking motivation by embedding items on live assessments. The magnitude of the observed item difficulty changes was relatively small, and the correlations of item difficulties were rather high. Furthermore, item position change did not impact change in item difficulty or item discrimination. By reducing the confounding variables in the present study, the effect of elapsed time was investigated more thoroughly here than in the previous research. Given these results, it was recommended that test constructors continue the general policy of building tests with items most recently field tested , but that they consider using up to 3-year-old items when those items meet their needs better than more recent items.

### **Importance of the Study**

In this testing program, hundreds of items are developed at each grade level and subject each academic year in order to ensure that enough items exist to build assessments that adhere to strict content and statistical requirements. Not all of the quality items developed each year end up making it onto an operational assessment. These good items accumulate over time and many are thought to expire because the integrity of the data is questioned. If, as the results of this study suggest, these items are still usable, two major benefits are possible. First, because developing items is an extremely costly endeavor, being able to use items longer after they were developed will save money and not waste development efforts. Secondly, the longer use of items could allow for higher quality assessments. Often test constructors have trouble locating an item with specific content or statistical properties. If there is an item that meets their needs, but is considered “outdated,” they might have to settle for an item that is not as good of a match to their content and statistical needs. Thus, having more available items to choose from should result in better assessments from both a psychometric and content perspective.

### **Suggestions for Future Research**

As a result of the current research, more in depth investigation of those items that were flagged as changing in difficulty more than the 0.30 criterion is warranted. A content review of these flagged items is planned by subject matter experts. Any patterns or features of the items that may make them differentially affected by elapsed time should be noted and taken into consideration in all test construction efforts. In addition, future research could investigate the stability of item statistics over an even longer time period to determine the point at which item statistics change substantially. Finally, additional empirical data are

needed to confirm or contradict the results observed in this study, particularly from testing programs that utilize different measurement models, field-test procedures or equating methods than those in the testing program analyzed here.

## References

- Bergstrom, B., Stahl, J.A., & Netzky, B. A. (2001) *Factors that influence item parameter drift*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- Meyers, J. L, Miller, G. E., & Way, W. D. (2006). *Item position and item difficulty change in an IRT-based common item equating design*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Miller, G. E., Rotou, O., & Twing, J.S. (2004). Evaluation of the 0.30 logits criterion in common item equating. *Journal of Applied Measurement, 5*(2), 172-177.
- Stahl, J. A., Bergstrom, B. A., & Shneyderman, O. (2002). *Impact of item drift on test-taker measurement*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77-87.
- 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*(2), 95-114.
- Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 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.

Table 1. Mean Rasch Item Difficulties for each grade and subject in the 2002 and 2004 field test administrations.

Subject	Grade	Number of Items	2002 RID				2004 RID			
			Mean	Min	Max	SD	Mean	Min	Max	SD
Mathematics	9	66	0.16	-1.691	2.16	0.936	0.18	-1.685	2.248	1.01
	10	76	0.184	-2.259	2.632	0.889	0.201	-2.64	2.833	1.101
	11	65	0.216	-1.688	3.001	0.926	0.411	-1.83	3.456	1.126
	Overall	207	0.187	-2.259	3.001	0.911	0.26	-2.64	3.456	1.081
Science	10	8	0.652	-2.611	2.524	1.452	0.795	-2.293	2.422	1.378
	11	47	0.651	-0.857	1.744	0.604	0.85	-1.024	1.946	0.715
	Overall	55	0.651	-2.611	2.524	0.764	0.842	-2.293	2.422	0.826
Social Studies	10	13	0.371	-2.669	2.25	1.618	0.689	-2.475	2.177	1.431
	11	36	1.023	-0.782	2.003	0.666	1.426	-0.83	2.924	0.845
	Overall	49	0.852	-2.669	2.25	1.031	1.231	-2.475	2.924	1.068
Reading and ELA	9	61	-0.253	-1.893	1.646	0.744	0.279	-1.397	2.094	0.78
	10	55	0.149	-0.971	1.192	0.538	0.265	-1.464	1.82	0.778
	11	145	-0.253	-1.984	2.317	0.667	-0.068	-2.243	1.906	0.896
	Overall	261	-0.169	-1.948	2.317	0.679	0.083	-2.423	2.094	0.86

Table 2. Summary of differences in RIDs between 2002 and 2004 field test administrations.

Subject	Grade	Number of Items	Mean RID 2002	Mean RID 2004	Change in Mean RID (2004-2002)	Correlation between RIDs	Number (%) of Items Flagged by 0.30 criterion	2004 Equating Cons	2005 Equating Cons
Mathematics	9	66	0.160	0.180	0.020	0.966	16(24%)	0.003	-0.027
	10	76	0.184	0.201	0.016	0.943	34(45%)	-0.057	-0.059
	11	65	0.216	0.411	0.195	0.929	35(54%)	-0.024	-0.024
	Overall	207	0.187	0.260	0.073	0.941	85(41%)	-	-
Science	10	8	0.652	0.795	0.143	0.983	5(63%)	0.049	0.030
	11	47	0.651	0.850	0.200	0.865	24(51%)	-0.060	0.017
	Overall	55	0.651	0.842	0.191	0.908	29(53%)	-	-
Social Studies	10	13	0.371	0.689	0.319	0.964	8(63%)	-0.116	0.008
	11	36	1.026	1.426	0.401	0.925	26(72%)	-0.068	0.017
	Overall	49	0.852	1.231	0.379	0.939	34(69%)	-	-
Reading/ELA	9	61	-0.253	0.280	0.533	0.951	51(84%)	0.985	0.394
	10	55	0.149	0.265	0.116	0.915	22(40%)	0.896	0.380
	11	145	-0.253	-0.068	0.186	0.878	76(52%)	0.102	0.333
	Overall	261	-0.169	0.084	0.252	0.879	149(57%)	-	-

Table 3. Average change in item discrimination between 2002 and 2004 administrations by grade and subject

<b>Subject</b>	<b>Grade</b>	<b>Number of Items</b>	<b>Change in Mean Point-Biserial</b>	<b>Min</b>	<b>Max</b>	<b>SD</b>
Mathematics	9	66	-0.017	-0.16	0.13	0.062
	10	76	-0.016	-0.2	0.21	0.075
	11	65	-0.01	-0.25	0.16	0.084
	Overall	207	-0.014	-0.25	0.21	0.073
Science	10	8	-0.033	-0.15	0.06	0.08
	11	47	-0.02	-0.2	0.1	0.076
	Overall	55	-0.022	-0.2	0.1	0.076
Social Studies	10	13	-0.041	-0.16	0.11	0.069
	11	36	-0.01	-0.15	0.12	0.06
	Overall	49	-0.018	-0.16	0.12	0.063
Reading and ELA	9	61	0.034	-0.09	0.14	0.036
	10	55	0.119	-0.04	0.23	0.059
	11	145	0.091	-0.31	0.42	0.083
	Overall	261	0.084	-0.31	0.42	0.075

Table 4. Mean Rasch Item Difficulties for each grade and subject in the 2003 and 2006 field test administrations.

Subject	Grade	Number of Items	2003 RID				2006 RID			
			Mean	Min	Max	SD	Mean	Min	Max	Std
Math	4	14	-0.307	-0.825	0.283	0.302	-0.201	-0.821	0.571	0.411
	7	18	-0.121	-1.188	0.544	0.45	0.018	-1.153	0.853	0.5
	Overall	32	-0.202	-1.188	0.544	0.398	-0.078	-1.153	0.853	0.469
Reading	4	20	-0.477	-1.7	1.295	0.789	-0.461	-2.057	1.272	0.897
	7	20	0.022	-1.761	1.142	0.761	-0.055	-1.793	1.122	0.809
	Overall	40	-0.227	-1.761	1.295	0.806	-0.258	-2.057	1.272	0.868
Science	8	20	0.376	-0.305	1.259	0.443	0.524	-0.369	1.435	0.517
Social Studies	10	18	-0.066	-1.112	1.05	0.681	0.106	-0.955	1.222	0.708

Table 5. Summary of differences in RIDs between 2003 and 2006 field test administrations.

<b>Subject</b>	<b>Grade</b>	<b>Number of Items</b>	<b>Mean RID 2003</b>	<b>Mean RID 2006</b>	<b>Change in Mean RID (2006-2003)</b>	<b>Correlation between RIDs</b>	<b>Number (%) of Items Flagged by 0.30 Criterion</b>
Mathematics	4	14	-0.307	-0.201	0.106	0.684	5(36%)
	7	18	-0.121	0.018	0.139	0.932	5(28%)
	Overall	32	-0.202	-0.078	0.125	0.862	10(31%)
Science	10	18	-0.066	0.106	0.172	0.957	6(33%)
	Overall	18	-0.066	0.106	0.172	0.957	6(33%)
Social Studies	8	20	0.376	0.524	0.148	0.889	5(25%)
	Overall	20	0.376	0.524	0.148	0.889	5(25%)
Reading	4	20	-0.477	-0.461	0.016	0.957	5(25%)
	7	20	0.022	-0.055	-0.077	0.966	5(25%)
	Overall	40	-0.227	-0.258	-0.031	0.961	10(25%)

Table 6. Change in item position between field testing in 2003 and 2006

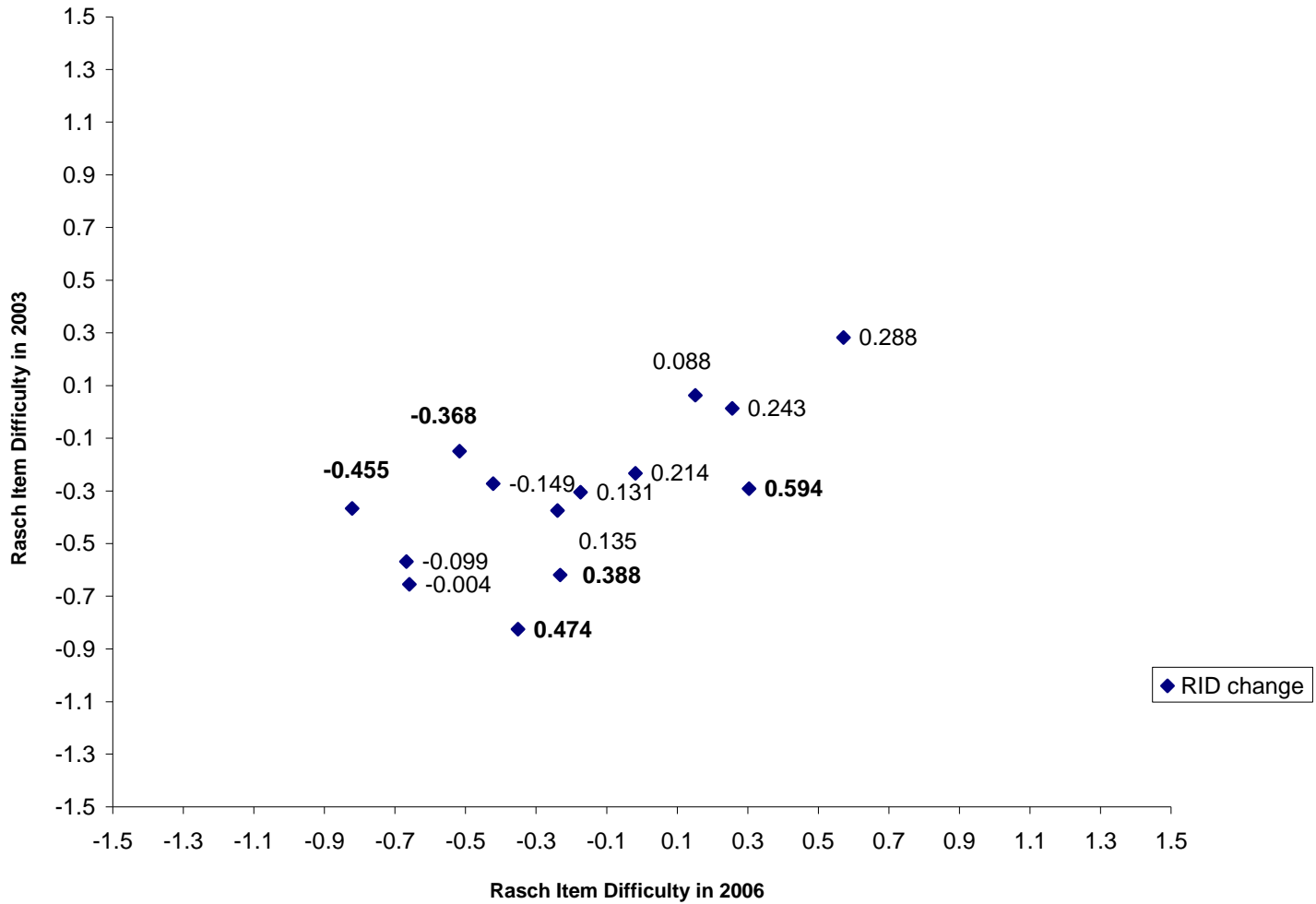
Subject	Grade	n Items	Change in Item Position			SD
			Mean	Min	Max	
Math	4	14	0	-5	6	3.245
Math	7	18	0	-7	8	3.879
Reading	4	20	2	-6	10	4.768
Reading	7	20	0	-6	5	3.093
Social Studies	8	20	1	-7	8	4.166
Science	10	18	0	-5	5	2.768

Note: The change in item position and RID was calculated by subtracting the 2003 re-field test value from the 2006 field test value

Table 7. Average change in item discrimination between 2003 and 2006 administrations by grade and subject

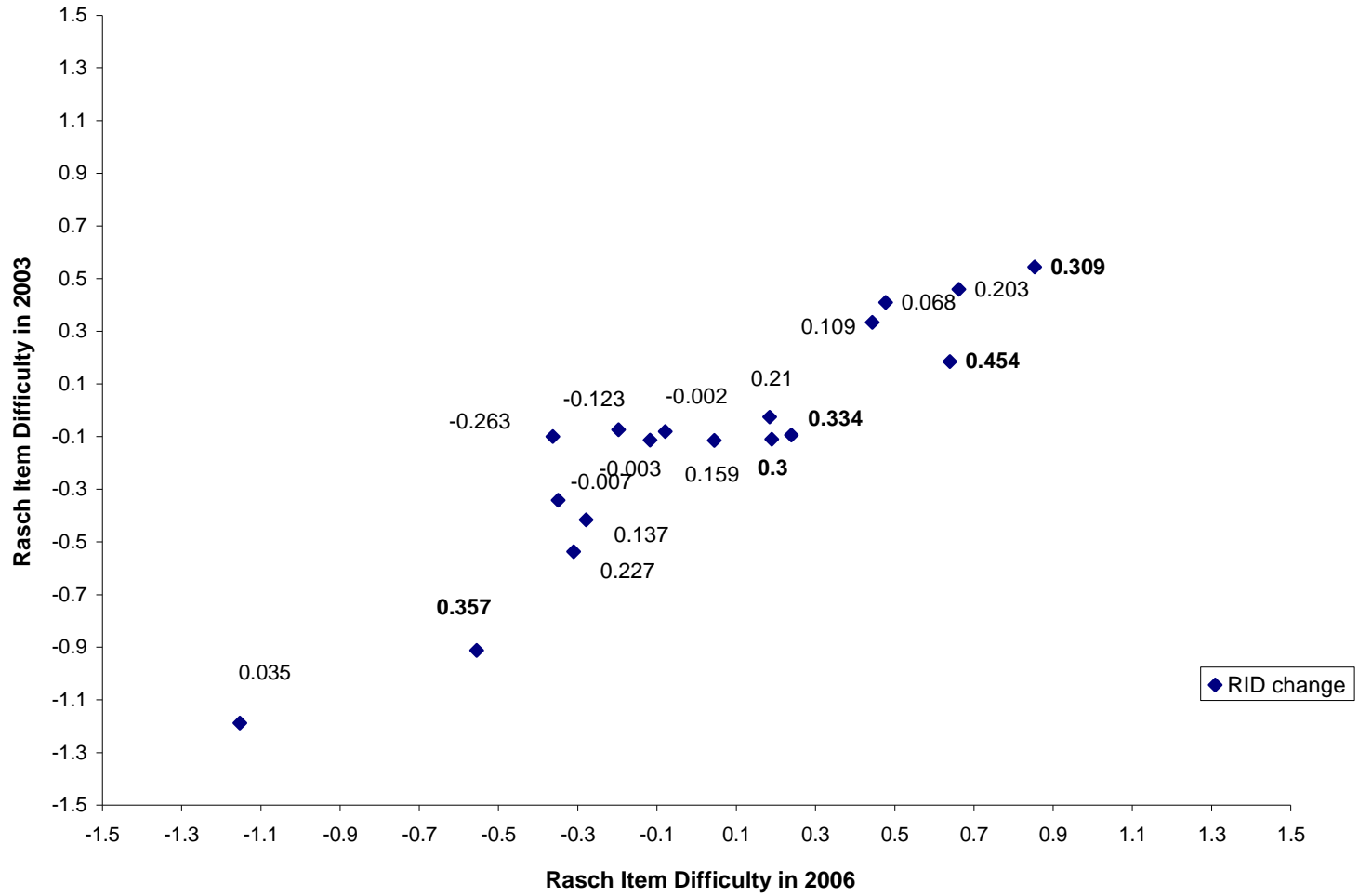
<b>Subject</b>	<b>Grade</b>	<b>Number of Items</b>	<b>Mean PBIS Change</b>	<b>Min</b>	<b>Max</b>	<b>SD</b>
Math	4	14	-0.012	-0.1	0.02	0.035
	7	18	0.022	-0.03	0.07	0.026
	Overall	32	0.007	-0.1	0.07	0.034
Reading	4	20	-0.024	-0.08	0.01	0.03
	7	20	-0.006	-0.06	0.06	0.027
	Overall	40	-0.015	-0.08	0.06	0.03
Science	8	20	0.002	-0.11	0.06	0.041
Social Studies	10	18	0.006	-0.04	0.06	0.03

Figure 1. Grade 4 Math: Changes in Rasch Item Difficulty 2003 to 2006



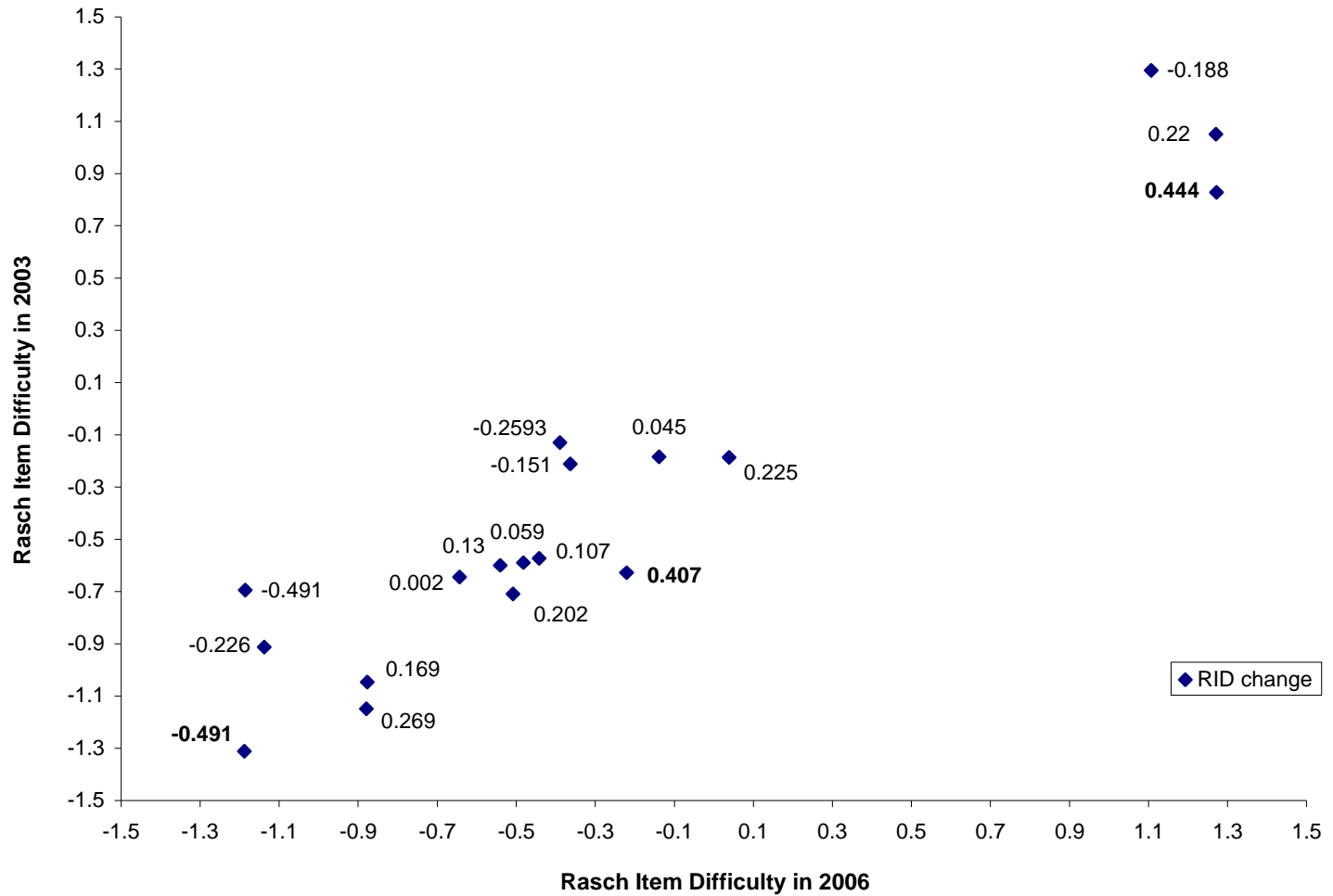
Note: Items in bold-faced type indicate differences larger than 0.30 logits.

Figure 2. Grade 7 Math: Changes in Rasch Item Difficulty 2003 to 2006



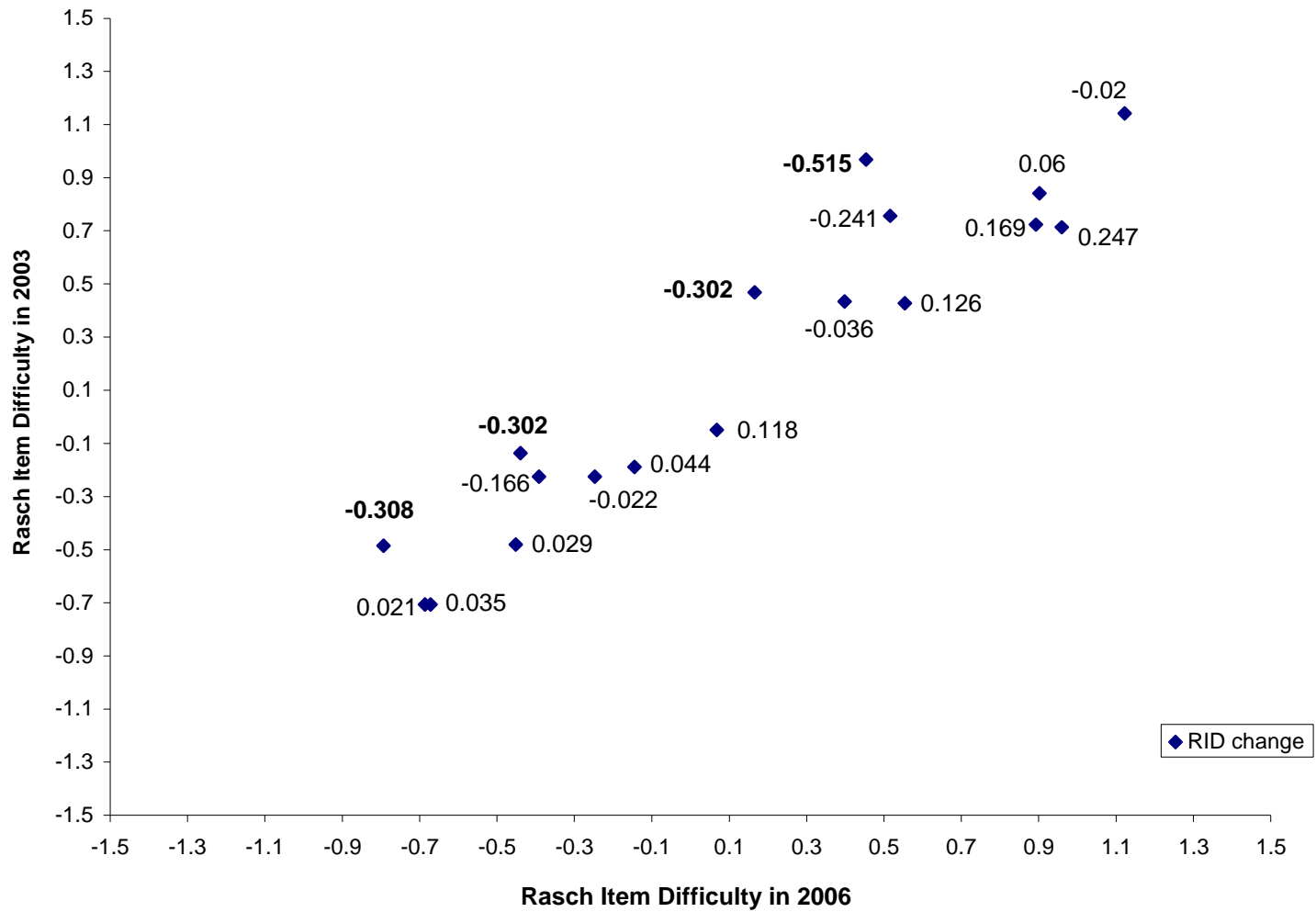
Note: Items in bold-faced type indicate differences larger than 0.30 logits.

**Figure 3. Grade 4 Reading: Changes in Rasch Item Difficulty 2003 to 2006**



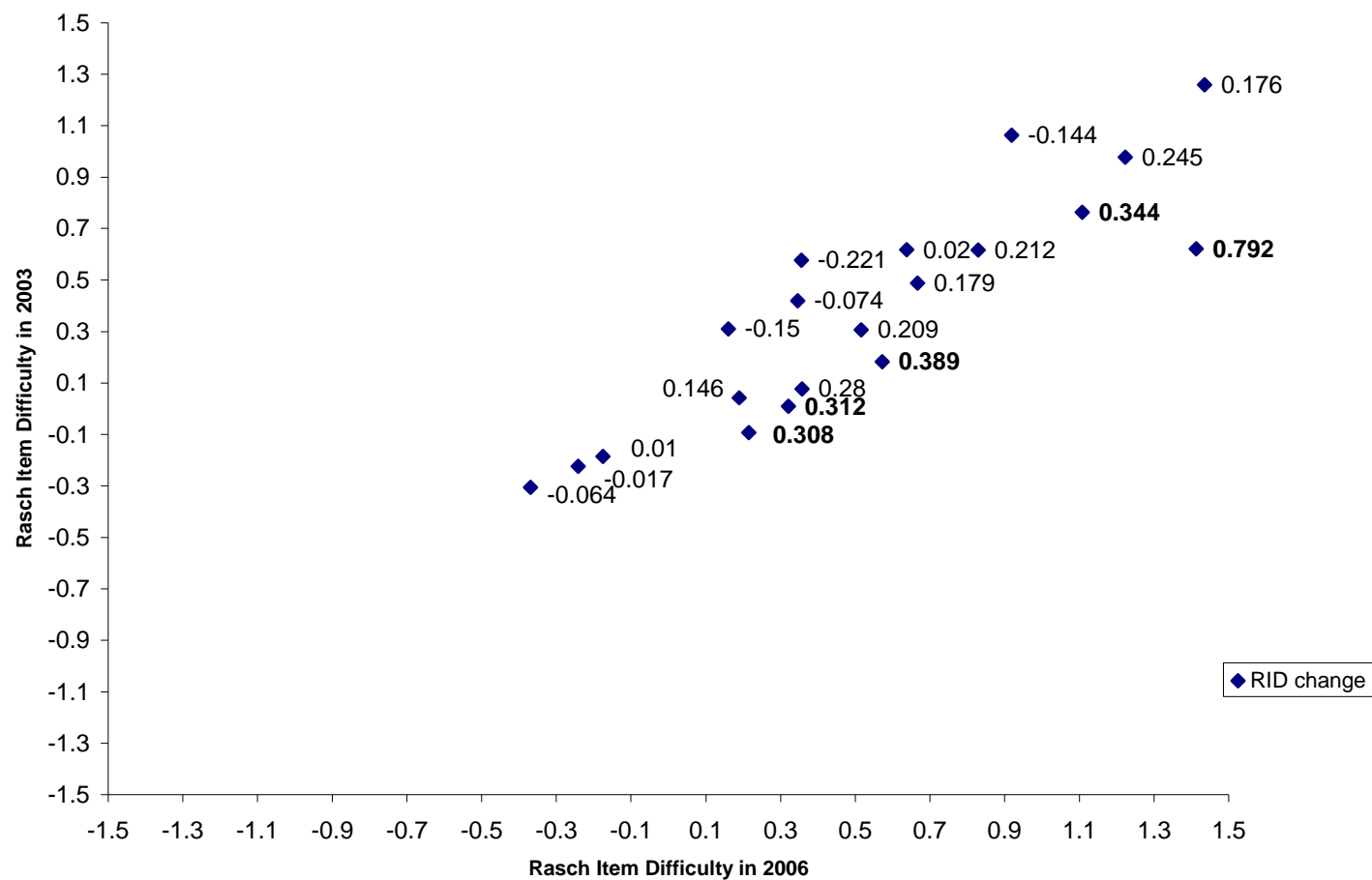
Note: Items in bold-faced type indicate differences larger than 0.30 logits.

Figure 4. Grade 7 Reading: Changes in Rasch Item Difficulty 2003 to 2006



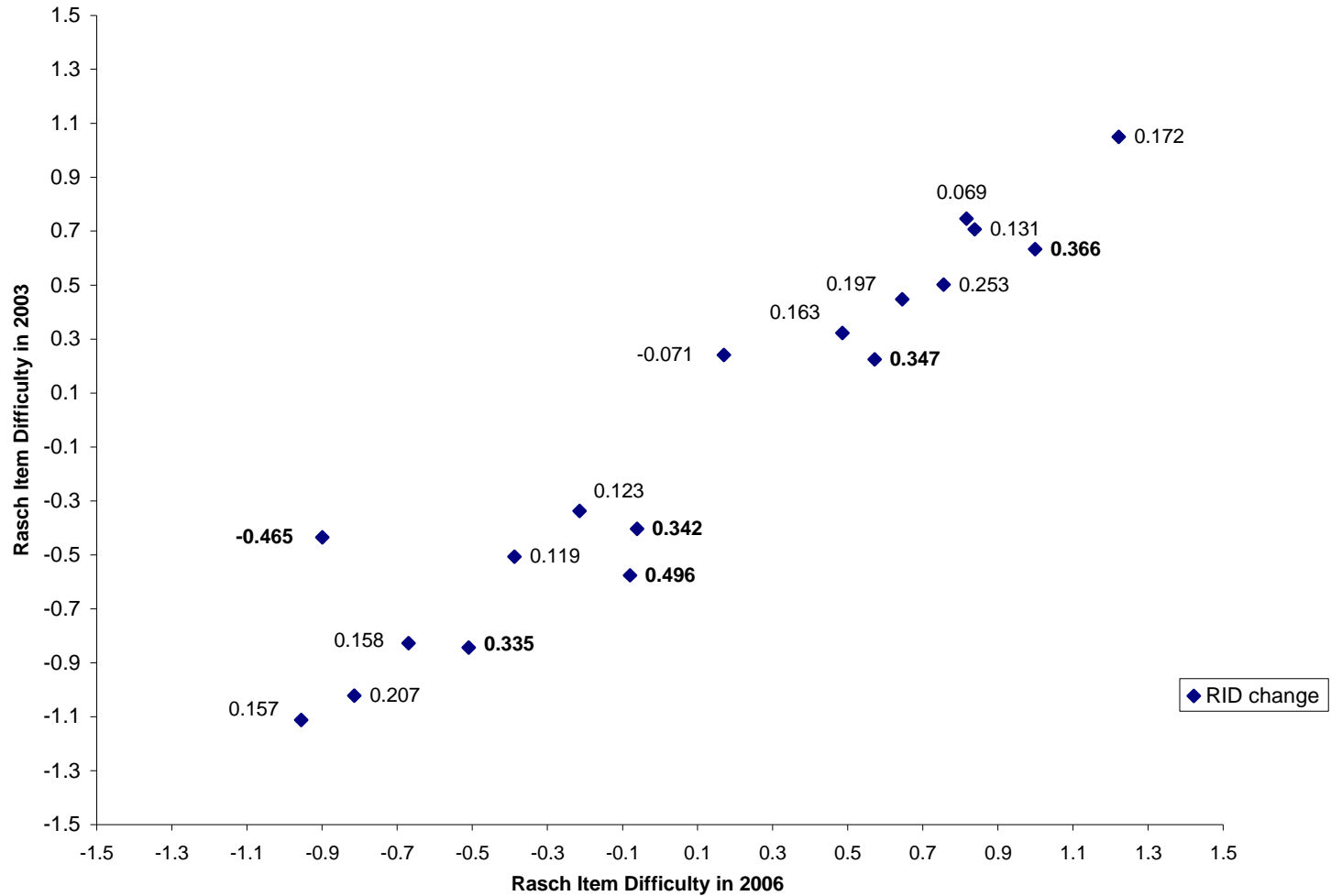
Note: Items in bold-faced type indicate differences larger than 0.30 logits.

**Figure 5. Grade 8 Social Studies: Changes in Rasch Item Difficulty 2003 to 2006**



Note: Items in bold-faced type indicate differences larger than 0.30 logits.

Figure 6. Grade 10 Science: Changes in Rasch Item Difficulty 2003 to 2006



Note: Items in bold-faced type indicate differences larger than 0.30 logits.