Listening. Learning. Leading.®

Research Report
ETS RR–14-31

Use of Longitudinal Regression in Quality Control

Ying Lu

Wendy M. Yen

December 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Use of Longitudinal Regression in Quality Control

Ying Lu[1] & Wendy M. Yen[2]

1 Educational Testing Service, Princeton, NJ
2 Wendy M. Yen Psychometrics, LLC, Pebble Beach, CA

This article explores the use of longitudinal regression as a tool for identifying scoring inaccuracies. Student progression patterns, as evaluated through longitudinal regressions, typically are more stable from year to year than are scale score distributions and statistics, which require representative samples to conduct credibility checks. Historical data from a large-scale K-12 testing program were used to evaluate the usefulness of several proposed longitudinal procedures. Results showed that the use of longitudinal regression in quality control was effective in detecting scoring errors, especially when the scoring errors were non-negligible.

Keywords Longitudinal data; scoring; large-scale assessment; accountability; equating

doi:10.1002/ets2.12032

Longitudinal regression refers to the application of regression procedure, linear or nonlinear, to longitudinal data where the outcome of an individual is measured at several different time points. For any educational assessment program that is intended to measure students' ability/achievement at different grade levels or different time points within a grade level, longitudinal data may be obtained by collecting multiple assessment results across an extended time span for the same cohort of students. Longitudinal regression may be conducted with a more recent measure being the dependent variable and earlier measure(s) being the independent variable(s). For instance, the regression of 2011 Grade 5 math score over 2010 Grade 4 math score for a specific group of students is regarded as a longitudinal regression. The pattern of math score change, for example, from Grade 4 to Grade 5, is generally referred to as a progression pattern in this article. Having vertically scaled test data is not required for the use of longitudinal regression.

The purpose of this article is to explore the use of longitudinal regression as a tool for identifying scoring inaccuracies, including problems with equating or item parameter estimates that contribute to inaccurate scoring tables. Specifically, a credibility check can be conducted through comparing new and historical longitudinal regression functions or comparing observed scores with predicted scores based on prior performance and historical regression functions. Based on historical results, flagging criteria can then be identified in terms of the size of pattern/score deviations for when operational scores do/do not pass the credibility check.

For almost all testing programs, quality control procedures are used to ensure that expected quality standards are achieved during scoring, equating, and reporting of test scores. Kolen and Brennan (2004, p. 309) list six quality controls with which to monitor equating:

1. Check that the administration conditions are followed properly.
2. Check that the answer key is correctly specified.
3. Check that the items appear as intended.
4. Check that the equating procedures specified are followed correctly.
5. Check that the score distribution and score statistics are consistent with those observed in the past.
6. Check that the correct conversion table or equation is used with the operational scoring. For example, if test forms are constructed based on known attributes and postequating is used, a conversion table developed based on post-equating should be reasonably close to the one that is developed based on previously known attributes.

Allalouf (2007) suggested additional quality control processes that compare student performance (scores or pass/fail rates) with prior expectations based on examinee background, exam date, and repeater data.

Corresponding author: Y. Lu, E-mail: ylu@ets.org

These quality control procedures, however, do not involve the use of longitudinal data. Previous research on the use of longitudinal data in quality control is very limited. Prior to the No Child Left Behind (NCLB) Act, longitudinal data were mostly used by state education agencies (as well as parents, teachers, and others) for the purpose of improved instruction. With NCLB and Race to the Top program, longitudinal data started to serve multiple purposes, including evaluating teacher performance and measuring student growth. But, so far, there has been little research on the use of longitudinal regression as one part of validating equating results.

The use of longitudinal regression in quality control can be especially helpful to K-12 assessments, for which it is commonplace that schedules force the release of scores before exhaustive credibility checks can be conducted. For K-12 programs with long testing windows (i.e., tests that are administered through an extended period of time), conversion tables usually need to be delivered for scoring when only a small percentage of responses, mostly from the early test takers, are available. Therefore, the equating sample is likely not fully representative of the state population. In such situations, the usefulness of comparing score distributions and score statistics over years as a credibility check before score release is limited due to the possible performance difference between the equating sample and the testing population. Quality control procedures based on longitudinal regressions have the advantage of being less stringent with respect to the requirement of sample representativeness. Student progression patterns, as evaluated through longitudinal regressions, typically are more stable from year to year (and group to group) than are scale score distributions and statistics. This is so because longitudinal regression is a conditional approach that evaluates the performance of the current year assessment by taking into consideration prior performance of the same cohort of students. Since it is less critical to obtain state-representative student samples for longitudinal analyses, longitudinal quality control procedures can be conducted at an earlier stage.

It should be noted, however, that there can exist legitimate differences between the new and historical progression patterns, or patterns of academic progress. The proposed procedure is intended to be used solely as an early warning flag to help analysts identify the most unexpected results for further quality control scrutiny.

This study proposes several longitudinal procedures for examining the potential inaccuracy of a set of test scores and evaluates the usefulness of these procedures using historical results from a large statewide K-12 testing program.

Data

In this study, longitudinal data for English-language arts (ELA) Grades 2 – 3, 5 – 6, and 8 – 11 and mathematics Grades 2 – 7 from 2005 to 2010 were used, with the 2010 administration assumed to be the new administration for which scoring quality control is needed. Score summary of the tests included in this study from 2005 to 2010 are presented in Table 1, including scale score range, mean, and standard deviation. Although these tests have the same scale score range of 150 – 600, their score scales were established independently and were not in any way related to each other. Note also that there is no vertical scale supporting the tests administered across different grade levels.

The unique statewide student identifier (SSID), which became available starting from 2006, was used to merge student records across the years from 2006 to 2010. The 2005 dataset was merged with 2006 data based on student name, date of birth, and school district.

Method

Longitudinal regression analyses to examine students' progression patterns from one grade level to the next grade level were conducted. Scale scores from the same grade across years were placed on the same scale after equating, and a key assumption of the proposed approach in this study is that the progression patterns should be reasonably stable from year to year. The approach involves two types of procedures: comparing regression functions and comparing prediction accuracy.

Examining Regression Functions

The first type of procedure involved comparing the linear regression functions or curves for the grade level progressions of interest with historical regressions. For instance, the regression function that defines the expected fifth grade math score given a fourth grade math score can be determined using merged Grades 5 and 4 math records from any two adjacent years. It is expected that some differences might exist, say, between the regression of 2009 Grade 5 math on 2008 Grade 4 math and the regression of 2008 Grade 5 math on 2007 Grade 4 math, but the differences are likely to be minor if equating

Table 1 Scale Score Distribution of the Studied Tests

		2010		2009		2008		2007		2006		2005	
Test	Range	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ELA G02	150 – 600	357	65	353	63	348	60	345	62	344	63	336	61
ELA G03	150 – 600	342	63	340	63	333	58	330	59	331	62	324	61
ELA G05	150 – 600	359	54	356	57	348	52	343	54	342	57	340	56
ELA G06	150 – 600	357	54	352	54	345	53	340	54	337	55	335	54
ELA G08	150 – 600	357	63	348	61	341	58	339	58	339	56	334	55
ELA G09	150 – 600	354	60	350	60	348	61	345	60	339	63	340	60
ELA G10	150 – 600	341	61	338	60	336	61	331	59	328	61	328	58
ELA G11	150 – 600	337	67	332	68	327	65	328	71	324	70	323	64
Math G02	150 – 600	382	86	377	81	372	81	369	82	372	86	366	83
Math G03	150 – 600	395	92	388	90	379	86	371	85	369	84	362	78
Math G04	150 – 600	390	79	383	77	374	75	366	73	361	74	354	70
Math G05	150 – 600	383	92	376	92	365	87	357	87	356	90	350	89
Math G06	150 – 600	361	75	354	75	348	71	343	69	341	68	340	70
Math G07	150 – 600	352	69	345	66	339	65	336	65	338	68	334	66

Note. ELA = English-language arts; G = grade.

results are valid and scoring is done correctly. As part of the evaluation process, regression curves were examined and compared graphically. The regression parameters were also compared across years. The variations among, or ranges of, the historical regression curves and parameters provide a reference for what should be expected for the regression curve and parameters developed based on the new administration data.

Examining the Scores Resulting From Applying the Regression Functions

The second type of procedure used the regression function developed based on previous years' testing populations to predict student performance in the target year. Prediction accuracy was examined and compared across years. For instance, the regression function developed based on predicting 2009 Grade 5 math scores from 2008 Grade 4 math scores was used to predict a student's 2010 Grade 5 math score given his or her 2009 Grade 4 math score.[1] The prediction results from this step were evaluated by examining summary statistics and cumulative distributions of residuals. Using regression functions developed from earlier datasets for prediction should provide reasonably accurate results, given stable progression patterns across years. Deviations of residuals from past patterns can act as a flag that leads to more in-depth quality control evaluations.

Flagging criteria were determined using the historical regression/prediction patterns for both statistical estimates and graphical evaluation. For statistical estimates, the pool of null parameter estimates was defined using the 2005 – 2009 testing populations. The mean and standard deviation of the historical values were calculated, and regression parameter estimates based on the new administration were flagged if they fell outside of the range, as defined by the historical mean, plus and minus 3 historical standard deviations. For graphical evaluation, visual inspection of the differences between the new administration pattern and historical patterns was used to flag abnormalities.

For ease of reading, the four procedures used in this study were assigned abbreviations, which were used in summarizing results. Specifically, *param* refers to the evaluation procedure that compared the new regression parameters with the historical regression parameters. *Curve* refers to the evaluation procedure that compared the new regression curve with the historical regression curves. *Resid CDF* refers to the evaluation procedure that compared the cumulative distribution of prediction residuals for the current year with those for previous years. *RMSD* refers to the evaluation procedure that compares the RMSD of the predicted values and observed values from the current year with those for previous years.

The power of the longitudinal regression approach was investigated by evaluating the extent to which any wrong conversion tables applied to the equating sample of the new administration can be effectively detected. In this study, the applied wrong conversion tables were the conversion tables for the same subject but for the wrong grade level. For instance, for the ELA Grade 9 2010 administration, the wrong conversion table used in this study was the ELA Grade 8 2010 conversion table. For the longitudinal regression approach to be useful for quality control purposes, it should be able to identify

Table 2 Root Mean Squared Deviations (RMSD) Between the Wrong Conversion Table and the Correct Conversion Table

	Wrong conversion source		Wrong conversion summary	
Target test (2010)	Test	Year	*RMSD*	Standardized *RMSD*
ELA03	ELA02	2010	13	0.20
ELA06	ELA05	2010	3	0.06
ELA09	ELA06	2010	7	0.12
ELA10	ELA09	2010	20	0.33
ELA11	ELA10	2010	8	0.12
Math03	Math02	2010	4	0.04
Math04	Math03	2010	19	0.24
Math05	Math04	2010	40	0.43
Math06	Math05	2010	19	0.25
Math07	Math06	2010	13	0.19

Note. ELA = English-language arts.

cases that were manipulated in the study to have wrong conversion tables applied, and to not flag cases where the correct conversion tables were applied to the equating sample.

Results

Results were organized into two major categories: comparing regression functions and comparing prediction accuracy. The power and the Type I error rate of the new approach were examined under each category. Using 2010 as the new administration, the new growth patterns displayed by 2010 equating samples were presented together with historical trends for comparison and evaluation. For each operational test in 2010, two conversion tables were applied, including one correct conversion table and one wrong conversion table.

Similar to hypothesis testing, the sensitivity of quality control procedures in detecting the wrong conversion tables is largely determined by how wrong the conversion tables are. To quantify this attribute, the deviation of the wrong conversion table from the correct conversion table was summarized by RMSD of the correct scale score and the wrong scale score assigned to each individual student in the early sample. The resulting RMSD between the correct and wrong conversion table was weighted by the number of examinees at each obtainable score point. Here a correct scale score refers to an observed scale score based on a valid equating procedure, not a true scale score (which would only be available under simulated conditions). Since the RMSD is most meaningful when interpreted together with the score scale, a standardized RMSD was also calculated by dividing the original RMSD by the standard deviation of the scores.

Table 2 summarizes the RMSD and standardized RMSD of the wrong conversion tables that were applied to each 2010 operational test included in the study. For each wrong conversion table, the source of the conversion table (i.e., the subject test it belonged to and the administration year it came from) is also presented. The table shows that RMSD values ranged from 3 to 40, and the standardized RMSD values ranged from 0.04 to 0.43. The quality control procedure was expected to be more capable of detecting wrong conversion tables with larger RMSD values.

Examining Regression Functions

Appendix A presents the regression parameter estimates and standard error of estimates for all adjacent test combinations included in the study. Regression root mean squared error (RMSE) and r^2 are also provided. The upper section of each table shows the historical regression functions based on the overall yearly testing populations. The lower section of each table shows the regression functions based on the 2010 equating sample with correct or wrong conversion tables applied to the 2010 test. As an example, in Table A1, the regression of *all0910* means the regression of ELA Grade 3 scores on ELA Grade 2 scores based on all matched cases from 2009 and 2010, and the regression of *all0506* means the regression of 2006 ELA Grade 3 scores on their matching 2005 ELA Grade 2 scores. In the lower section of Table A1, *eq0910* means the regression function based on the 2010 ELA Grade 3 equating sample with matching 2009 ELA Grade 2 scores and with the correct conversion table applied to the 2010 ELA Grade 3 test. The last row of the table gives similar information with *eq0910* but under the condition of the wrong conversion table being applied to the 2010 ELA Grade 3 test. For instance,

regression estimates under *eq0910_wrong* were based on the 2010 ELA Grade 3 equating sample with matching 2009 ELA Grade 2 scores and with the wrong conversion table (i.e., ELA Grade 2 conversion table from 2010) applied to score the 2010 ELA Grade 3 responses.

Note that the case counts for the *all* samples are close to 400,000 or greater, while the case counts for the equating samples range from about 68,000 to 90,000. The standard errors for the regression parameters are, therefore, about half the size for the *all* samples than the equating samples.

Appendix A shows that the regression pattern, or regression parameter estimates, varied to some degree across the years. In fact, year-to-year variation accounted for much more of the variance than sampling errors within years. Standard errors of parameter estimates were small because of the very large sample sizes. For some tests, equating samples also showed a slightly different progression pattern compared to the overall testing populations, as demonstrated by the minor difference in parameter estimates under *all0910* and *eq0910*. Because of year-to-year growth variation and the minor difference in progression pattern between the equating sample and the overall testing population, the flagging of wrong conversion tables could be difficult when the wrong conversion tables deviate minimally from the correct conversion tables.

A simple rule was used to test if the regression parameter estimates based on the 2010 equating sample came from the pool of parameter estimates based on historical data. In our case, the pool of estimates could be best defined using the 2005 – 2009 testing populations (i.e., estimates under *all0809*, *all0708*, *all0607*, and *all0506*). The mean and standard deviation of the slope and intercept parameter estimates in the pool were calculated, and the parameter estimates based on the 2010 equating samples were flagged if they fell outside of the range, as defined by the historical mean, plus and minus 3 historical standard deviations. As an example, consider the regression of ELA Grade 6 scores over ELA Grade 5 scores. Table A2 presents the historical slope and intercept parameter estimates based on 2005 – 2009 data. The range for the complete samples was calculated to be (36.27, 72.95) and (0.77, 0.92) for the intercept and slope, respectively. An examination of the lower section of Table A2 shows that parameter estimates under *eq0910* fell within the ranges for both slope and intercept, and the intercept parameter estimates under *eq0910_wrong*, which was based on the 2010 equating sample with the wrong conversion table applied, fell outside of the range defined by historical data for the intercept parameter.

For the longitudinal regression approach to be useful for quality control purposes, we would like the flagging criteria to be able to identify all cases manipulated in the study to have wrong conversion tables applied (i.e., the last row in Tables A1 – A10) and to not flag cases where the correct conversion tables were applied to the equating sample (i.e., the second row from the last in Tables A1 – A10). All regression parameter estimates based on the 2010 equating samples for all tests included in the study were evaluated using this rule, and results were summarized in Table 3. The conversion tables that were identified as problematic were flagged using *X* in the table. It was noted that only one of the 10 correct conversion tables was flagged. Although the sample size of 10 datasets was too small to establish an accurate Type I error rate, the

Table 3 Conversion Tables Identified as Incorrect

	Correct conversion				Wrong conversion			
Test (2010)	Param	Curve	Resid CDF	*RMSD*	Param	Curve	Resid CDF	*RMSD*
ELA03				X			X	X
ELA06					X			X
ELA09					X	X	X	X
ELA10					X	X	X	X
ELA11					X	X		
Math03								
Math04					X	X	X	X
Math05					X	X	X	X
Math06					X	X	X	X
Math07	X				X	X	X	

Note. Param = the evaluation procedure that compares the new regression parameters with the historical regression parameters; curve = the evaluation procedure that compares the new regression curve with the historical regression curves; resid CDF = the evaluation procedure that compares the cumulative distribution of prediction residuals for the current year with those for previous years; RMSD = the evaluation procedure that compares the RMSDs of the predicted values and observed values from the current year with those for previous years.

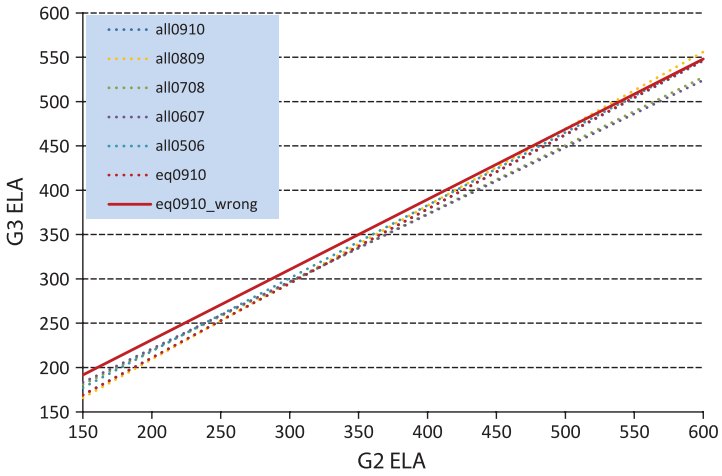

Figure 1 Linear regression of Grade 3 English-language art (ELA) on Grade 2 ELA.

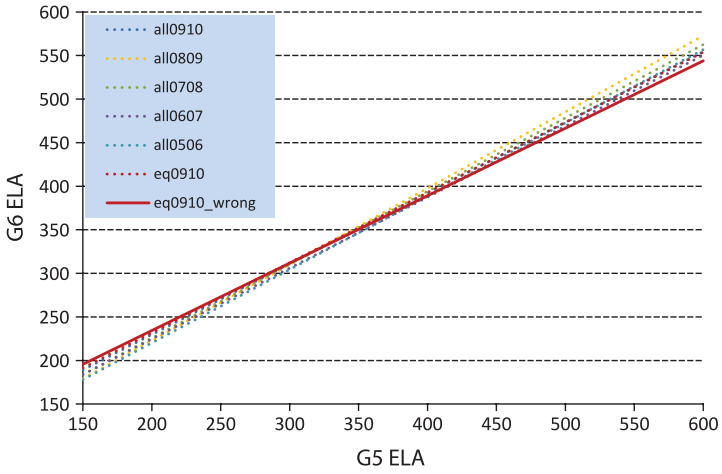

Figure 2 Linear regression of Grade 6 English-language art (ELA) on Grade 5 ELA.

results show that the Type I error rate was quite low using the proposed flagging criteria. Eight of the 10 wrong conversion tables were flagged, showing reasonable power in detecting scoring inaccuracy.

As an alternative way to compare regression functions, historical regression lines were presented in Figures 1–10 together with the new regression lines based on the equating sample for verification. Historical regression lines were based on the total population with about 400,000 observations. New regression lines were based on the equating samples with 68,000–90,000 observations. Graphical evaluation, although less objective than a statistical test on some occasions, can be useful in evaluating the discrepancy among the regression curves. The naming conventions provided in the figure legends are consistent with those used in Tables A1–A10. Again, *all0910* did not constitute a historical pattern, as the information would not have been available at the time the new 2010 conversion was developed based on the equating sample needed to be verified. Rather, it was used to evaluate how the equating sample deviated from the testing population in terms of progression pattern. These plots were evaluated with the focus being to compare the two regression lines with names starting with *eq0910* to the historical regression curves. Note that the results based on the wrong conversions are represented by solid lines. The regression lines were flagged in Table 3 under the column heading, *Curve*, if they were noticeably different from the historical curves (i.e., the regression line visually fell outside the collection of correct, historical regression lines). The graphs suggested that none of the correct conversions would be flagged. More than half of the wrong conversions were flagged, especially those that were known to deviate considerably from the correct conversion tables, as indicated by the standardized deviation measures in Table 2. The method of visual comparison of regression lines was not powerful enough to detect small errors in applied conversions. Comparison between the regression lines

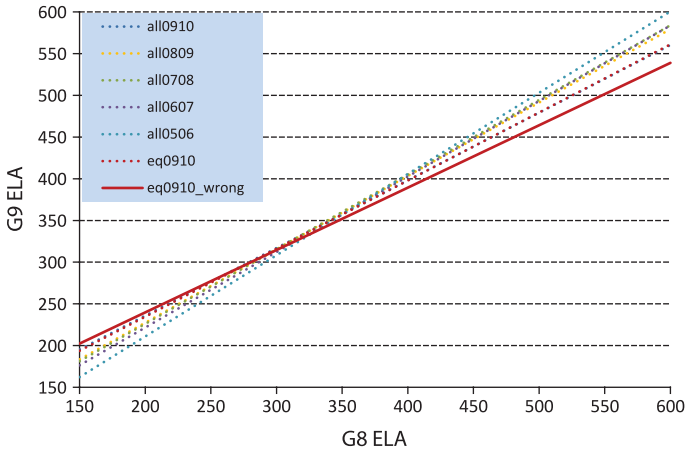

Figure 3 Linear regression of Grade 9 English-language art (ELA) on Grade 8 ELA.

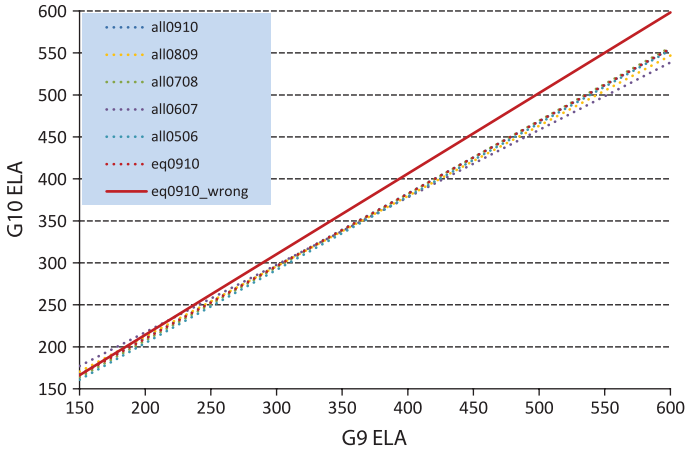

Figure 4 Linear regression of Grade 10 English-language art (ELA) on Grade 9 ELA.

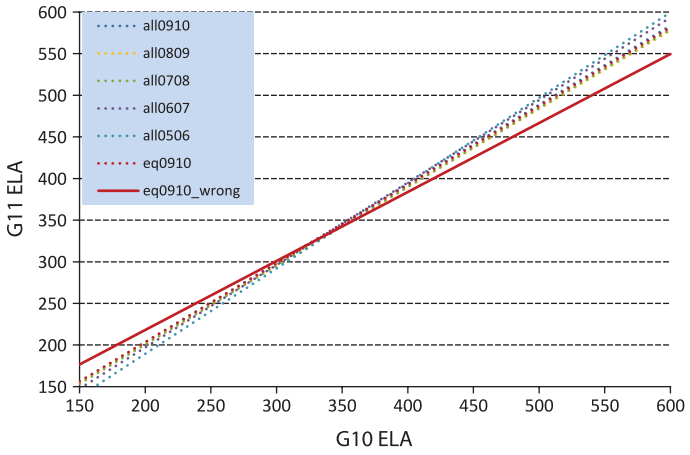

Figure 5 Linear regression of Grade 11 English-language art (ELA) on Grade 10 ELA.

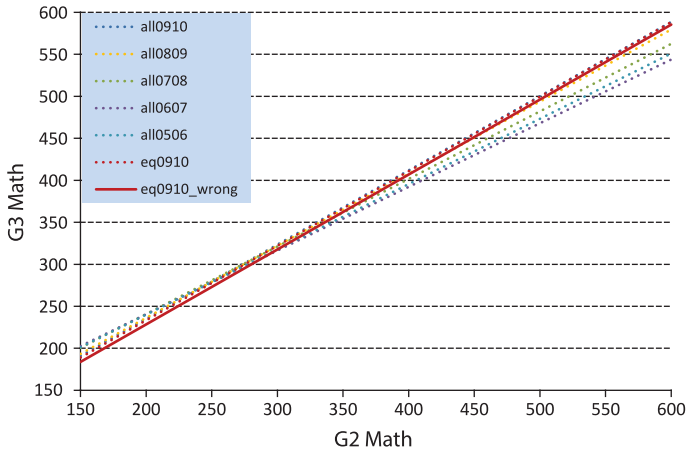

Figure 6 Linear regression of Grade 3 math on Grade 2 math.

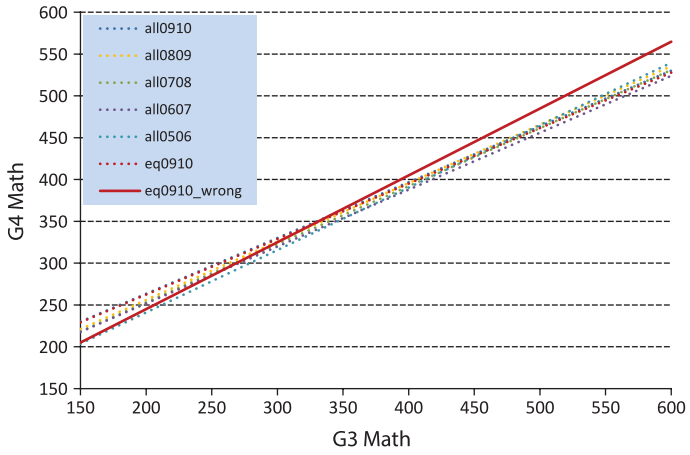

Figure 7 Linear regression of Grade 4 math on Grade 3 math.

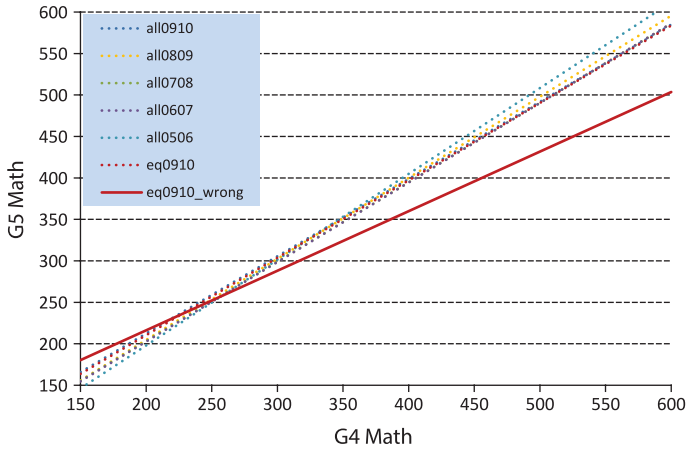

Figure 8 Linear regression of Grade 5 math on Grade 4 math.

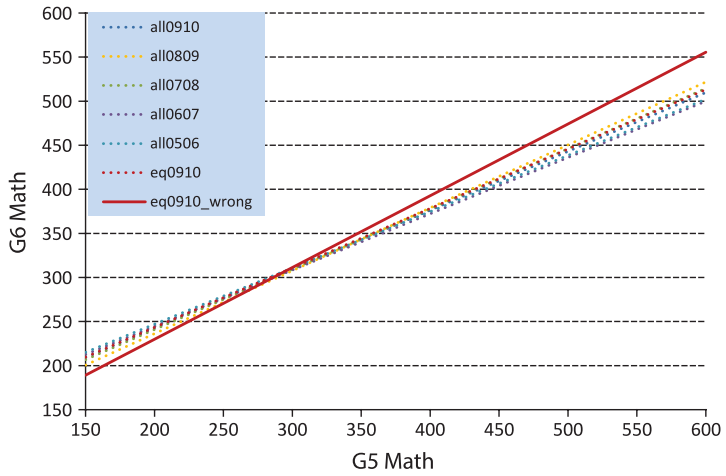

Figure 9 Linear regression of Grade 6 math on Grade 5 math.

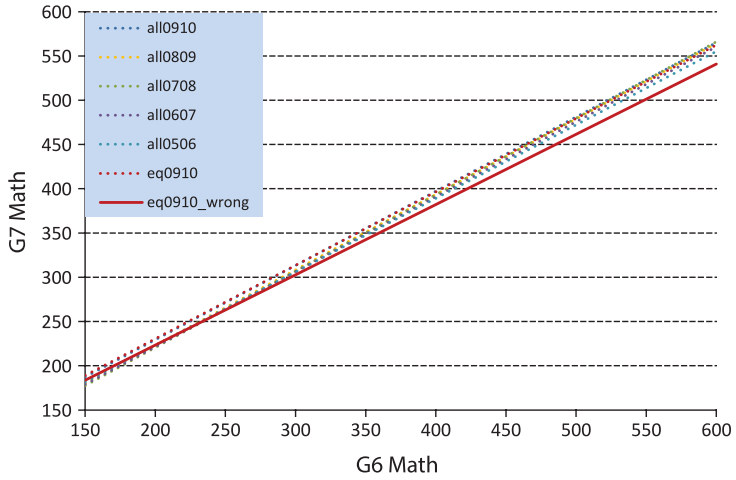

Figure 10 Linear regression of Grade 7 math on Grade 6 math.

under *all0910* and *eq0910* showed that, in general, the equating sample and the testing population yielded similar growth patterns, although some minor difference in regression lines could be observed.

Examining the Scores Resulting From Applying the Regression Functions

In addition to comparing regression functions, progression patterns can be indirectly compared by examining the accuracy of predicting student performance in 1 year using the regression function developed from previous years. For instance, the regression of 2009 Grade 3 ELA on 2008 Grade 2 ELA can be used to predict the 2010 Grade 3 ELA score for students who took Grade 2 ELA in 2009. If the student progression pattern stays similar from year to year, the size and distribution of prediction residuals should also be similar to the prediction residuals from previous years. To examine if the prediction residuals were reasonable in terms of size and range, the residual cumulative distributions for predicting the current year's scores were compared with cumulative distributions of the same type of prediction residuals from previous years. The residuals were computed as the observed value subtracted from the predicted value.

Figures 11 – 20 show the cumulative distributions of residuals resulting from predicting 2010, 2009, 2008, and 2007 test scores. For example, Figure 11 shows the cumulative distribution of residuals for the prediction of 2010 Grade 3 ELA using the 2008 – 2009 regression function based on all students (*pred_all10*), the prediction of 2009 Grade 3 ELA using the 2007 – 2008 regression function (*pred_all09*), the prediction of 2008 Grade 3 ELA using the 2006 – 2007 regression

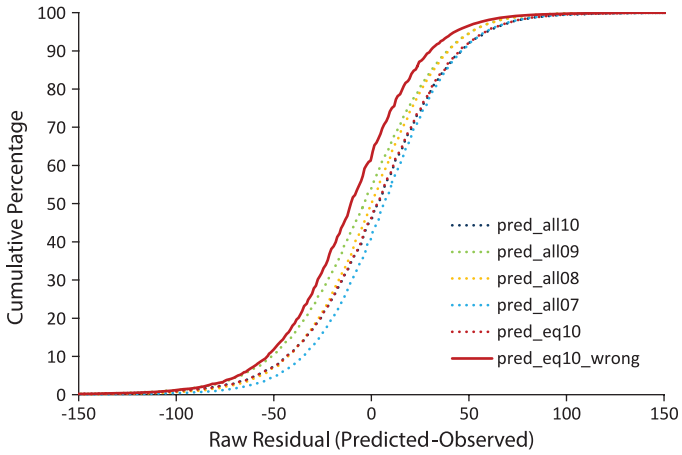

Figure 11 Cumulative distribution of residuals for predicting Grade 3 English-language arts (ELA).

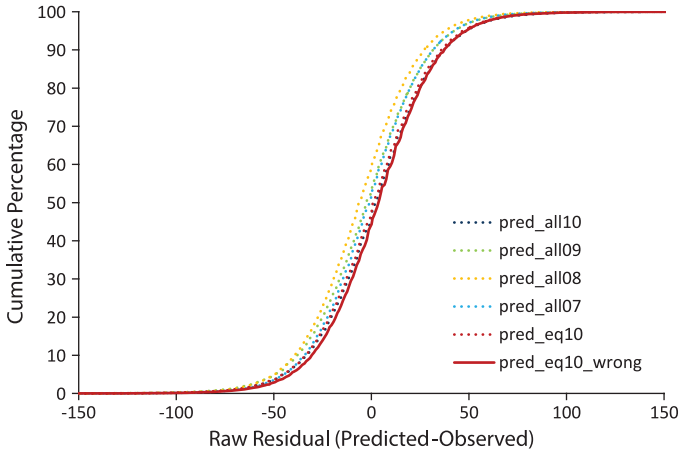

Figure 12 Cumulative distribution of residuals for predicting Grade 6 English-language arts (ELA).

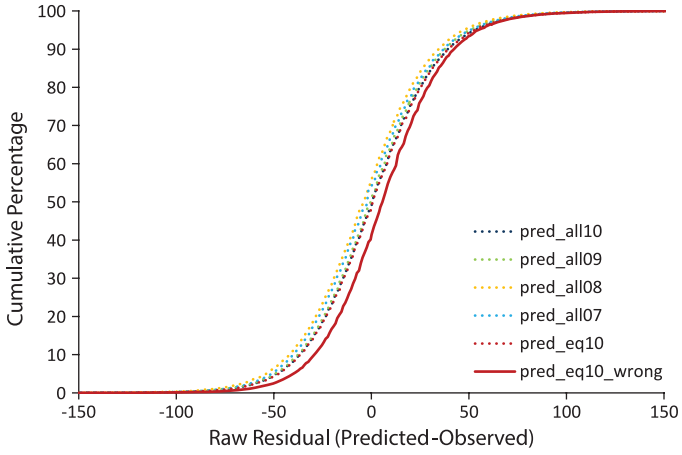

Figure 13 Cumulative distribution of residuals for predicting Grade 9 English-language arts (ELA).

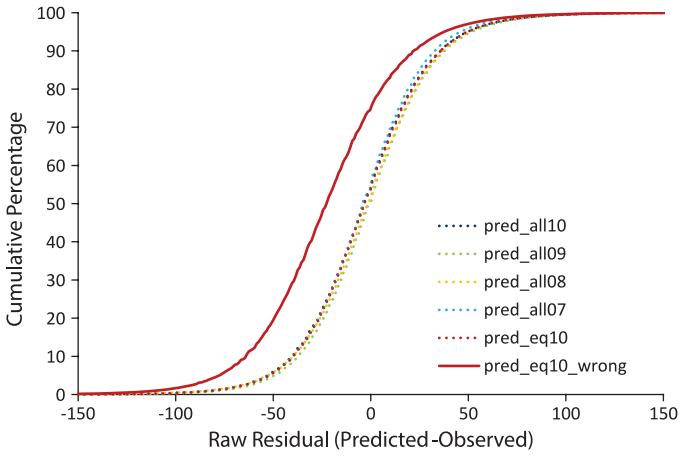

Figure 14 Cumulative distribution of residuals for predicting Grade 10 English-language arts (ELA).

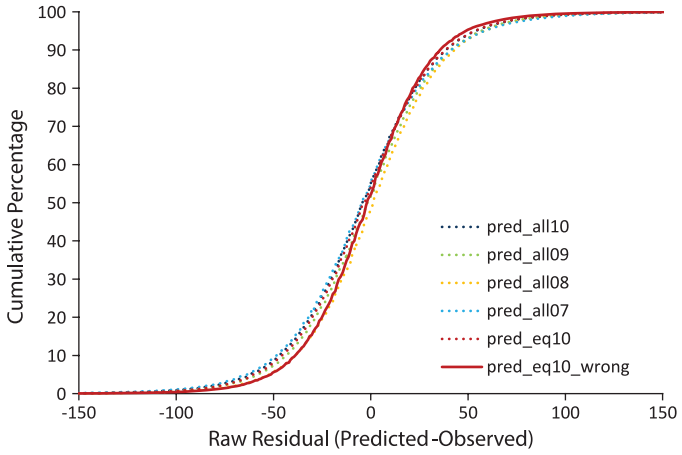

Figure 15 Cumulative distribution of residuals for predicting Grade 11 English-language arts (ELA).

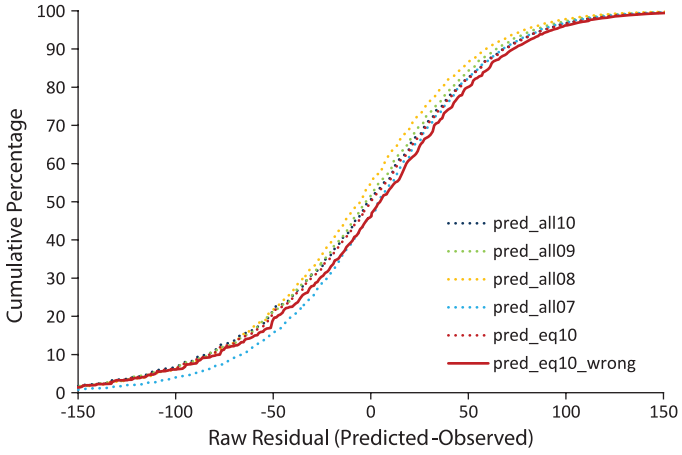

Figure 16 Cumulative distribution of residuals for predicting Grade 3 math.

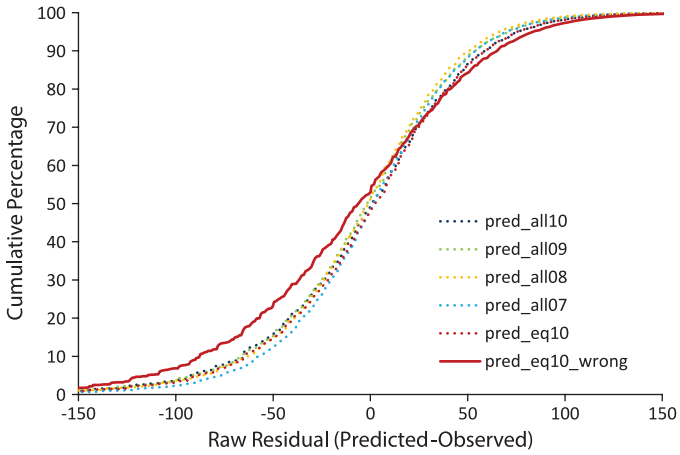

Figure 17 Cumulative distribution of residuals for predicting Grade 4 math.

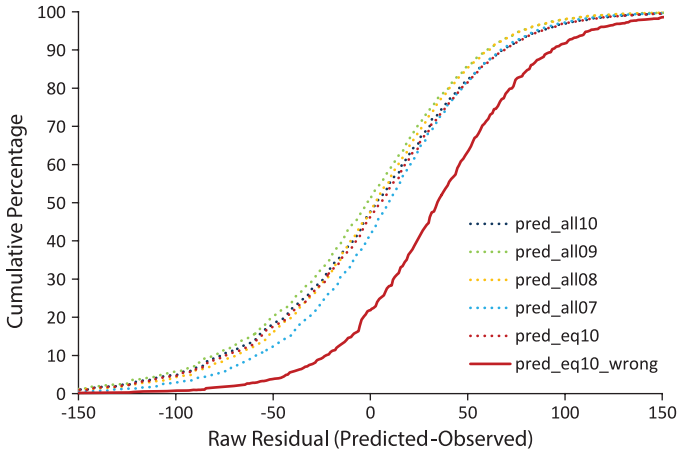

Figure 18 Cumulative distribution of residuals for predicting Grade 5 math.

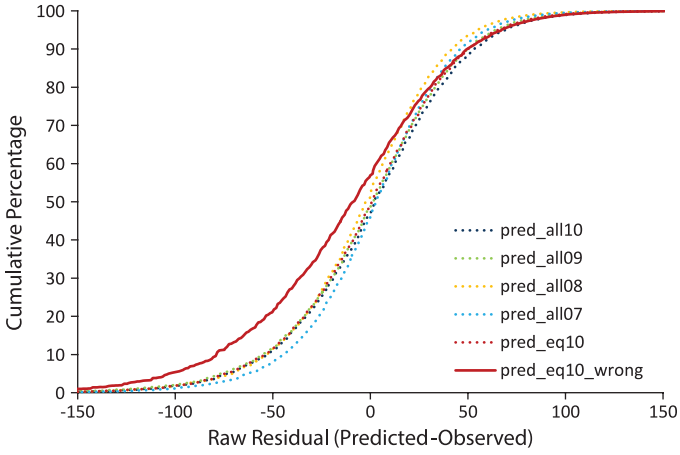

Figure 19 Cumulative distribution of residuals for predicting Grade 6 math.

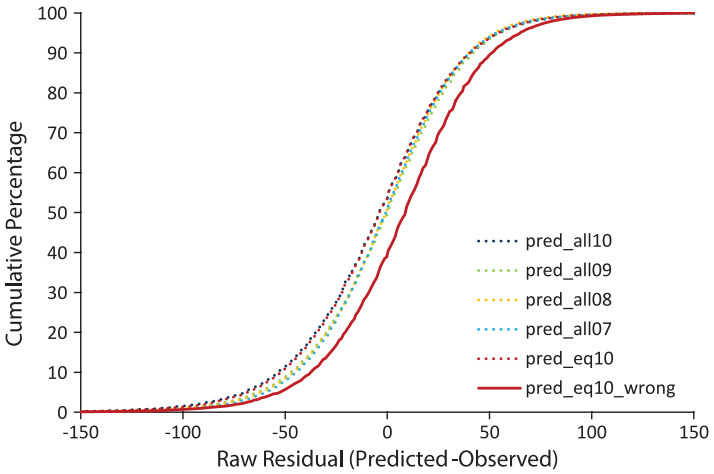

Figure 20 Cumulative distribution of residuals for predicting Grade 7 math.

function (*pred_all08*), and the prediction of 2007 Grade 3 ELA using the 2005–2006 regression function (*pred_all07*). These results are displayed along with the prediction results based on the 2010 equating sample with two conditions applied: with the correct conversion applied to the 2010 test (*pred_eq10*) and with the wrong conversion table applied to the 2010 test (*pred_eq10_wrong*). In the plots, the results based on the wrong conversion tables are represented by solid lines. The cumulative distributions of the residuals for the two predictions using the 2010 equating sample were visually compared to the historical prediction residual distributions (i.e., *pred_all09*, *pred_all08*, and *pred_all07*) for a reasonableness check. The ones that showed noticeable differences from the historical patterns were flagged in Table 3 under the heading of *Resid CDF*. The results were consistent with the earlier analysis results in that the majority of the wrong conversion tables were flagged. None of the correct conversion tables were flagged, indicating a minimum Type I error rate.

To summarize the prediction residuals, the RMSD of the predicted values and observed values were calculated and presented in Appendix B for the overall group, as well as for subgroups categorized based on the predictor variable, with Group 1 consisting of students with low ability (predictor scores below 250), Group 2 consisting of students with medium ability (predictor scores between 250 and 450), and Group 3 consisting of students with high ability (predictor scores above 450). Sample sizes for the overall group and each examinee subgroup are also presented in these tables. Note that while a smaller RMSD indicates more accurate predictions, it does not necessarily mean that a more accurate conversion table has been applied. This is due to the change of the distribution of dependent variables. In most model application studies, the datasets are fixed with the statistical model being the varying component. In this study, a fixed linear regression model is applied to varying datasets with different conversions applied. And RMSDs are determined not only by model-data fit but also by the variance of the dependent variable itself. The interpretation of the RMSD was also complicated by the confounding effect of change in growth pattern and equating error. For instance, if the regression function used underestimates the students' performance by 5 score points on average due to more growth made by the current testing population, and if there is a systematic error of positive 5 scale score point assigned to each student, then the conversion table with the error will lead to a smaller RMSD compared to a conversion table without the error. Given these considerations, RMSDs should be interpreted with caution. Instead of RMSDs that take on small values, RMSDs that fall within a normal range as defined by historical data provide validity evidence for equating results.

A rule that was similar to what was used to identify regression parameter estimates outside of the historical range was used here to identify unusual RMSDs, either at the overall or subgroup level. The mean and standard deviation were calculated for RMSDs based on historical data, and the historical range was defined to be the mean plus and minus 3 standard deviations. As an example, consider the evaluation of the 2010 Grade 3 ELA equating results. Table B1 shows the RMSDs for predicting 2009 (*pred_all09*), 2008 (*pred_all08*), and 2007 (*pred_all07*) scores, as well as the historical range of RMSDs defined by these values. The RMSDs for predicting 2010 scores for the equating sample were flagged if they fell outside of the historical range. And a conversion table was flagged if the overall group level RMSD and/or the subgroup level RMSD was flagged. Table 3 records all conversions that were flagged due to unusual RMSDs under *RMSD*. The results

were consistent with the earlier analysis results in that the majority of the wrong conversion tables were flagged. Only one correct conversion table was flagged.

In comparing the procedures that examine regression functions (i.e., param and curve) with those examining the scores resulting from applying the regression functions (i.e., resid CDF and RMSD), very similar results have been produced. It is hard to identify one type of procedure as performing better than the other type. In fact, all the four procedures produced similar results, as shown in Table 3. Type I error rate was consistently low. None or only one of the correct conversions was flagged by each procedure. The procedures that involve hypothesis testing (i.e., param and RMSD) were slightly more powerful than that of the others, but they are also associated with slightly higher Type I error rate.

As with hypothesis testing, the procedures described in this study were found to be more powerful in detecting errors in scoring tables when the errors are of more than negligible size. Tables 2 and 3 were evaluated together to establish the relationship between the degree of deviation of the wrong conversion tables from the correct conversion tables and capability of the procedures to detect conversion table errors. It was found that all wrong conversion tables with root mean squared scoring deviations over 7 points, or with standardized deviation over 0.12 points, were detected by one or more procedures used in this study. The concept of standardized RMSD was similar to that of the effect size for hypothesis testing. It can be concluded that regardless of the score scale the investigation is based upon, scoring errors with standardized RMSD of over 0.12 could be effectively detected using procedures described in this study. With all other factors held constant, the procedures were more powerful when the standardized RMSD was larger.

Recommendations and Limitations

The use of longitudinal regression in quality control is quite effective in detecting scoring errors, especially when the scoring errors are non-negligible (i.e., with standardized RMSD of 0.12 or above). Given that the results produced by all four procedures were quite consistent, and that score reporting usually has tight timelines, it is suggested that one or two instead of all four procedures be implemented operationally as part of the quality control process. For example, if an automated flagging procedure was desired, then a selection from the numerical param or RMSD procedures would be appropriate. It is recommended that any automated procedure be augmented by a graphical procedure (curve or resid CDF) that can be readily visually evaluated for magnitude by human (psychometric) review.

It should be noted that the procedures that examine the scores resulting from applying the regression functions set a higher requirement on the number of years of historical data needed to conduct analyses. For instance, for a quality control process of 2010 equating results, regression comparison procedures (i.e., param and curve) require longitudinal data from 2008 to 2010 at least, while the residual distribution comparison procedures (resid CDF and RMSD) require longitudinal data from 2007 to 2010 at least. For a testing program with limited historical data that can be linked longitudinally, it may be simpler and more straightforward to use the present and past regression curve comparison. Depending on scheduling and data source availability, other procedures may be conducted to verify the results of the regression curve comparison procedure.

Compared to the traditional equating quality control procedures, the procedures described in this study are more focused on examining the reasonableness of the observed progression patterns for examinees at all ability levels. Therefore, some type of scoring errors, especially nonuniform errors, may be able to be caught by procedures described in this study but not the traditional procedures. For instance, a wrong scoring table may have positive bias (i.e., with assigned score higher than the correct score) for examinees of low ability and negative bias (i.e., with assigned score lower than the correct score) for examinees of high ability. It may happen that these biases can be canceled out and that the wrong scoring table can still produce an average score mean for the new administration that is consistent with the historical trend in means. Such a case, however, will be flagged due to a lower-than-expected regression slope or out-of-range residual distribution by the longitudinal quality control procedures.

One limitation of the procedures described in this study is that they may not be able to detect minor scoring errors, given that historical patterns of accurate regressions contain some variation. In other words, if a scoring error is smaller than what would be expected within a normal range of variation due to year-to-year growth differences, it will not be detected using the procedure described in this study. The power of these procedures may be improved when there are multiple years of historical data. While the historical data serves as the reference to judge the reasonableness of the new test results, the enlargement of the historical pool improves the effectiveness of the flagging criteria.

This study focused on a simple linear regression method, which is probably the easiest way to make use of longitudinal data for quality control purpose. Variations of the method may be investigated in future research to see how powerful they are in detecting scoring errors. For instance, multiple prior years' assessment scores may be used as predictors. As an example, the regression function developed based on predicting 2009 Grade 5 math scores from 2008 Grade 4 and 2007 Grade 3 scores can be used to predict a student's 2010 Grade 5 math score, given his or her 2009 Grade 4 and 2008 Grade 3 scores. While multiple predictors make it difficult to present and compare regression curves, it may be possible to use a single predictor that incorporates information from multiple prior years' scores. The single predictor may be the average of 2 prior years' scores, or the weighted average of 2 prior years' scores with weights determined arbitrarily or based on multiple regression results. As another variation to simple linear regression, nonlinear regression may be applied if it is found to improve model-data fit significantly.

For testing programs that do not have longitudinal data, the described approach might be applied using demographic variables as predictor variables. The relationship between demographic variables and test scores may be determined from one testing cycle and then be used to predict results for another testing cycle and determine if the relationship is sufficiently consistent and strong as to be useful for identifying quality control issues. It is acknowledged, however, that test scores are likely to have much weaker relationships to demographic variables than to longitudinal test scores.

Note

1 It may also be possible to use past regression functions with more than 1 year's lag for prediction. For instance, the regression function developed based on predicting 2007 Grade 5 math scores from 2006 Grade 4 math scores can be used to predict a student's 2010 Grade 5 math score given his or her 2009 Grade 4 score. However, this option is not pursued in this study, as student progression patterns are expected to be more similar in adjacent years and, therefore, in this study the regression function with only 1 year's lag is expected to be more accurate for prediction.

References

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, 26, 36–46.

Kolen, M. J., & Brennan, R. L. (2004). *Testing equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Appendix A

Linear Regression Comparison Results

Table A1 Linear Regression of English-Language Art (ELA) Grade 3 Scores From ELA Grade 2 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	413153	43.55	0.35	0.84	0.00	37	0.65
	all0809	416,814	35.78	0.35	0.87	0.00	37	0.66
	all0708	412,723	65.34	0.31	0.77	0.00	34	0.65
	all0607	409,827	69.52	0.29	0.76	0.00	33	0.67
	all0506	413,438	53.86	0.31	0.82	0.00	35	0.67
	Historical range		(10.84, 101.41)		(0.65, 0.96)			
Equating samples	*Correct conversion*							
	eq0910	89,933	42.83	0.74	0.84	0.00	37	0.65
	Wrong conversion							
	eq0910_wrong	89,933	72.79	0.70	0.79	0.00	35	0.65

Table A2 Linear Regression of English-Language Art (ELA) Grade 6 Scores Over ELA Grade 5 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	411,349	67.34	0.29	0.81	0.00	29	0.71
	all0809	415,280	47.71	0.31	0.88	0.00	29	0.71
	all0708	438,836	57.35	0.27	0.84	0.00	27	0.73
	all0607	407,631	61.61	0.26	0.81	0.00	27	0.74
	all0506	410,005	51.77	0.27	0.84	0.00	28	0.74
	Historical range		(36.27, 72.95)		(0.77, 0.92)			
Equating samples	*Correct conversion*							
	eq0910	87,029	70.70	0.63	0.81	0.00	28	0.71
	Wrong conversion							
	eq0910_wrong	87,029	79.54	0.60	0.77	0.00	27	0.71

Table A3 Linear Regression of English-Language Art (ELA) Grade 9 Scores Over ELA Grade 8 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	419,927	73.62	0.29	0.81	0.00	31	0.71
	all0809	436,469	51.41	0.28	0.88	0.00	31	0.71
	all0708	432,372	47.00	0.29	0.90	0.00	32	0.73
	all0607	407,432	40.51	0.31	0.91	0.00	31	0.74
	all0506	413,290	16.07	0.32	0.97	0.00	33	0.74
	Historical range		(−8.55, 86.05)		(0.79, 1.04)			
Equating samples	*Correct conversion*							
	eq0910	72,246	71.28	0.71	0.82	0.00	31	0.71
	Wrong conversion							
	eq0910_wrong	72,246	90.17	0.65	0.75	0.00	29	0.71

Table A4 Linear Regression of English-Language Art (ELA) Grade 10 Scores Over ELA Grade 9 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	438,053	38.22	0.30	0.86	0.00	32	0.71
	all0809	433,180	44.78	0.29	0.84	0.00	32	0.72
	all0708	430,774	32.54	0.30	0.87	0.00	32	0.72
	all0607	416,146	56.56	0.27	0.80	0.00	31	0.72
	all0506	417,959	29.83	0.29	0.87	0.00	32	0.73
	Historical range		(4.07, 77.78)		(0.75, 0.95)			
Equating samples	*Correct conversion*							
	eq0910	74,055	36.57	0.73	0.86	0.00	32	0.71
	Wrong conversion							
	eq0910_wrong	74,055	22.16	0.81	0.96	0.00	35	0.71

Table A5 Linear Regression of English-Language Art (ELA) Grade 11 Scores Over ELA Grade 10 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	410,771	13.22	0.34	0.95	0.00	36	0.70
	all0809	404,134	13.77	0.33	0.94	0.00	36	0.71
	all0708	403,537	11.33	0.32	0.95	0.00	35	0.71
	all0607	385,804	−1.17	0.35	0.99	0.00	38	0.71
	all0506	365,705	−15.74	0.38	1.03	0.00	39	0.69
	Historical range		(−38.58, 42.68)		(0.86, 1.10)			
Equating samples	*Correct conversion*							
	eq0910	70,212	14.48	0.80	0.95	0.00	35	0.70
	Wrong conversion							
	eq0910_wrong	70,212	52.25	0.70	0.83	0.00	31	0.70

Table A6 Linear Regression of Math Grade 3 Scores Over Math Grade 2 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	415,260	58.30	0.48	0.88	0.00	61	0.56
	all0809	418,896	64.53	0.45	0.86	0.00	59	0.57
	all0708	414,733	79.80	0.43	0.80	0.00	57	0.56
	all0607	411,357	88.85	0.39	0.76	0.00	55	0.58
	all0506	414,427	84.54	0.39	0.78	0.00	54	0.58
	Historical range		(47.63, 111.22)		(0.67, 0.93)			
Equating samples	*Correct conversion*							
	eq0910	90,463	55.67	1.02	0.89	0.00	61	0.56
	Wrong conversion							
	eq0910_wrong	90,463	49.86	1.03	0.89	0.00	61	0.56

Table A7 Linear Regression of Math Grade 4 Scores Over Math Grade 3 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	417,776	129.96	0.37	0.67	0.00	52	0.56
	all0809	411,531	116.29	0.36	0.70	0.00	50	0.58
	all0708	418,943	114.15	0.34	0.69	0.00	48	0.59
	all0607	418,649	115.63	0.32	0.68	0.00	46	0.61
	all0506	424,949	91.67	0.34	0.75	0.00	46	0.62
	Historical range		(73.80, 145.07)		(0.62, 0.79)			
Equating samples	*Correct conversion*							
	eq0910	90,473	129.70	0.77	0.66	0.00	51	0.56
	Wrong conversion							
	eq0910_wrong	90,473	85.16	0.91	0.80	0.00	60	0.58

Table A8 Linear Regression of Math Grade 5 Scores Over Math Grade 4 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	413,019	26.08	0.46	0.93	0.00	57	0.61
	all0809	416,276	9.07	0.45	0.98	0.00	56	0.62
	all0708	425,152	11.73	0.43	0.96	0.00	53	0.63
	all0607	428,984	10.64	0.39	0.96	0.00	51	0.66
	all0506	435,724	−9.12	0.41	1.04	0.00	52	0.66
	Historical range		(−24.00, 35.16)		(0.87, 1.09)			
Equating samples	*Correct conversion*							
	eq0910	89,561	23.63	0.99	0.93	0.00	57	0.60
	Wrong conversion							
	eq0910_wrong	89,561	72.51	0.78	0.72	0.00	45	0.59

Table A9 Linear Regression of Math Grade 6 Scores Over Math Grade 5 Scores

			B0		B1			
Samples	Regression	*N*	Par	*SE*	Par	*SE*	*RMSE*	R^2
Complete samples	all0910	412,605	108.83	0.29	0.67	0.00	43	0.66
	all0809	415,901	93.43	0.29	0.71	0.00	43	0.67
	all0708	440,135	104.01	0.25	0.68	0.00	39	0.70
	all0607	409,016	116.79	0.24	0.64	0.00	38	0.70
	all0506	410,371	119.35	0.24	0.64	0.00	38	0.69
	Historical range		(72.32, 144.47)		(0.56, 0.78)			
Equating samples	*Correct conversion*							
	eq0910	87,305	108.33	0.62	0.67	0.00	43	0.67
	Wrong conversion							
	eq0910_wrong	87,305	66.95	0.74	0.81	0.00	51	0.67

Table A10 Linear Regression of Math Grade 7 Scores Over Math Grade 6 Scores

Samples	Regression	*N*	B0		B1		*RMSE*	R^2
			Par	*SE*	Par	*SE*		
Complete samples	all0910	383,656	61.46	0.33	0.84	0.00	39	0.69
	all0809	397,651	52.52	0.32	0.85	0.00	37	0.69
	all0708	413,119	47.88	0.30	0.86	0.00	35	0.70
	all0607	409,390	51.89	0.30	0.85	0.00	36	0.69
	all0506	428,982	57.01	0.27	0.83	0.00	36	0.72
	Historical range		(41.10, 63.55)		(0.81, 0.89)			
Equating samples	*Correct conversion*							
	eq0910	68,280	64.61	0.78	0.83	0.00	38	0.68
	Wrong conversion							
	eq0910_w2_M06_10	68,280	64.63	0.74	0.79	0.00	36	0.68

Appendix B

Summary of Prediction Residuals

Table B1 Summary of Prediction Residuals for English-Language Art (ELA) Grade 3

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	413,153	37	10,548	33	373,629	36	28,976	55
	pred_all09	416,814	37	12,318	31	385,128	37	19,368	50
	pred_all08	412,723	34	17,447	30	372,817	33	22,459	47
	pred_all07	409,827	34	18,739	31	365,413	33	25,675	52
	Historical range		(30, 41)		(30, 31)		(28, 41)		(42, 58)
Equating samples	*Correct conversion*								
	pred_eq10	89,933	37	2,575	32	81,645	35	5,713	54
	Wrong conversion								
	pred_eq10_wrong	89,933	36	2,575	39	81,645	35	5,713	51

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B2 Summary of Prediction Residuals for English-Language Art (ELA) Grade 6

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	411,349	29	8,473	31	380,902	28	21,974	42
	pred_all09	415,280	29	7,329	29	398,006	28	9,945	40
	pred_all08	438,836	28	11,604	27	418,922	28	8,310	38
	pred_all07	407,631	27	13,265	27	379,775	26	14,591	39
	Historical range		(25, 30)		(24, 31)		(25, 30)		(36, 41)
Equating samples	*Correct conversion*								
	pred_eq10	87,029	29	1,877	31	80,705	28	4,447	41
	Wrong conversion								
	pred_eq10_wrong	87,029	28	1,877	30	80,705	27	4,447	43

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B3 Summary of Prediction Residuals for English-Language Art (ELA) Grade 9

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	419,927	32	14,797	34	386,131	31	18,999	43
	pred_all09	436,469	31	20,774	31	401,987	31	13,708	39
	pred_all08	432,372	32	21,282	32	400,272	31	10,818	41
	pred_all07	407,432	32	13,572	34	383,964	31	9,896	43
	Historical range		(30, 33)		(27, 38)		(30, 32)		(36, 46)
Equating samples	*Correct conversion*								
	pred_eq10	72,246	32	2,520	33	67,129	31	2,597	43
	Wrong conversion								
	pred_eq10_wrong	72,246	31	2,520	32	67,129	29	2,597	50

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B4 Summary of Prediction Residuals for English-Language Art (ELA) Grade 10

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	438,053	32	12,904	33	405,149	32	20,000	39
	pred_all09	433,180	32	12,071	35	398,512	31	22,597	40
	pred_all08	430,774	32	13,374	30	397,631	32	19,769	42
	pred_all07	416,146	31	22,681	37	376,923	30	16,542	40
	Historical range		(30, 34)		(23, 45)		(28, 34)		(38, 44)
Equating samples	*Correct conversion*								
	pred_eq10	74,055	32	2,282	32	69,109	32	2,664	38
	Wrong conversion								
	pred_eq10_wrong	74,055	42	2,282	39	69,109	41	2,664	49

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B5 Summary of Prediction Residuals for English-Language Art (ELA) Grade 11

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	410,771	36	24,497	40	375,237	36	11,037	47
	pred_all09	404,134	36	24,596	37	366,689	36	12,849	47
	pred_all08	403,537	35	21,659	36	372,566	35	9,312	50
	pred_all07	385,804	39	27,236	41	348,766	38	9,802	52
	Historical range		(31, 42)		(31, 45)		(31, 41)		(43, 56)
Equating samples	*Correct conversion*								
	pred_eq10	70,212	35	4,440	38	64,278	35	1,494	47
	Wrong conversion								
	pred_eq10_wrong	70,212	32	4,440	39	64,278	31	1,494	50

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B6 Summary of Prediction Residuals for Math Grade 3

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	415,260	61	13,675	44	317,428	59	84,157	71
	pred_all09	418,896	59	18,617	42	327,651	57	72,628	70
	pred_all08	414,733	58	19,339	43	325,869	56	69,525	68
	pred_all07	411,357	55	23,970	40	312,551	53	74,836	66
	Historical range		(50, 64)		(38, 45)		(48, 62)		(63, 74)
Equating samples	*Correct conversion*								
	pred_eq10	90,463	61	3,081	43	70,237	58	17,145	72
	Wrong conversion								
	pred_eq10_wrong	90,463	61	3,081	42	70,237	59	17,145	73

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B7 Summary of Prediction Residuals for Math Grade 4

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	417,776	52	14,678	37	312,287	49	90,811	63
	pred_all09	411,531	50	16,363	35	302,250	47	92,918	61
	pred_all08	418,943	48	21,067	33	329,321	45	68,555	62
	pred_all07	418,649	46	25,435	33	322,922	43	70,292	61
	Historical range		(42, 54)		(30, 37)		(39, 51)		(59, 64)
Equating samples	*Correct conversion*								
	pred_eq10	90,473	51	3,443	35	68,465	48	18,565	62
	Wrong conversion								
	pred_eq10_wrong	90,473	61	3,443	50	68,465	60	18,565	65

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B8 Summary of Prediction Residuals for Math Grade 5

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	413,019	57	9,411	44	328,682	54	74,926	70
	pred_all09	416,276	56	8,373	40	350,021	54	57,882	69
	pred_all08	425,152	53	12,368	38	361,602	51	51,182	69
	pred_all07	428,984	52	19,616	37	362,789	49	46,579	70
	Historical range		(46, 61)		(33, 44)		(44, 59)		(68, 71)
Equating samples	*Correct conversion*								
	pred_eq10	89,561	57	2,190	42	72,156	54	15,215	71
	Wrong conversion								
	pred_eq10_wrong	89,561	61	2,190	32	72,156	53	15,215	92

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B9 Summary of Prediction Residuals for Math Grade 6

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	412,605	44	21,367	34	303,842	41	87,396	54
	pred_all09	415,901	43	20,617	32	320,127	40	75,157	55
	pred_all08	440,135	39	38,655	29	331,372	37	70,108	52
	pred_all07	409,016	38	41,689	29	302,768	36	64,559	51
	Historical range		(32, 48)		(24, 36)		(31, 45)		(46, 60)
Equating samples	*Correct conversion*								
	pred_eq10	87,305	43	4,723	35	65,110	40	17,472	54
	Wrong conversion								
	pred_eq10_wrong	87,305	53	4,723	41	65,110	50	17,472	64

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Table B10 Summary of Prediction Residuals for Math Grade 7

Samples predicted	Variables predicted	All		Group 1		Group 2		Group 3	
		N	*RMSD*	*N*	*RMSD*	*N*	*RMSD*	*N*	*RMSD*
Complete samples	pred_all10	383,656	39	17,761	34	333,615	38	32,280	53
	pred_all09	397,651	37	17,808	31	353,431	36	26,412	53
	pred_all08	413,119	35	24,389	30	367,530	35	21,200	51
	pred_all07	409,390	36	24,607	29	366,530	35	18,253	59
	Historical range		(34, 38)		(26, 34)		(34, 37)		(42, 67)
Equating samples	*Correct conversion*								
	pred_eq10	68,280	38	3,225	33	59,724	37	5,331	52
	Wrong conversion								
	pred_eq10_wrong	68,280	38	3,225	30	59,724	36	5,331	54

Note. Group 1 has predictor scores below 250. Group 2 has predictor scores between 250 and 450. Group 3 has predictor scores above 450.

Suggested citation:

Lu, Y., & Yen, W. M. (2014). *Use of longitudinal regression in quality control* (ETS Research Report No. RR-14-31). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12032

Action Editor: Gautam Puhan

Reviewers: Hongwen Guo and Samuel Livingston