JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
DOI: 10.1111/jedm.12228
Journal of Educational Measurement
Fall 2019, Vol. 56, No. 3, pp. 626–652

Scoring Stability in a Large-Scale Assessment Program: A Longitudinal Analysis of Leniency/Severity Effects

Corey Palermo, Michael B. Bunch, and Kirk Ridge
Measurement Incorporated

Although much attention has been given to rater effects in rater-mediated assessment contexts, little research has examined the overall stability of leniency and severity effects over time. This study examined longitudinal scoring data collected during three consecutive administrations of a large-scale, multi-state summative assessment program. Multilevel models were used to assess the overall extent of rater leniency/severity during scoring and examine the extent to which leniency/severity effects were stable across the three administrations. Model results were then applied to scaled scores to estimate the impact of the stability of leniency/severity effects on students' scores. Results showed relative scoring stability across administrations in mathematics. In English language arts, short constructed response items showed evidence of slightly increasing severity across administrations, while essays showed mixed results: evidence of both slightly increasing severity and moderately increasing leniency over time, depending on trait. However, when model results were applied to scaled scores, results revealed rater effects had minimal impact on students' scores.

For hand-scored responses, the validity of scores depends on both inter- and intrarater accuracy and consistency. Much attention has been given to rater effects such as leniency and severity (Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980; Wolfe, 2004). Raters have also been found to demonstrate halo (Ridge, 2001a, 2001b; Rudner, 1992), central tendency (Engelhard, 1994), and sequential effects (Attali, 2011; Zhao, Andersson, Guo, & Xin, 2017) during scoring. Researchers have examined the impact of rater background characteristics, such as demographics and experience (Song et al., 2014), and rater training methods (Raczynski, Cohen, Engelhard & Lu, 2015; Wolfe, Matthews, & Vickers, 2010) on scoring accuracy.

Less attention has been given to the stability of rater effects over time. Lunz and O'Neill (1997) analyzed 10 years of score data from the American Society of Clinical Pathologists' histology examinations to examine rater leniency and consistency across administrations. There was sufficient overlap across administrations for the authors to use a multifaceted Rasch model (MFRM; Linacre, 1989) to calibrate all raters, projects, tasks, and examinees onto a benchmark scale. Raters varied in their extent of leniency but were relatively consistent in the application of this leniency across administrations.

Myford and Wolfe (2009) used a MFRM to examine rater accuracy (which the authors defined as the degree of consistency between a given rater's rank ordering of students' performances and the rank ordering provided by the rest of the raters) and central tendency during essay scoring associated with the College Board's 2002 Advanced Placement English Literature and Composition Examination. Raters

scored benchmark essays twice per day over 4 days of scoring. Most raters exhibited little change in scale category use and accuracy over time. Those who did show evidence of change tended to become more consistent in their use of scale categories and more accurate over time. Findings are limited by the fact that the benchmark essays were not administered inconspicuously to raters, that is, they visibly differed from operational responses.

Leckie and Baird (2011) examined raters' severity and central tendency during essay scoring associated with England's 2008 national curriculum English writing test. Raters scored sets of benchmark essays, selected and prescored by an expert committee, at the completion of training and then periodically throughout operational scoring. The authors used cross-classified multilevel models (MLMs) to measure raters' overall leniency/severity (defined as the extent to which raters over- or underscored essays relative to the expert committee consensus scores) and examine the stability of raters' leniency/severity over time. Raters' overall leniency/severity was relatively stable over time, although the authors found significant within-rater variability in leniency/severity from check to check, suggesting that rater leniency/severity was not stable and could be influenced by essays. Raters also exhibited central tendency effects. Findings were limited by the fact that check essays were presented to raters onscreen while operational essays were paper-based, so raters were aware of the monitoring tool. Despite this limitation, this study is noteworthy due to the number of raters involved (689) and the fact that raters scored common essays at each check.

Several researchers have used MFRM to examine rater drift and rater accuracy drift in the context of music performance assessment. Wind and Wesolowski (2018) evaluated rater accuracy across 5 days of ratings, using criterion-referenced accuracy indicators derived by comparing operational and expert ratings. Results showed that on average rater accuracy decreased over time, and further varied within domains over time, indicating that raters found particular components of performance more/less difficult to score across the 5 days of ratings. In a related study, Wesolowski, Wind, and Engelhard (2017) examined raters' differential leniency/severity across 5 days of ratings. Raters exhibited decreasing severity over time, and results further showed between- and within-rater differences in leniency/severity over time.

Zupanc and Štrumbelj (2018) used a Bayesian hierarchical model to investigate rater bias and variability over 5 years of essay scoring associated with Slovenia's nationwide external examination. Two raters evaluated each essay using separate rubrics for content, syntax, style, and structure. Essays could earn a maximum of 50 points across rubrics; a third rater provided a resolution score if the two raters' scores differed by more than 10 points. Results indicated within-year rater bias and variability effects and showed relative stability of effects across years, suggesting that rater effects partially persist over administrations. The authors estimated that 10% of essays were assigned a score $\pm$ five points of the true score, which equated to 3.6% of the total exam grade.

In sum, researchers have adopted a range of approaches to investigate scoring stability in rater-mediated assessment contexts. These can be broadly grouped as calibration approaches (e.g., MFRM), sampling approaches (e.g., generalizability studies), and explanatory approaches (e.g., MLMs). Each approach offers unique

affordances (and imposes unique limitations) to examine rater effects: MFRM allows raters and responses to be placed on a common scale; generalizability studies isolate error variance by source and combinations of sources (e.g., raters, responses, and Raters × Responses); and MLMs structure and model complex data to explain the variance components attributable to each level of the data hierarchy.[1]

Despite the variety of approaches, previous research examining the stability of rater effects over time supports a limited understanding of the stability of scoring across assessment program administrations. Most related work has examined a small number of raters, a small pool of items and responses, and a short period of time. In an age of assessment consortia and computer-adaptive assessment, the reality of large-scale summative assessment scoring in the United States involves thousands of raters, thousands of hand-scored items, and millions of hand-scored student responses. In order to explain the overall extent of rater leniency/severity and determine the extent to which these effects were stable across administrations of a large-scale assessment program, we adopted MLM analysis techniques to model raters' scores and the responses and items with which they were associated.

The Present Study

The present study analyzed scoring data associated with large-scale summative assessments in the United States. Data were collected across three operational scoring administrations each spring from 2016 to 2018. Validity responses (i.e., expert-scored benchmark responses selected to represent the range of responses that raters encounter operationally) from a common pool were distributed to raters throughout the three operational scoring administrations. We were thus able to use the difference between the scores raters assigned to the validity responses throughout operational scoring and the benchmark scores to determine leniency/severity (i.e., the extent to which responses were assigned lower or higher scores than warranted given an external criterion of performance; see Saal, Downey, & Lahey, 1980). The present study addressed three research questions:

1. What was the overall extent of rater leniency/severity during scoring?
2. To what extent were leniency/severity effects stable across the three administrations?
3. What was the impact of the stability of leniency/severity effects on students' scaled scores?

Methods

The Summative Assessments

The summative assessments with which the scoring data are associated were developed by a multi-state assessment consortium. The summative assessments were designed to measure students' progress toward college and career readiness in English language arts (ELA) and mathematics. Member states administered some form of the summative assessments at the end of each school year. Assessments emphasize deep knowledge of core concepts and ideas within and across disciplines, analysis, synthesis, problem solving, communication, and critical thinking

(Darling-Hammond, 2010). The assessments include two parts: a computer-adaptive test (CAT) and a performance task (PT); the latter requires students to produce source-based writing and solve multi-step, real-world mathematics problems. Test items are aligned with the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) via content claims (reading, writing, listening, and research in ELA; concepts and procedures, problem solving, modeling and data analysis, and communicating reasoning in mathematics) and assessment targets, following an evidence-centered design (Mislevy, Steinberg, & Almond, 2003) approach. Items include such response types as multiple choice, multi-select, matching, table, equation, evidence-based selected response, short answer, and writing extended response. Summative assessment results describe student achievement and growth and are used to inform school, district, and state accountability systems.

The CAT part of the assessments demands a much larger item pool than a test that presents the same items to every examinee. In order to vary item difficulty during testing to adapt to a student's current performance, a CAT requires a sufficiently large pool of items so that every content category has an array of items of varying difficulty. For example, a fixed form test assessing 12 content standards might have a pool consisting of five items per content standard for a total of 60 items. A computer-adaptive version of that same test might require a pool of 20–30 items of widely varying difficulty for each standard, for a total of 240–360 items, even though any one student might only encounter 30–40 of those items. This requirement extends to human-scored, constructed-response items as well as machine-scored, selected response items.

In a computer-adaptive environment, each student takes a test tailored to his or her ability level. This tailoring means it is theoretically possible for every student to take a unique version of the test. Without a common set of items by which to compare student performances, the need for comparability of assessment versions within and consistency across years is paramount.

The ELA and mathematics summative assessments include hand-scored items in both the CAT and PT parts of the assessments.[2] Specifically, the ELA CAT includes hand-scored, constructed response items aligned with the reading and writing content claims, and the ELA PT includes hand-scored, constructed response and essay items aligned with the research and writing content claims. The mathematics CAT and PT include constructed response hand-scored items aligned with the content claims of problem solving, modeling and data analysis, and communicating reasoning.

Hand-Scoring Sources of Variance

There are numerous potential sources of variance associated with the process of hand-scoring responses to open-ended items. Scoring rubrics are written at the time of item development, based on content standards. During rangefinding, experts then interpret and operationalize rubrics by applying them to student responses. Ideally, the responses reviewed will reflect the full depth and breadth of responses raters will encounter operationally; however, there are many practical factors that

narrow the range. As responses vary in scoring complexity (Engelhard, 1996), rangefinding can involve considerable discussion and negotiation. Ultimately, even experts do not consistently agree on the true score of responses (Sulsky & Balzer, 1988).

Following rangefinding, scoring leaders prepare and group the expert-scored responses for rater training. During this process, scoring leaders make decisions about response selection, presentation sequence, and scoring rationale.

Raters come to a scoring project with varying backgrounds, education, and potential biases. Furthermore, raters vary in how they apply the training materials when scoring operational responses, and may differ in their interpretations of and reactions to feedback during training and scoring (Eckes, 2008). Some responses may reflect unanticipated student approaches. Operational scoring is a repetitive and often cognitively demanding activity, and rater attitude and fatigue may impact scoring accuracy (Lunz & Stahl, 1990).

Procedure

Rater recruitment. Each year, prospective raters were recruited from two pools: (1) a pool of experienced raters and (2) a pool of new applicants. Priority was given to experienced raters. All prospective raters had a minimum of a 4-year college degree, successfully completed an interview, and provided references. Scoring data were collected on total of 6,191 raters during the three administrations. Approximately 58% of raters scored during a single administration, 30% of raters scored during two administrations, and 12% scored during all three administrations.

Training materials. Training materials were a product of rangefinding meetings held to reach consensus on the scoring of student responses. Participants included educators, consortium staff, and service provider staff with hand-scoring expertise. During rangefinding, participants identified the full range of student responses that represented each score point. Following rangefinding, the consensus-scored responses were arranged into training materials that included the scoring rubric for each item, an anchor set of annotated exemplar responses at each score point, practice sets, and qualifying sets.

Rater training and qualification. Training was designed to communicate the scoring decisions made during rangefinding to raters with the goal of maximizing the reliability and validity of operational scoring. Training was primarily self-paced; the scoring system required that raters complete all training, practice, and qualifying activities sequentially. Training duration varied based on content area and item type, although all training followed a standard process. Prospective raters first studied the item/prompt and any source materials, each score point of the rubric, and the underlying criteria. Following, individuals studied various applications of each score point via an anchor set, learning the characteristics of responses associated with each score point. Anchor responses were selected to exemplify the diversity of responses raters would encounter in operational scoring and included responses that were close to the score point lines, that is, included characteristics of more than one score point. Prospective raters were counseled to weigh elemental strengths and weaknesses in

light of the overall effectiveness of a response in determining the correct score point. Training examples further included atypical student responses; raters were advised on how to evaluate anomalous task approaches based on the rubric and decisions made by the rangefinding committee. Prospective raters then applied what they had learned by assigning scores to one or more practice sets of responses. Prospective raters were guided to compare each practice response with comparable anchor responses to ensure accuracy and consistency. After each practice set, individuals were led through a process of reviewing the detailed annotations of each practice response that identified the strengths and weaknesses therein. Next, prospective raters scored two qualifying sets to provide formal evidence of their ability to apply the scoring criteria reliably. Scoring leaders provided feedback following each qualifying set. Only those prospective raters who met the qualifying criteria established by the assessment consortium were eligible for operational scoring.

During training and throughout operational scoring, scoring leaders emphasized the scoring process as an application of the criteria defined by the rubrics and anchor responses that had been validated during rangefinding. This criterion-based process was essential to limiting common sources of rater bias such as severity and leniency, halo, and central tendency effects.

Rater monitoring and feedback. Evidence of raters' challenges during training and qualifying was used to focus initial monitoring efforts. Documentation included the characteristics of responses scored incorrectly during training as well as the nature of raters' errors. Scoring leaders used this information to establish a feedback loop with individual raters that would be maintained throughout the scoring effort.

Ongoing monitoring efforts were informed by raters' validity performance, score point distributions, and interrater reliability results, all of which were available to scoring leaders via a suite of real-time reports. Scoring leaders analyzed these data to evaluate accuracy, diagnose scoring issues, and provide feedback to raters. These data guided intervention strategies throughout operational scoring and informed the extent of monitoring and the type of feedback most appropriate for each rater. Further, these data were used to evaluate whether particular raters required dismissal or particular responses required rescoring.

Feedback provided to raters based on these data most frequently involved referring raters back to the essential criteria of the rubric, the appropriate anchor responses, and the precise scoring consideration for the unique aspects of the student response(s) in question. Additionally, scoring leaders addressed response-specific questions raised by raters during scoring. Any scoring decisions made when unique and/or unusual responses were encountered were immediately communicated to all raters.

As a monitoring and evaluation tool, validity responses provide an external frame of reference (Myford & Wolfe, 2009) by depicting rater performance against an external criterion. In the present study, the criterion was the consensus benchmark score of each validity response as determined by scoring leaders and/or consortium staff. The pool of validity responses represented all score points and a wide variety of response approaches for each item. Validity responses were selected to be representative of responses that raters encountered in operational scoring. To this end, the

validity pool included responses that were clear representations of score points as well as responses that were close to the score point lines.

The use of a secure, online scoring system allowed for validity responses to be embedded among operational responses and inconspicuously distributed to raters. During operational scoring, raters were assigned validity responses at a ratio of 1:20 operational responses. Validity responses were inserted randomly into scoring sets of operational responses, and raters had no way to distinguish between the two.

For the purpose of the present study, data associated with validity responses common to the 2016, 2017, and 2018 operational administrations were analyzed. The dependent variable in all analyses was the signed score difference between the score assigned by the rater and the benchmark score as determined by scoring leadership and consortium staff. A score difference of zero indicated the rater-assigned score matched the benchmark score of the response. A positive score difference indicated leniency, whereas a negative score difference indicated severity (Leckie & Baird, 2011). We examine all hand-scored items: ELA short writing task, research, and reading items scored on a 0–2 scale; ELA essays scored on three traits (1–4, 1–4, and 0–2); and mathematics items ranging from 0–1 to 0–3. Separate analyses were conducted based on the content area, type, and score point range of items to support interpretability of score differences.

Analyses

Due to the explanatory nature of our research questions, MLMs—also known as hierarchical linear models (Raudenbush & Bryk, 2002), random effects models, and mixed models—were used to estimate the overall extent of rater leniency/severity during scoring and determine the extent to which leniency/severity effects were stable across scoring administrations. Although MLMs assume normally distributed and homoscedastic random effects, MLMs avoid some of the more stringent assumptions of MFRM, in particular unidimensionality, equal discrimination, and response independence. MLMs account for the hierarchal nesting of residuals, in this case, the nesting of score differences within validity responses within items. In our data set, in which all items and validity responses were common across the three administrations, each score difference was associated with a validity response and each validity response was associated with an item. Thus, MLMs were used to accommodate a hierarchical data structure and adjust for and model observations that were not independent.

In our data set, repeated observations of the dependent variable (the raw score difference between the assigned and benchmark ratings of each validity response) represent Level 1 of the data hierarchy. Each score difference is associated with a validity response, representing Level 2 of the hierarchy. Finally, validity responses are nested within items at Level 3 of the data hierarchy. Figure 1 presents a classification diagram depicting the multilevel structure of the data. Model fit was examined using negative log likelihood (–2LL). Smaller values indicate better fit provided a significant chi-square difference test based on the number of estimated parameters. All analyses were conducted using MLwiN v3.02 (Charlton, Rasbash, Browne, Healy, & Cameron, 2017).

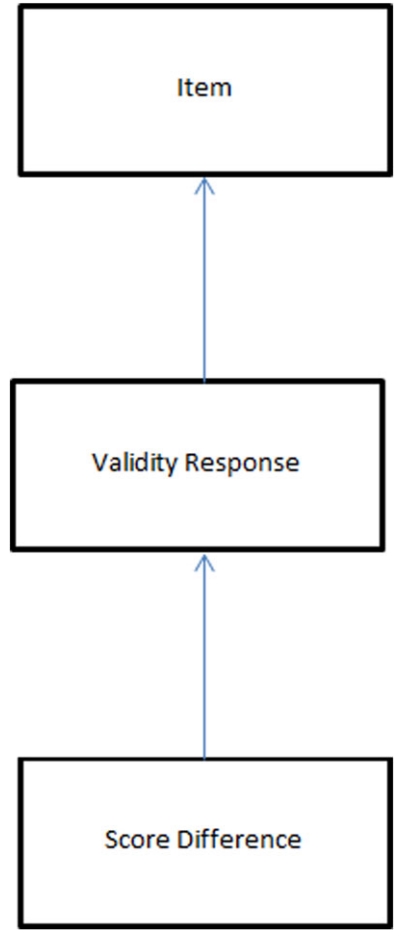

Figure 1. Classification diagram. (Color figure can be viewed at wileyonlinelibrary.com)

Analyses involved fitting a series of models for each item group. The first model was an unconditional model that ignored scoring administrations and assumed effects were stable across all 3 years:

$$y_{ij} = \beta_0 + u_j + e_{ij}. \quad (1)$$

In Equation 1, the dependent variable y_{ij} is the score difference between the assigned score and the true score for validity response i associated with item j. The intercept parameter β_0 provides the predicted mean score difference across items. This model also includes item random effects u_j and validity response random effects e_{ij}. The random effects decompose the variation in score differences into separate variance components, indicating the variation in score differences attributable to items (i.e., between validity response variation) and to validity responses (i.e., within

validity response variation). Both u_j and e_{ij} are assumed normally distributed with a mean of zero and variance σ^2. These assumptions were evaluated by examining (1) normal plots of standardized residuals and (2) standardized residuals plotted against predicted random effects. The unconditional model determined the overall extent of leniency/severity during scoring (research question 1) and served as the baseline against which subsequent conditional models were compared.

A second model (Model 2) added fixed and random scoring administration effects, allowing the predicted score differences to vary across the 3 years:

$$y_{ij} = \beta_0 + \beta_1 Year_{ij} + u_{0j} + u_{1j} Year_{ij} + e_{ij}. \quad (2)$$

In Equation 2, the intercept parameter β_0 represents score differences in 2016 (year = 0), while the slope β_1 represents the mean change in score differences per year in 2017 and 2018. This model includes random effects for items (u_{0j}), item intercept/slope covariance (u_{1j}), and validity responses (e_{ij}). Thus, this model determined the extent to which leniency/severity effects were stable across administrations (research question 2). In the case of the multi-trait ELA essays, this model was expanded to include additional terms (β_2–β_5) that tested for trait effects and trait × year interactions.

In presenting results, we address the substantive implications of findings in two ways. First, we report the proportional reduction in mean squared prediction error when comparing Models 1 and 2. This indicates the amount of residual variance explained by the addition of scoring administration effects to the model. Second, we apply model coefficients to scaled scores to estimate the impact of the stability of leniency/severity effects on students' scores (research question 3).

Results

Overall reliability (Pearson's correlation) between the raters' scores and the benchmark scores in mathematics was .94 for the 0–1 items, .96 for the 0–2 items, and .97 for the 0–3 items. In ELA, reliability for the short constructed response items was .86 for the research items, .89 for the short writing task items, and .93 for the reading items. Reliability for the essay traits was .91 for evidence/elaboration, .90 for development/elaboration, .91 for purpose/organization, and .86 for conventions.

Table 1 presents the descriptive statistics for study variables. This table provides counts of items (Level 3), validity responses (Level 2), and score differences (Level 1) for each item group. The number of items per item group, which ranges from 21 in ELA reading to 242 in ELA research, was a function of the assessment blueprints and adaptive algorithm. Across all item groups, there was a median of 33 validity responses per item. The median number of score differences per validity response was 106 in 2016, 43 in 2017, and 36 in 2018. This number decreased over time as additional validity responses were added to the pool in 2017 and 2018. Table 1 also provides score difference means and standard deviations, disaggregated by scoring administration.

The mean score differences for ELA items are larger in absolute value than those for mathematics items, suggesting overall greater rater error in the form of leniency/severity in ELA scoring. Items with larger score point ranges (i.e.,

Table 1
Descriptive Statistics for Study Variables

Content Area/Item Type/Trait	Score Point Range	Items (*n*)	Validity Responses (*n*)	Raters *n* 2016	Raters *n* 2017	Raters *n* 2018	Score Differences *n* 2016	Score Differences *n* 2017	Score Differences *n* 2018	Score Differences *M* (SD) 2016	Score Differences *M* (SD) 2017	Score Differences *M* (SD) 2018
Mathematics	0–1	73	1,822	733	957	720	110,641	119,729	108,587	−.009 (.166)	−.015 (.175)	−.017 (.174)
Mathematics	0–2	199	4,796	1,236	1,341	1,033	485,745	588,124	432,351	−.010 (.262)	−.009 (.250)	−.010 (.231)
Mathematics	0–3	24	527	276	474	378	58,028	56,119	56,317	.000 (.294)	.007 (.287)	.006 (.266)
ELA research	0–2	242	8,270	1,106	932	520	253,460	185,701	145,552	−.060 (.410)	−.096 (.445)	−.090 (.423)
ELA short writing task	0–2	105	3,515	415	338	274	96,325	47,620	42,207	−.006 (.366)	−.018 (.369)	−.033 (.375)
ELA reading	0–2	21	313	43	51	27	5,526	2,793	1,118	−.020 (.187)	−.039 (.238)	−.051 (.251)
ELA essay: evidence/elaboration	1–4	95	3,144	598	692	565	586,350	163,981	147,271	−.071 (.438)	−.095 (.435)	−.115 (.445)
ELA essay: development/elaboration	1–4	33	1,265	274	249	109	162,593	55,431	45,974	−.042 (.462)	−.041 (.458)	−.054 (.468)
ELA essay: purpose/organization	1–4	128	5,299	872	941	755	745,364	219,421	193,251	−.050 (.439)	−.070 (.440)	−.091 (.448)
ELA essay: conventions	0–2	128	5,289	872	941	755	747,882	219,022	192,578	.004 (.397)	.061 (.406)	.052 (.395)

Table 2
Unstandardized Coefficients (and Standard Errors) of Mathematics Constructed Response 1-Point Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	−.013	.007	<.001	−.010	.000	<.001
Year slope (β_1)				−.004	.000	<.001
Random effects						
Item intercept variance (σ^2_{u0})	.000	.000				
Item intercept/slope covariance (σ^2_{u01})						
Item slope variance (σ^2_{u1})						
Validity response intercept variance (σ^2_{e0})	.029	.000		.028	.000	
Validity response intercept/slope covariance (σ^2_{e01})				.002	.000	
Validity response slope variance (σ^2_{e1})				−.002	.000	
Goodness of fit						
Deviance (−2LL)	−234,322.331			−231,399.227		
Difference test				$\chi^2_5 = 2{,}923.104$		<.001

Note. *P*-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

ELA essay 4-point traits) have mean score differences that are relatively larger in absolute value, which is expected, given that wider score point ranges provide less opportunity for chance agreement. With the exception of ELA essay conventions and 3-point mathematics items, all mean score differences across years are negative values, suggesting that, on average, raters exhibited severity to a greater extent than they exhibited leniency during scoring. To more precisely describe the stability of scoring across administrations, results of MLMs are presented next.

Mathematics

Table 2 presents MLM results for the 1-point mathematics items. Results of the unconditional model indicated that after accounting for the multilevel data structure, the 1-point mathematics responses were scored slightly severely overall ($\beta_0 = -.013$, $p < .001$). All of the variation in score differences fell within validity

Table 3
Unstandardized Coefficients (and Standard Errors) of Mathematics Constructed Response 2-Point Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	−.011	.002	<.001	−.010	.002	<.001
Year slope (β_1)				−.001	.001	.254
Random effects						
Item intercept variance (σ^2_{u0})	.001	.000		.001	.000	
Item intercept/slope covariance (σ^2_{u01})				−.000	.000	
Item slope variance (σ^2_{u1})				.000	.000	
Validity response intercept variance (σ^2_{e0})	.061	.000		.061	.000	
Goodness of fit						
Deviance (−2LL)	72,322.012			71,029.885		
Difference test				$\chi^2_3 = 1{,}292.127$		<.001

Note. *P*-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

responses; therefore, no item random effects were included in the subsequent model. A likelihood ratio test contrasting Model 2 with the unconditional model indicated that allowing the predicted score differences to vary across the three administrations improved model fit ($\chi^2_5 = 2{,}923.104$, $p < .001$). Model 2 results showed that the predicted mean score difference for the 1-point mathematics responses in 2016 was −.010. These responses were scored slightly more severely each year in 2017 and 2018 ($\beta_1 = -.004$, $p < .001$). Adding year in Model 2 reduced the validity response intercept variance (σ^2_{e0}), and so explained (.029−.028)/.029 = 3.4% of the total variance in score differences.

Table 3 presents MLM results for the 2-point mathematics items. Unconditional model results showed the 2-point mathematics responses were scored slightly severely overall ($\beta_0 = -.011$, $p < .001$). Of the total variation in score differences, 1.6% fell between validity responses within items, whereas 98.4% fell within validity responses. A likelihood ratio test contrasting Model 2 with the unconditional model indicated that allowing the predicted score differences to vary across the three

Table 4
Unstandardized Coefficients (and Standard Errors) of Mathematics Constructed Response 3-point Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	.004	.001	<.001	.002	.001	.055
Year slope (β_1)				.003	.001	.001
Random effects						
Validity response intercept variance (σ^2_{e0})	.080	.000		.080	.000	
Goodness of fit						
Deviance (–2LL)	52,905.208			52,894.718		
Difference test				$\chi^2_1 = 10.49$		<.001

Note. *P*-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

administrations improved model fit ($\chi^2_3 = 1{,}292.127$, $p < .001$). Results of Model 2 indicate that the predicted mean score difference for the 2-point mathematics responses in 2016 was $-.010$. The change in scoring severity associated with these responses in subsequent administrations was insignificant ($\beta_1 = -.001$, $p = .254$). Additionally, adding year to this model did not explain any of the variance in score differences.

Table 4 presents MLM results for the 3-point mathematics items. Due to the small number of 3-point items (24), item-level effects were excluded from the model. Results of the unconditional model indicated the 3-point mathematics responses were scored slightly leniently overall ($\beta_0 = .004$, $p < .001$). A likelihood ratio test contrasting Model 2 with the unconditional model indicated that allowing the predicted score differences to vary across the three administrations improved model fit ($\chi^2_1 = 10.49$, $p < .001$). Model 2 results indicate that the predicted mean score difference for the 3-point mathematics responses in 2016 was .002 points. These responses were scored slightly more leniently in 2017 and 2018 ($\beta_1 = .003$, $p = .001$). However, the addition of year in Model 2 did not explain any of the variance in score differences.

English Language Arts

Table 5 presents MLM results for the 2-point research items. Unconditional model results indicated that the research responses were scored slightly severely overall ($\beta_0 = -.077$, $p < .001$). The random effects indicated that 2.2% of the total variation in score differences fell between validity responses within items, whereas 97.8%

Table 5
Unstandardized Coefficients and Standard Errors of ELA Research 2-Point Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	−.077	.004	<.001	−.064	.005	<.001
Year slope (β_1)				−.014	.002	<.001
Random effects						
Item intercept variance (σ^2_{u0})	.004	.000		.006	.000	
Item intercept/slope covariance (σ^2_{u01})				−.001	.000	
Item slope variance (σ^2_{u1})				.001	.000	
Validity response intercept variance (σ^2_{e0})	.177	.000		.164	.000	
Validity response intercept/slope covariance (σ^2_{e01})				.027	.001	
Validity response slope variance (σ^2_{e1})				−.025	.001	
Goodness of fit						
Deviance (–2LL)	648,282.002			644,514.813		
Difference test				$\chi^2_5 = 3{,}767.189$		<.001

Note. *P*-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

fell within validity responses. A likelihood ratio test contrasting Model 2 with the unconditional model indicated that allowing the predicted score differences to vary across the three administrations improved model fit ($\chi^2_5 = 3{,}767.189$, $p < .001$). Model 2 results show that the predicted mean score difference for research responses in 2016 was −.064 points. These responses were scored slightly more severely each year in 2017 and 2018 ($\beta_1 = -.014$, $p < .001$). Adding year in Model 2 reduced the validity response intercept variance (σ^2_{e0}), and so explained (.177−.164)/.177 = 7.3% of the within validity response variance in score differences.

Table 6 presents MLM results for the 2-point short writing task items. Results of the unconditional model indicated the short writing task responses were scored slightly severely overall ($\beta_0 = -.022$, $p < .001$). Random effects show that items accounted for 2.2% and validity responses for 97.8% of the total variation in score differences. A likelihood ratio test contrasting Model 2 with the unconditional model indicated that allowing the predicted score differences to vary across the three administrations improved model fit ($\chi^2_5 = 863.795$, $p < .001$). Results of Model 2 indicate

Table 6
Unstandardized Coefficients and Standard Errors of ELA Short Writing Task 2-Point Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	−.022	.006	<.001	−.010	.006	<.041
Year slope (β_1)				−.018	.004	<.001
Random effects						
Item intercept variance (σ^2_{u0})	.003	.001		.003	.000	
Item intercept/slope covariance (σ^2_{u01})				.000	.000	
Item slope variance (σ^2_{u1})				.001	.000	
Validity response intercept variance (σ^2_{e0})	.133	.000		.132	.001	
Validity response intercept/slope covariance (σ^2_{e01})				−.003	.001	
Validity response slope variance (σ^2_{e1})				.004	.001	
Goodness of fit						
Deviance (–2LL)	152,736.688			151,872.893		
Difference test				$\chi^2_5 = 863.795$		<.001

Note. P-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

that in 2016, the predicted mean score difference for the short writing task responses was −.010 points. These responses were scored slightly more severely in subsequent years ($\beta_1 = -.018$, $p < .001$). Adding year in Model 2 reduced the validity response intercept variance (σ^2_{e0}), and so explained (.133−.132)/.133 = .8% of the within validity response variance in score differences.

Table 7 presents MLM results for the 2-point reading items. Due to the small number of reading items (21), item effects were excluded from the model. Unconditional model results indicated that the reading responses were scored slightly severely overall ($\beta_0 = -.029$, $p < .001$). A likelihood ratio test contrasting Model 2 with the unconditional model indicated that allowing the predicted score differences to vary across the three administrations improved model fit ($\chi^2_3 = 332.946$, $p < .001$). Model 2 results show that the predicted mean score difference for the reading responses in 2016 was −.020. These responses were scored slightly more severely each year in 2017 and 2018 ($\beta_1 = -.017$, $p < .001$). Adding year in Model 2 reduced the validity response intercept variance (σ^2_{e0}), and so explained (.045−.035)/.045 = 22.2% of the total variance in score differences.

Table 8 presents MLM results for the 4-point essay traits. Results of the unconditional model showed that the 4-point ELA essay traits were scored slightly severely

Table 7
Unstandardized Coefficients and Standard Errors of ELA Reading 2-Point Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	−.029	.002	<.001	−.020	.002	<.001
Year slope (β_1)				−.017	.003	<.001
Random effects						
Validity response intercept variance (σ^2_{e0})	.045	.001		.035	.001	
Validity response intercept/slope covariance (σ^2_{e01})				.014	.002	
Validity response slope variance (σ^2_{e1})				−.008	.002	
Goodness of fit						
Deviance (–2LL)	−2,523.955			−2,856.901		
Difference test				$\chi^2_3 = 332.946$		<.001

Note. *P*-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

overall ($\beta_0 = -.061$, $p < .001$). Of the total variation in score differences, 3% fell between validity responses within items, whereas 97% fell within validity responses. As there were three unique 4-point essay traits (essays were evaluated for organization/purpose and either evidence/elaboration or development/elaboration, depending on the writing purpose), Model 2 tested for trait main effects and trait by year interactions to determine (1) whether mean score differences varied by trait in 2016 and (2) whether changes in mean score differences across scoring administrations varied by trait. A likelihood ratio test contrasting Model 2 with the unconditional model indicated that adding trait main effects, trait by year interactions, and allowing the predicted score differences to vary across the three administrations improved model fit ($\chi^2_8 = 657.604$, $p < .001$). In Model 2, the intercept ($\beta_0 = -.047$) indicates the predicted mean score difference for the organization/purpose trait in 2016. The mean score differences for the evidence/elaboration and development/elaboration traits in 2016 were therefore .065 points ($-.047 + -.018$) and $-.051$ points ($-.047 + -.004$), respectively. Model 2 slope coefficients indicated that all 4-point traits were scored slightly more severely in subsequent years that in 2016. Specifically, the organization/purpose trait was scored increasingly severely by $-.018$ points per year ($\beta_1 = -.018$, $p < .001$). The same was true for the development/elaboration trait, as its slope did not differ significantly from that of the organization/purpose trait ($\beta_5 = .003$, $p = .092$). Finally, the evidence/elaboration trait also reflected a minor increase in scoring severity across administrations, on the order of $-.018 + .003 = -.015$ points per year ($\beta_4 = .003$, $p < .001$). Adding year, trait main effects, and trait by year interactions in Model 2 reduced the validity

Table 8
Unstandardized Coefficients and Standard Errors of ELA Essay 1–4-Point Trait Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	−.061	.007	<.001	−.047	.006	<.001
Year slope (β_1)				−.018	.005	<.001
Evid/Elab (ref. Org/Purp, β_2)				−.018	.001	<.001
Dev/Elab (ref. Org/Purp, β_3)				−.004	.002	.016
Evid/Elab × Year (β_4)				.003	.001	<.001
Dev/Elab × Year (β_5)				.003	.002	.092
Random effects						
Item intercept variance (σ^2_{u0})	.006	.001		.004	.001	
Item intercept/slope covariance (σ^2_{u01})				.001	.000	
Item slope variance (σ^2_{u1})				.003	.000	
Validity response intercept variance (σ^2_{e0})	.192	.000		.191	.000	
Validity response intercept/slope covariance (σ^2_{e01})				−.003	.000	
Validity response slope variance (σ^2_{e1})				.003	.001	
Goodness of fit						
Deviance (–2LL)	2,757,321.600			2,743,277.632		
Difference test				$\chi^2_8 = 657.604$		<.001

Note. *P*-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

response intercept variance (σ^2_{e0}), and so explained (.192−.191)/.192 = .5% of the within validity response variance in score differences.

Table 9 presents MLM results for the 2-point essay trait, conventions. Unconditional model results indicated that the conventions trait was scored slightly leniently overall ($\beta_0 = .029$, $p < .001$). Random effects estimates show that items accounted for 3.7% and validity responses for 96.3% of the total variation in score differences. A likelihood ratio test contrasting Model 2 with the unconditional model indicated that allowing the predicted score differences to vary across the three administrations

Table 9
Unstandardized Coefficients (and Standard Errors) of ELA Essay 0–2-Point Trait (Conventions) Multilevel Models

	Unconditional Model			Model 2		
Parameter	Est.	*SE*	*p*-Value	Est.	*SE*	*p*-Value
Fixed effects						
Score difference intercept (β_0)	.029	.007	<.001	.011	.007	.117
Year slope (β_1)				.036	.004	<.001
Random effects						
Item intercept variance (σ^2_{u0})	.006	.001		.006	.001	
Item intercept/slope covariance (σ^2_{u01})				−.001	.000	
Item slope variance (σ^2_{u1})				.002	.000	
Validity response intercept variance (σ^2_{e0})	.156	.000		.154	.000	
Validity response intercept/slope covariance (σ^2_{e01})				.008	.001	
Validity response slope variance (σ^2_{e1})				−.009	.001	
Goodness of fit						
Deviance (–2LL)	1,136,041.132			1,128,947.572		
Difference test				$\chi^2_5 = 7{,}093.56$		<.001

Note. *P*-values are not provided for random effects as the Wald test for these parameters is only approximate. Est., unstandardized parameter estimate; *SE*, standard error.

improved model fit ($\chi^2_5 = 7{,}093.56$, $p < .001$). Model 2 results show that in 2016, the predicted mean score difference for the conventions trait was .011. The conventions trait was scored moderately more leniently in subsequent years ($\beta_1 = .036$, $p < .001$). Adding year in Model 2 reduced the validity response intercept variance (σ^2_{e0}), and so explained (.156−.154)/.156 = 1.3% of the within validity response variance in score differences.

Impact on Students' Scaled Scores

The computer-adaptive nature of the assessments meant that each student received a blueprint-compliant set of items selected to maximize test information near the student's ability. The blueprints specified a minimum number of hand-scored items. On average, hand-scored items represented 14–16 points in ELA and 4–9 points in mathematics, out of an average total of 65 raw score points per test. Note that the adaptive algorithm allowed slight variations in the number of hand-scored items and

total items assigned to a given student based on blueprint match and measurement precision tolerances.

The slope coefficients presented in Tables 2–9 quantify drift in leniency/severity across scoring administrations in raw score terms. Maximum likelihood estimation was used to estimate student ability for the assessments; this employed a two-parameter logistic model for 0–1 items and the generalized partial credit model (Muraki, 1992) for all other items. Student ability was reported in theta (θ) units. Scaled score conversions were linear transformations of the θ values and therefore not linearly related to the raw scores. In general, tests contained approximately 65 raw score points and had scaled score ranges of 432–582 points. Although raw scores and scaled scores are not linearly related, through most of the ability range scaled scores increased by 7.5–9.0 points per raw score point. To further examine the practical significance of the MLM results, we used the estimated slope coefficients, mean number of hand-scored items from the blueprints, and the scaled score/raw score ratio (i.e., 7.5–9.0) to estimate the net impact of leniency/severity effects across scoring administrations on students' scaled scores by grade and content area.

Table 10 summarizes the estimated scaled score changes from 2016 to 2018 due to drift in leniency/severity across scoring administrations. The raw score changes for each item group sum to an estimated net scaled score change for mathematics and ELA. Note that the table does not contain entries for the 2-point ELA reading items or the 4-point ELA development/elaboration trait because these items were exposed to a minority of students. (Specifically, 2–26% of students received a reading item and 20–40% of students received a narrative prompt associated with the development/elaboration trait). In sum, we estimated that instability in leniency/severity effects impacted students' scaled scores by less than one scaled score point out of a total of 432–582 points. For mathematics, estimated scaled score changes range from −.110 points in Grade 6 to .00 points in Grade 5. For ELA, estimated scaled score changes range from −.704 points in Grade 4 to −.443 points in Grade 11.

Discussion

The present study analyzed scoring data collected across three operational administrations of a large-scale summative assessment program in the United States. Expert-scored validity responses from a common pool were distributed to raters at frequent intervals throughout the 3 years of operational scoring. MLMs were used to examine the extent to which leniency/severity effects were stable across the three administrations. Model results were then applied to scaled scores in the critical region of the proficient cut to estimate the net impact of leniency/severity effects across scoring administrations on students' scaled scores.

Mathematics

Results of the 1-point and 3-point mathematics models showed relative stability in leniency/severity across scoring administrations. Specifically, results showed a slight increase in severity associated with the 1-point items, on the order of −.004 points per year. The conditional model with year effects yielded a 3.6% reduction in the within validity response variance in score differences compared to the unconditional

Table 10
Estimated Scaled Score Changes From 2016 to 2018 Due to Drift in Leniency/Severity Across Scoring Administrations

Content Area/Item Type/Trait	Score Point Range	Typical Number of Hand-Scored Items per Grade							Scaled Score Change from 2016 to 2018 by Grade						
		3	4	5	6	7	8	11	3	4	5	6	7	8	11
Mathematics	0–1	1	0	0	1	0	1	1	−.053	.000	.000	−.063	.000	−.066	−.072
Mathematics	0–2	2	2	3	3	2	2	2	−.027	−.028	−.044	−.047	−.032	−.033	−.036
Mathematics	0–3	0	0	1	0	0	0	1	.000	.000	.044	.000	.000	.000	.054
Total mathematics points		5	4	9	7	4	5	8	−.080	−.028	.000	−.110	−.032	−.099	−.054
ELA research	0–2	2	2	2	2	2	2	1	−.439	−.458	−.431	−.443	−.420	−.414	−.214
ELA short writing task	0–2	1	1	1	1	1	1	1	−.282	−.295	−.277	−.285	−.270	−.266	−.275
ELA essay: purpose/organization	1–4	1	1	1	1	1	1	1	−.282	−.295	−.277	−.285	−.270	−.266	−.275
ELA essay: evidence/elaboration	1–4	1	1	1	1	1	1	1	−.235	−.246	−.231	−.237	−.225	−.222	−.229
ELA essay: conventions	0–2	1	1	1	1	1	1	1	.564	.589	.554	.569	.539	.533	.549
Total ELA points		16	16	16	16	16	16	14	−.673	−.704	−.662	−.680	−.644	−.636	−.443

Note. Results assume: Differences are equal across grades; slope translates into raw score points; 65 raw score points per test; raw score to scaled score conversion is linear in vicinity of cut score; linear year slope.

model. Results also showed a slight increase in severity associated with the 3-point items, on the order of .003 points per year. However, adding time to this model did not reduce any of the within validity response variance in score differences compared to the unconditional model. Although the increase in severity was statistically significant, it was slight, and the magnitude of the effects and the amount of variance explained by time suggest this extent of drift is likely trivial unless it is to persist for many years.

Results of the 2-point mathematics model showed that overall rater effects were stable across administrations. This stability is notable given the number of 2-point items (199) and churn in the rater pool over the 3 years, and suggests the effectiveness of rater training at limiting variability (Raczynski et al., 2015; Woehr & Huffcutt, 1994). This finding aligns with a previous body of research that has reported relative consistency in scoring over time despite rater variation in leniency/severity, albeit in subjects other than mathematics (Leckie & Baird, 2011; Lunz & O'Neill, 1997; Zupanc & Štrumbelj, 2018).

English Language Arts

Results of the short constructed response models showed a slight increase in severity across administrations associated with scoring of the research, short writing task, and reading items. These responses were scored increasingly severely by −.014 points, −.018 points, and −.017 points per year, respectively. This may suggest that, over time, raters increasingly focused on the weaknesses found in responses rather than the positive characteristics indicative of score points 1 and 2. For the research items, adding year effects in Model 2 yielded a 6.5% reduction in the within validity response variance in score differences. The reduction in variance explained by time was .7% and 28.6% for the short writing task and reading items, respectively. The magnitude of these effects was larger than observed in mathematics, although time explained a relatively small amount of the variance in score differences for the research and short writing task responses. Furthermore, time explained nearly 29% of the variance in score differences for the reading items. Results of the reading model should be interpreted cautiously, however. As shown in Table 1, the number of reading items is small (21), as is the number of raters in any given year. Additionally, due to variation in reading item exposure, the score differences are extremely unbalanced within items such that nearly 80% of the reading score differences were clustered within three of the 21 items. Results may not generalize to a larger pool of reading items or larger number of raters. Collectively, results of the short constructed response models suggest the overall drift in rater error is of limited consequence at this point in time.

Results of the 4-point essay trait model provided evidence of slight drift across administrations associated with all three traits. The organization/purpose and development/elaboration traits reflected a comparable increase in severity over time, on the order of −.018 points per year. The evidence/elaboration trait also reflected increased severity but of a smaller magnitude (by −.015 points per year). The conditional model with year effects yielded a 1.5% reduction in the within validity response variance in score differences compared to the unconditional model.

Results of the 2-point essay trait model showed a modest increase in leniency across administrations associated with scoring of the conventions trait, on the order of .036 points per year. Adding year to the model yielded a 1.3% reduction in the within validity response variance in score differences compared to the unconditional model.

Students' total essay scores are calculated as the average of the 4-point trait scores, rounded up to the nearest integer, plus the 2-point trait score, for a maximum of six points. This means the increased severity in organization/purpose and evidence/elaboration scoring counterbalanced over 90% of the increased leniency in the conventions scoring at the total essay score level. These relations may be in part due to dependency among the trait scores, as for a given essay all three traits are scored by the same rater.

Previous research examining leniency/severity in essay scoring over time has shown mixed results. For example, findings reported in the literature include relative stability (Leckie & Baird, 2011), increased severity (Congdon & McQueen, 2000), and increased leniency (Lunz & Stahl, 1990). Collective essay model results identified divergent between-trait differences in trends, showing that, after controlling for prompts (e.g., items) and validity responses, scoring trends reflected increased severity across administrations in two traits (organization/purpose and evidence/elaboration) but increased leniency in another (conventions). To evaluate an essay, raters must make complex judgments of organization, elaboration, and conventions of a response, taking into consideration such factors as focus appropriate to the assigned writing purpose, the ability to connect ideas, use of relevant information from source materials to support the chosen focus, development of ideas related to the chosen focus, and grade-appropriate conventions such as spelling and sentence formation. Raters are known to vary in their decision-making behaviors to the extent that they consider components of multiple writing criteria (Cumming, Kantor, & Powers, 2001). Results suggest that raters' consideration of components shifted across administrations, toward a slightly higher standard for content-level components of writing (i.e., organization/purpose and evidence/elaboration) and a moderately lower standard for surface-level components of writing (e.g., conventions). These findings suggest the importance of expanding investigations of leniency/severity to incorporate longer time frames, a greater number of prompts, and multiple scoring dimensions. The role of scoring criteria becomes even more critical when used to make multiple judgments of performance (Eckes, 2008).

Impact on Students' Scaled Scores

MLM results were applied to scaled scores in the critical region of the proficient cut to estimate the net impact of leniency/severity effects across scoring administrations on students' scaled scores. This exercise revealed that, in general, raters scored increasingly severely over time. However, the cumulative effect of drift was estimated to be change of less than one scaled score point for tests having scaled score ranges of 432–582 points. Across all grades and content areas, the largest estimated change to students' scaled scores was −.704 scaled score points related to an assessment (Grade 4 ELA) with a range of 532 scaled score points. Taken

together, these results suggest that overall leniency/severity effects had little impact on students' scores.

General Discussion

Study results showed greater variation in score differences at the item level in ELA than in mathematics, suggesting more objective scoring criteria in mathematics. Across all item groups, however, items accounted for relatively little variation in score differences, explaining no more than approximately 4% of the total variance. For all item groups, the vast majority of the variance in score differences fell between validity responses within items. Previous research has shown scoring difficulty varies widely, depending on the nature of the response (Engelhard, 1996), and results align with a number of prior studies that have found individual responses to be one of the largest explainable sources of variation in rater accuracy (Baird et al., 2013; Leckie & Baird, 2011; Pinot de Moira, Massey, Baird, & Morrissey, 2002; Zhao et al., 2017).

Although some evidence of instable leniency/severity effects across administrations was identified, changes in leniency/severity across administrations were of a relatively small magnitude, with the maximum leniency change being .036 points per year relative to the essay conventions trait and the maximum severity change being −.018 points per year relative to the short writing task responses and essay organization/purpose trait. Scoring administrations explained 0–28.6% of the overall variability in score differences, depending on item group. When the cumulative effect of drift was applied to students' scaled scores, the estimated change was .000 to −.704 scaled score points depending on grade and content area. In many cases, these estimated changes are not substantially different from the changes introduced when scaled score rounding rules are applied for reporting purposes. Of primary interest to stakeholders in the assessment program is that collectively, findings suggest that the overall variability in leniency/severity effects across administrations present limited substantive concern.

Study results further provide implications for rater training and monitoring. By comprehensively analyzing all hand-scored items across all content areas, item types, and traits, results can inform resource allocation and training enhancements. For example, findings showed that ELA scoring was less stable over time than math scoring. Particularly, although ELA short constructed response items reflected similar severity drift across administrations, the application of MLM results to scaled scores illustrated that minimizing the increased severity associated with the research responses would most appreciably mitigate the overall impact of severity effects on students' scores (see Table 10). Combining comprehensive, longitudinal evaluations of scoring stability with real-time, rater-specific evaluations of performance (e.g., MFRM) allows greater insight into training and monitoring needs at all levels of the hand-scoring program.

Limitations

Study results should be interpreted in the context of several limitations. Validity responses were inserted randomly into scoring sets and assigned to raters at a ratio of one validity response for every 20 operational responses. However, we

did not directly examine rater effects, opting instead to analyze the complete data set to investigate scoring stability at the aggregate level. Additional research is needed to examine within- and between-rater change in leniency/severity across administrations, and to estimate the effects of individual raters on student scores. Second, in examining drift across administrations, we tested linear time effects only. It is possible that logarithmic, exponential, or quadratic time relationships may be appropriate. Third, estimates of leniency/severity were based on validity responses common to all three scoring administrations. We did not analyze score data associated with validity responses added to the pool in 2017 and 2018. Future research could examine longitudinal effects by analyzing validity responses administered in any two years, in order to accommodate additional responses over time and avoid relying on validity responses being administered in the baseline year. Fourth, despite the quality assurances and expert scoring process associated with validity responses, the absolute precision of true scores cannot be assumed (Sulsky & Balzer, 1988). Fifth, with regard to estimating impact on scaled scores, we applied the models uniformly within subject. Separate models by grade might reveal different results. Future research should employ each examinee's scaled score to estimate impact on students' scaled scores. Sixth, the validity pool is not assumed to comprise a random sample of validity responses or equal representation of score points. Average score differences in 2016 may reflect the nature or distribution of responses in the pool rather than absolute leniency/severity. Similarly, the validity pool score distribution is not assumed to match the score distribution of the operational sample. For this reason, we did not investigate central tendency effects. Finally, as this study examined the scoring of a nascent summative assessment program aligned with a relatively new set of performance standards, some amount of drift could be anticipated as the assessment consortium clarified the scoring criteria and provided guidance related to interpretation during the initial operational years. Changes in leniency/severity attributable to intentional refinements in application of the scoring criteria cannot be disentangled from leniency/severity drift. Further, teachers' familiarity with the standards and students' opportunity to learn and exposure to items aligned with the standards over time may have resulted in changes to the nature of student responses.

Conclusion

Both policy changes and technological improvements are changing the expectations around hand-scoring of high-stakes, large-scale assessments in the United States. The availability of technological devices has made online testing ubiquitous. Moreover, as the cost of technology has fallen and 1:1 devices become the norm, schools have been more able to administer summative assessments to students concurrently. This maximizes instruction but results in an increasing number of tests being submitted for scoring in an increasingly narrow timeframe each year. Online testing has also brought with it expectations for expeditious reporting, and a 10-day turnaround has become the norm in contracts for state testing programs. These changes mean more and more raters are required to score a fixed number of responses in a relatively short period of time, and call for novel approaches to

monitoring and evaluating the stability of scoring in the context of large-scale assessment programs.

This study was the first of its kind to provide a comprehensive examination of the longitudinal scoring stability of a large-scale assessment program. To our knowledge, this was also the first study to examine longitudinal scoring stability of mathematics and ELA short constructed response items. Future research should expand on this examination by investigating within- and between-rater change in leniency/severity across administrations and estimating the effects of individual raters on student scores.

Acknowledgments

Thanks to Timothy Davey, Claudia Flower, Brian Gong, Ronald Hambleton, Suzanne Lane, Richard Patz, and Martha Thurlow for encouraging this research. Thanks also to Susan Lottridge for feedback on earlier drafts of this article. Any remaining errors are the sole responsibility of the authors. The opinions expressed are those of the authors and do not represent the positions or policies of Measurement Incorporated.

Notes

[1]See Baird, Hayes, Johnson, Johnson, and Lamprianou (2013) for a detailed review of MRFM, generalizability theory, and MLM applied to rater effects research.

[2]While hand-scored items are included in the CAT part of the test to meet blueprint requirements, the adaptive algorithm adjusts the difficulty of questions during testing based exclusively on students' responses to machine-scored items.

References

Attali, Y. (2011). Sequential effects in essay ratings. *Educational and Psychological Measurement*, *71*, 68–79. https://doi.org/10.1177/0013164410387344

Baird, J.-A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). Marker effects and examination reliability. A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling (Ofqual/13/5261). Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378059/2013-01-21-marker-effects-and-examination-reliability.pdf

Charlton, C., Rasbash, J., Browne, W. J., Healy, M., & Cameron, B. (2017). MLwiN version 3.02. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, 163–178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Cumming, A., Kantor, R. & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series, MS-22). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RM-01-04.pdf

Darling-Hammond, L. (2010). *Performance counts: Assessment systems that support high-quality learning*. Washington, DC: Council of Chief State School Officers.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185. https://doi.org/10.1177/0265532207086780

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*, 56–70. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*, 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Lunz, M. E., & O'Neill, T.R. (1997, March). *A longitudinal study of judge leniency and consistency*. Paper presented at the Annual meeting of the American Educational Research Association, Chicago, IL.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*, 425–444. https://doi.org/10.1177/016327879001300405

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. https://doi.org/10.1177/014662169201600206

Myford, C. M., and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). Common core state standards. Washington D.C.: Author.

Pinot de Moira, A., Massey, C., Baird, J., & Morrissey, M. (2002). Marking consistency over time. *Research in Education*, *67*, 79–87. https://doi.org/10.7227/RIE.67.8

Raczynski, K. R., Cohen, A. S., Engelhard, G., Jr., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*, 301–318.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Ridge, J. K. (2001a, April). *Rater halo error and accuracy in a mathematics performance assessment*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Ridge, J. K. (2001b, March). *Do raters demonstrate halo error when scoring a series of responses?* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research and Evaluation*, *3*(3). Retrieved from http://PAREonline.net/getvn.asp?v=3&n=3

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428. https://doi.org/10.1037/0033-2909.88.2.413

Song, T., Wolfe, E. W., Hahn, L., Less-Petersen, M., Sanders, R., & Vickers, D. (2014). Relationship between rater background and rater performance. Retrieved from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/022_Song_RaterBackground_04_21_2014.pdf

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497–506.

Wesolowski, B. C., Wind, S. A., & Engelhard, G., Jr. (2017). Evaluating differential rater functioning over time in the context of solo music performance assessment. *Bulletin of the Council for Research in Music Education*, *212*, 75–98. https://doi.org/10.5406/bulcouresmusedu.212.0075

Wind, S. A., & Wesolowski, B. C. (2018). Evaluating differential rater accuracy over time in solo music performance assessment. *Bulletin of the Council for Research in Music Education*, *215*, 33–55. https://doi.org/10.5406/bulcouresmusedu.215.0033

Woehr, D., & Huffcutt, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189–205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35–51.

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *Journal of Technology, Learning, and Assessment*, *10*(1), 4–21. Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1601/1457

Zhao, H., Andersson, B., Guo, B., & Xin, T. (2017). Sequential effects in essay ratings: Evidence of assimilation effects using cross-classified models. *Frontiers in Psychology*, *8*, 933. https://doi.org/10.3389/fpsyg.2017.00933

Zupanc, K., & Štrumbelj, E. (2018). A Bayesian hierarchical latent trait model for estimating rater bias and reliability in large-scale performance assessment. *PloS ONE*, *13*(4), e0195297. https://doi.org/10.1371/journal.pone.0195297

Authors

COREY PALERMO is Vice President of Performance Assessment Scoring at Measurement Incorporated, 423 Morris St., Durham, NC 27701; cpalermo@measinc.com. His primary research interests include rater effects in large-scale assessment contexts and automated writing evaluation applications to improve the teaching and learning of writing.

MICHAEL B. BUNCH is Senior Vice President at Measurement Incorporated, 423 Morris St., Durham, NC 27701; mbunch@measinc.com. His primary research interests include large-scale assessment, standard setting, and automated scoring.

KIRK RIDGE is Senior Vice President at Measurement Incorporated, 423 Morris St., Durham, NC 27701; kridge@measinc.com. His primary research interests include performance assessment, rater error, and e-learning.