Briefs

Educational Evaluation and Policy Analysis
September 2024, Vol. 46, No. 3, pp. 581–590
DOI: 10.3102/01623737231165027
Article reuse guidelines: sagepub.com/journals-permissions

Testing an Explanation for Summer Learning Loss: Differential Examinee Effort Between Spring and Fall

Megan Kuhfeld
NWEA

James Soland
NWEA
University of Virginia

Brennan Register
University of Maryland, College Park

Andrew McEachin
NWEA

Summer learning loss is a perennial concern for educators and parents alike. However, researchers have recently questioned whether summer learning loss is just a statistical artifact driven by how achievement is measured across the school year. In this study, we empirically investigated a plausible critique of summer learning loss research, namely that students do not put forth their best effort on the fall test compared with the spring test. While we cannot conclude based on our findings that students do in fact lose ground during the summer, we did not find evidence that seasonal differences in test effort are a main driver of summer learning patterns estimated with MAP Growth assessments.

Keywords: *achievement, testing, educational policy, hierarchical linear modeling, summer learning loss, test effort*

SINCE Cooper et al.'s (1996) landmark study, a common understanding among the educators and public alike is that students routinely experience summer learning loss[1] (SLL; sometimes also referred to as "summer slide" or "summer setback") and that the phenomenon helps explain some of the most vexing sources of educational inequality in the United States (e.g., Alexander et al., 2007; Entwisle et al., 2000). More recently, nationally representative data collected by the Early Childhood Longitudinal Study (ECLS-K), 2010–2011 provides our most comprehensive look at SLL in the early grades. On average, this work found near-zero levels of growth during the summers following kindergarten and first grade, which reflects more of slowdown than learning loss (von Hippel et al., 2018). To understand patterns of SLL beyond first grade, researchers have used NWEA's MAP Growth assessments, which are administered multiple times a year to millions of U.S. K–8 students. For example, Atteberry and McEachin (2021) used data from almost 18 million students in Grades 1 to 8 and found that the average student lost 17% to 34% of the prior year's learning gains during summer break. In addition, Kuhfeld et al. (2021) documented racial/ethnic gaps in SLL using test scores from 2.5 million K–8 students. They found that SLL varies greatly across students, and that race/ethnicity and school-level poverty do not explain much of the variance in summer learning patterns. Based in part on this body of research, policymakers and educators have implemented a host of policies to mitigate SLL (National Summer Learning Association, 2019; von Hippel, 2016).

However, there are long-standing questions about whether fall-to-spring growth (and correspondingly, SLL) is actually a statistical artifact driven by differences in fall and spring testing conditions (Borman & D'Agostino, 1996; Keesling, 1984; Slavin et al., 1989; von Hippel, 2019). Examples of these critiques include the use of nonlinked test forms between spring and fall of adjacent grade levels in early SLL research (von Hippel & Hamrock, 2019), the role of teachers' coaching to potentially boost spring performance and depress fall test scores (Slavin et al., 1989), and other potential differences in students' motivation and test effort across the fall and spring assessments (Baird & Pane, 2018; Slavin, 2020). Researchers have also pointed to a lack of replication in findings about the magnitude of SLL as evidence that SLL estimates may be very sensitive to how measures are constructed (von Hippel, 2019; von Hippel & Hamrock, 2019). Given the high cost of providing summer learning programs designed to provide enriching summer experiences, typically with a goal of mitigating SLL (Schwartz et al., 2021), it is important to understand whether our concerns about SLL are distorted or even unfounded.

In this study, we assess the relationship between SLL patterns and differential test effort using measures of test effort and score quality. For test effort, we use a test effort metric supported by decades of validity evidence: The proportion of items on which a student responded so quickly, they could not have understood the item's content (a behavior referred to as "rapid guessing" [Wise, 2017]). For score quality, we are referring to metrics associated with whether individual scores are trustworthy, such as overall test duration, students' standard error of measurement (SEM), and percentage of items answered correctly. If, say, the SEM is very high, or the individual is getting far more items wrong than would be expected on an adaptive test, then the validity of the scores for their intended uses may be called into question. Note that, whereas our effort metrics are supported by extensive research as being valid for quantifying test effort (summarized by Wise, 2015), deviations in the score quality metrics could be caused by factors other than low effort and are therefore used as evidence to corroborate effort results. We use MAP Growth assessment data that are often the basis for analysis in the SLL literature (e.g., Atteberry & McEachin, 2021; Kuhfeld et al., 2021; von Hippel & Hamrock, 2019) to address two research questions:

Research Question 1: Do we see seasonal patterns in test effort and other score quality metrics that may indicate that students are not putting forth similar effort on fall and spring tests?

Research Question 2: Are SLL estimates sensitive to accounting for test effort and score quality metrics?

Background

SLL and Test Effort

Differences in summer learning patterns have been widely assumed to reflect disparities in family resources and access to enriching summer activities, such as summer camps, libraries, or other summer learning programs. However, family background explains a very small percentage of variation in SLL patterns (Burkam et al., 2004; von Hippel et al., 2018). Given the lack of documented evidence connecting student/community factors and SLL, researchers have speculated whether SLL reflects construct-irrelevant variance. For example, Baird and Pane (2018) suggested summer test score declines may actually reflect differences in testing conditions, including implicit or explicit pressures on students or educators to do well on spring assessments. For example, interim achievement tests given in the spring are often used as benchmarks for how students will do on state summative tests a few weeks later, with teachers (and perhaps students as well) potentially incentivizing students to try harder on the spring administrations given the scores' use as an accountability benchmark. Alternatively, interim tests given in the spring are likely given lower priority than summative tests, which could actually reduce effort on the interim measure if it is more of an afterthought than in the fall. Slavin (2020) expanded on this theory, noting that fall-winter gains on MAP Growth assessments are much larger than winter-spring gains, which he takes as evidence that the fall score is downwardly biased.

Metrics Relevant to Understanding Test Effort

The validity of an inference from a given test score assumes students' responses to all items reflect their knowledge of the domain of interest. Low effort is a violation to such an assumption. Currently, several metrics can be used to help understand a given student's test effort, most of which rely on metadata from computer-based tests (CBTs) and, in particular, computer-adaptive tests (CATs). While some metrics have been designed specifically to quantify test effort (e.g., response time effort, described below), others are more general indicators of the quality of a test score (e.g., percentage of correct responses to a computer-adaptive assessment and the SEM). Regardless, both types of metrics can be used to substantiate conclusions related to differential effort across testing periods (e.g., fall vs. spring). In the following sections, we review the body of literature on test effort measures and other metrics of score quality that can be used to evaluate whether estimates of SLL are sensitive to construct-irrelevant factors (such as differential effort) on fall tests.

Response-Time Effort

Schnipke and Scrams (2002) divided test examinees into two categories: those exhibiting "solution behavior" and those exhibiting "rapid-guessing behavior." Students in the latter category, who respond to a test item without sufficient time to have understood the question, are not engaged with the test during that item (Wise & Kong, 2005). Wise and Kong (2005) used an empirical approach based on the response time distribution for a given item to identify rapid-guessing behavior and generate an overall measure of a student's test-taking effort, which they term response time effort (RTE). RTE scores range from 1 to 0 and represent the proportion of test items on which the student exhibited solution behavior. Supplemental Appendix A (in the online version of the journal) provides a description of how individual items are flagged as disengaged and the validity evidence supporting the use of this metric to flag disengaged test-takers. While RTE is supported by extensive validity evidence for its intended use in CBT/CAT settings (see Wise, 2015), the methods on which it relies nonetheless have limitations. For example, individuals could be disengaged in ways unrelated to effort, an issue we discuss more in the study's conclusion. In addition, RTE was specifically designed to avoid overidentifying low-effort test-takers and therefore can be a conservative measure of disengagement (Wise & Kuhfeld, 2021). There is the definite possibility for Type 2 errors when identifying disengaged students.

Test Duration

Overall, test duration, or the minutes that students took to complete the test, is another approach to measure score quality (e.g., Kuhfeld & Soland, 2020; Soland, 2018). Prior research suggests students who complete a test much faster than is typical are not fully engaged with the material (e.g., Wise, 2015). Furthermore, there is evidence that students who spend longer on items are more motivated and conscientious, suggesting that test duration is likely related to test effort (Soland et al., 2019).

Percent Correct

On a CAT like MAP Growth, items are optimally targeted to a student's estimated achievement (e.g., items are selected to be maximally informative based on the students' prior correct/incorrect responses). As a result, students should get questions correct about 50% of the time. When students have a proportion correct on the test that is far lower than 50%, one might worry that the student is giving less than complete effort and randomly guessing across the test items. That is, a proportion correct below 50% indicates that students frequently responded incorrectly to items matched to their estimated achievement level.

Standard Error of Measurement

When tests are scored using item response theory (IRT), each student's score is accompanied by an SEM that helps quantify the precision of that score. SEMs may also help identify instances when a student did not provide full effort on the test. On MAP Growth in particular, students often have an SEM of around 3 scale score points. While deviations from that typical SEM can occur for a variety of reasons, low effort is one possible explanation. Thus, while an anomalously high SEM is not necessarily a sign of low effort, it could raise suspicions if other metrics such as RTE also indicate low motivation.

Lingering Questions Related to Test Effort and SLL

The question of whether test effort explains SLL patterns has received little empirical investigation so far. Kuhfeld and Soland (2020) found that SLL estimates were not significantly different under two approaches for adjusting for rapid guessing, including (a) filtering out students with low RTE and (b) adjusting students' test scores using an approach called effort-moderated scoring (Wise & DeMars, 2006). However, that study only examined reading test scores for the summer after fourth grade, and one cannot therefore be sure results generalize to other grades and subjects.

Methods

Sample

The data for this study are from NWEA's anonymized longitudinal student achievement database. School districts use NWEA's MAP Growth assessments to monitor elementary and secondary students' reading and mathematics growth throughout the school year, with assessments typically administered in the fall, winter, and spring. MAP Growth is a CAT that precisely measures achievement even for students above or below grade level and is vertically scaled to allow for the estimation of gains across time. Test scores are reported on the RIT (Rasch unIT) scale, which is a linear transformation of the logit scale units from the Rasch IRT model. Reliability and validity evidence to support the use of MAP Growth to monitor student achievement and growth within and across grades is described in NWEA (2019). MAP Growth is used by districts for a range of purposes, including progress monitoring, universal screening and placement decisions, predicting student performance on state assessments, evaluating programs, and occasionally in school/teacher evaluation systems.

We use the test scores of approximately 2.7 million kindergarten to seventh-grade students in 12,957 public schools across the United States. In this study, we follow students across two school years (2017–2018 and 2018–2019) and one summer break (summer of 2018). We chose these school years because they are the most recent years of data collected that were not disrupted by the COVID-19 pandemic. The NWEA data also include demographic information, including student race/ethnicity and gender, although student-level SES is not available. Table 1 provides descriptive statistics for the sample by subject and grade. A comparison of the 12,957 schools in our sample relative to the U.S. population of public schools (78,153 schools serving Grades K–8) is provided in Appendix B of Supplemental Materials (in the online version of the journal). Overall, the sample closely aligns to the characteristics of U.S. public schools, with a slight overrepresentation of Black students, underrepresentation of Hispanic students, and slight overrepresentation of urban schools.

Analytic Approach

We employ the four metrics described in the "Background" section to quantify students' test effort (or score quality more generally): (a) RTE, (b) overall test duration, (c) percent correct, and (d) SEM. Based on each of these measures, we created filters to remove students who showed signs of low effort on a given test event. We describe the thresholds employed for filtering in Appendix A of Supplemental Materials (in the online version of the journal). Supplemental Table A1 (in the online version of the journal) describes the characteristics of each of the

TABLE 1

Sample Characteristics of the Full Analytic Sample

			Sample size			Student race/ethnicity						
Subject	Grade (2017–2018)	Grade (2018–2019)	Test events	Students	Schools	White	Black	Asian	Hispanic	Other Race	Male	% FRPL
Math	0	1	1,555,656	256,394	6,245	0.48	0.18	0.04	0.15	0.09	0.51	0.54
Math	1	2	2,257,737	372,229	7,862	0.48	0.17	0.04	0.17	0.08	0.51	0.53
Math	2	3	2,388,745	394,212	8,624	0.49	0.17	0.04	0.17	0.09	0.51	0.52
Math	3	4	2,437,382	403,607	8,435	0.49	0.17	0.04	0.17	0.09	0.51	0.52
Math	4	5	2,420,610	401,296	8,191	0.50	0.17	0.03	0.17	0.08	0.51	0.51
Math	5	6	2,118,270	351,123	8,033	0.50	0.17	0.03	0.17	0.08	0.51	0.53
Math	6	7	1,906,396	316,241	4,640	0.51	0.17	0.04	0.16	0.08	0.50	0.51
Math	7	8	1,742,404	289,389	3,681	0.51	0.17	0.03	0.16	0.08	0.51	0.51
Reading	0	1	1,489,491	244,166	5,986	0.50	0.18	0.03	0.14	0.10	0.51	0.54
Reading	1	2	2,154,530	353,558	7,540	0.50	0.17	0.03	0.15	0.09	0.51	0.52
Reading	2	3	2,392,354	393,254	8,626	0.50	0.17	0.03	0.16	0.09	0.51	0.52
Reading	3	4	2,477,149	408,600	8,548	0.50	0.17	0.04	0.15	0.09	0.50	0.52
Reading	4	5	2,449,093	404,638	8,246	0.50	0.17	0.04	0.16	0.08	0.51	0.51
Reading	5	6	2,117,610	350,099	8,067	0.50	0.17	0.04	0.16	0.08	0.51	0.52
Reading	6	7	1,911,765	316,507	4,607	0.51	0.17	0.04	0.16	0.08	0.50	0.51
Reading	7	8	1,752,909	290,704	3,638	0.51	0.17	0.04	0.16	0.08	0.50	0.51

Note. FRPL = free or reduced-price lunch.

filtered samples. Supplemental Appendix C (in the online version of the journal) describes the methodology used to produce the SLL estimates within each subsample. As a further sensitivity test, we also examine the effects of low effort on summer loss using an item-level rescoring approach to remove noneffortful responses (prior research conducted by Rios et al., 2017, shows that removing examinees can be problematic if test effort is correlated with the student's true achievement). SLL results were not sensitive to this alternative approach for accounting for disengagement (see Supplemental Appendix D in the online version of the journal).

Results

Question 1. Descriptive Patterns by Season

We first examine descriptive patterns of student test scores and effort/quality metrics across term and grade cohort. Table 2 provides these descriptive statistics for math using the full analytic sample (see Supplemental Table A2 in the online version of the journal for the reading results). Overall, the average duration, SEM, RTE, and percentage of correct responses are highly similar between the spring and subsequent fall term. For example, students got an average of 51% of items correct in the spring of fourth grade relative to an average of 50% in the fall of fifth grade. The one exception is overall test duration. In Grades K–3, students spent longer on average on their fall assessment than the spring assessment in the prior grade (a difference of 3–8 minutes, which works out to approximately 4–10 seconds longer per item), while in the later grades test duration was 1 to 2 minutes shorter in the fall than the prior spring (1–3 seconds per item).

While we did not see strong seasonal patterns in test effort in our overall sample, it is possible that decreased test effort in the fall occurs in a subset of schools that may have attempted to manipulate testing conditions to artificially promote fall-spring growth (such as coaching students to give less effort in the fall as a means of boosting fall-spring gains). In Supplemental Appendix E (in the online version of the journal), we examined whether there are seasonal patterns in test effort and score quality in a subset of schools. School-level patterns were consistent with the overall sample results.

Question 2. The Relationship Between Test Effort and SLL Estimates

In addition, we estimate the sensitivity of SLL estimates to excluding students who showed patterns of lower test effort/quality based on RTE, test duration, SEM, and percent correct in a term. Figure 1 provides a comparison of the SLL estimates (reported as change in RIT score for each month of summer break) for the overall sample as well as our four restricted samples that removed disengaged test-takers (see Supplemental Table C2 in the online version of the journal for the full set of parameter estimates). Across all grades and subjects, we do not find that SLL is sensitive to various ways of removing students with low test effort or generally low-quality test events from our sample. Without filtering, we see an average SLL across grades of 1 to 3 RIT points per summer month. Regardless of whether students are filtered for RTE, test duration, percent correct, SEM, or all four criteria combined, the estimates of SLL remain between 1 and 3 RIT points per summer month.

Discussion

The idea that students lose ground academically over the summer has grown stronger among educators and policymakers. An increasing share of time and resources are spent on providing students summer programming to mitigate SLL from relatively hands-off programs that provide students reading material over the summer (Kim, 2006) to robust, multi-year interventions (Augustine et al., 2016). However, recently researchers have called into question whether SLL is a real phenomenon—oftentimes positing that differential test effort between fall and spring (due to differences in student motivation and/or explicit teacher coaching) is to blame—and, if not, suggested resources to prevent SLL would be better spent on different policies and programs.

In this brief, we tackle one of the main critiques against the SLL literature: Students spend a differential amount of effort on the fall test compared with the spring test. Regardless of the metric of test effort examined, we did not find evidence that seasonal differences in test effort/quality are a main driver of the SLL phenomenon. An important limitation is that our effort metric (RTE) is conservative in the sense that it is designed to avoid overidentification of lower effort. Furthermore, it

TABLE 2

Averages by Grade/Term/Cohort of RIT Scores and Test Effort for Math

Grade cohort)	Grade by term	Term	*N*	Months (prior to testing)	RIT	SEM	Duration (in minutes)	RTE	RTE <.90	% Correct
C1	0	F17	291,998	0.94	138.47	3.23	24.96	1.00	0.00	0.49
C1	0	W18	293,431	4.64	151.15	3.25	21.67	1.00	0.00	0.52
C1	0	S18	293,326	8.53	162.46	3.27	25.22	1.00	0.00	0.52
C1	1	F18	291,294	0.74	162.03	3.22	28.95	1.00	0.00	0.50
C1	1	W19	291,646	4.47	173.39	3.25	32.58	1.00	0.00	0.52
C1	1	S19	292,312	8.45	183.29	3.24	36.71	1.00	0.00	0.53
C2	1	F17	428,304	0.84	161.50	3.22	29.19	1.00	0.00	0.50
C2	1	W18	428,526	4.63	173.36	3.24	29.20	1.00	0.00	0.52
C2	1	S18	428,853	8.54	182.69	3.27	31.30	1.00	0.00	0.53
C2	2	F18	428,749	0.71	177.52	3.03	39.62	0.99	0.01	0.49
C2	2	W19	427,732	4.49	186.41	3.03	43.39	0.99	0.01	0.51
C2	2	S19	429,022	8.47	193.76	3.04	48.65	0.99	0.01	0.52
C3	2	F17	453,308	0.79	177.36	3.04	39.47	1.00	0.01	0.49
C3	2	W18	451,855	4.58	185.91	3.03	41.14	1.00	0.01	0.51
C3	2	S18	452,379	8.52	192.88	3.07	45.47	0.99	0.01	0.52
C3	3	F18	448,423	0.64	188.97	2.92	49.75	0.99	0.02	0.49
C3	3	W19	448,341	4.40	196.51	2.92	55.49	0.99	0.02	0.51
C3	3	S19	448,719	8.44	202.49	2.92	62.73	0.99	0.02	0.51
C4	3	F17	465,681	0.72	189.13	2.92	48.20	0.99	0.02	0.49
C4	3	W18	464,810	4.49	196.48	2.92	52.59	0.99	0.01	0.51
C4	3	S18	465,746	8.48	202.44	2.94	59.19	0.99	0.02	0.51
C4	4	F18	463,426	0.62	200.89	2.92	56.70	0.99	0.02	0.49
C4	4	W19	463,665	4.42	206.56	2.92	60.95	0.99	0.02	0.51
C4	4	S19	464,577	8.45	212.24	2.93	67.61	0.99	0.02	0.51
C5	4	F17	463,092	0.69	200.79	2.92	53.87	0.99	0.02	0.49
C5	4	W18	462,697	4.49	206.42	2.92	57.44	0.99	0.02	0.51
C5	4	S18	462,567	8.48	212.22	2.95	63.66	0.99	0.02	0.51
C5	5	F18	461,166	0.60	210.48	2.93	61.16	0.99	0.02	0.50
C5	5	W19	461,451	4.41	215.41	2.94	66.82	0.99	0.01	0.51
C5	5	S19	463,428	8.45	220.38	2.96	71.63	0.99	0.02	0.52
C6	5	F17	405,585	0.67	210.15	2.93	57.80	0.99	0.02	0.50
C6	5	W18	405,693	4.46	215.22	2.94	62.95	0.99	0.01	0.51
C6	5	S18	406,877	8.48	220.31	2.99	67.57	0.99	0.02	0.53
C6	6	F18	404,367	0.62	214.95	2.93	65.94	0.99	0.03	0.48
C6	6	W19	404,177	4.39	219.05	2.94	71.88	0.99	0.03	0.50
C6	6	S19	404,860	8.42	222.93	2.93	76.43	0.98	0.04	0.50
C7	6	F17	368,733	0.68	214.89	2.94	62.10	0.99	0.04	0.49
C7	6	W18	367,431	4.45	219.10	2.93	67.44	0.99	0.03	0.50
C7	6	S18	368,324	8.43	223.03	2.94	72.67	0.99	0.04	0.50
C7	7	F18	366,298	0.61	221.89	2.94	69.01	0.99	0.04	0.49
C7	7	W19	366,079	4.38	225.15	2.95	74.45	0.98	0.04	0.50
C7	7	S19	367,322	8.39	228.31	2.94	75.89	0.98	0.05	0.50
C8	7	F17	334,189	0.68	221.51	2.94	64.42	0.98	0.05	0.49
C8	7	W18	333,111	4.45	224.99	2.94	69.73	0.98	0.04	0.50
C8	7	S18	334,286	8.42	228.21	2.95	72.36	0.98	0.05	0.50
C8	8	F18	333,028	0.61	227.52	2.96	70.09	0.98	0.04	0.49
C8	8	W19	332,215	4.36	230.57	2.97	74.85	0.98	0.04	0.50
C8	8	S19	333,681	8.36	233.11	2.96	74.23	0.98	0.05	0.50

Note. Reading results are presented in Supplemental Appendix A (in the online version of the journal). C1 = Cohort 1 (students in Grades K–1); C8 = Cohort 8 (students in Grades 7–8); RIT = average test score; SEM = standard error of measurement; RTE = response time effort (% of responses that were effortful).

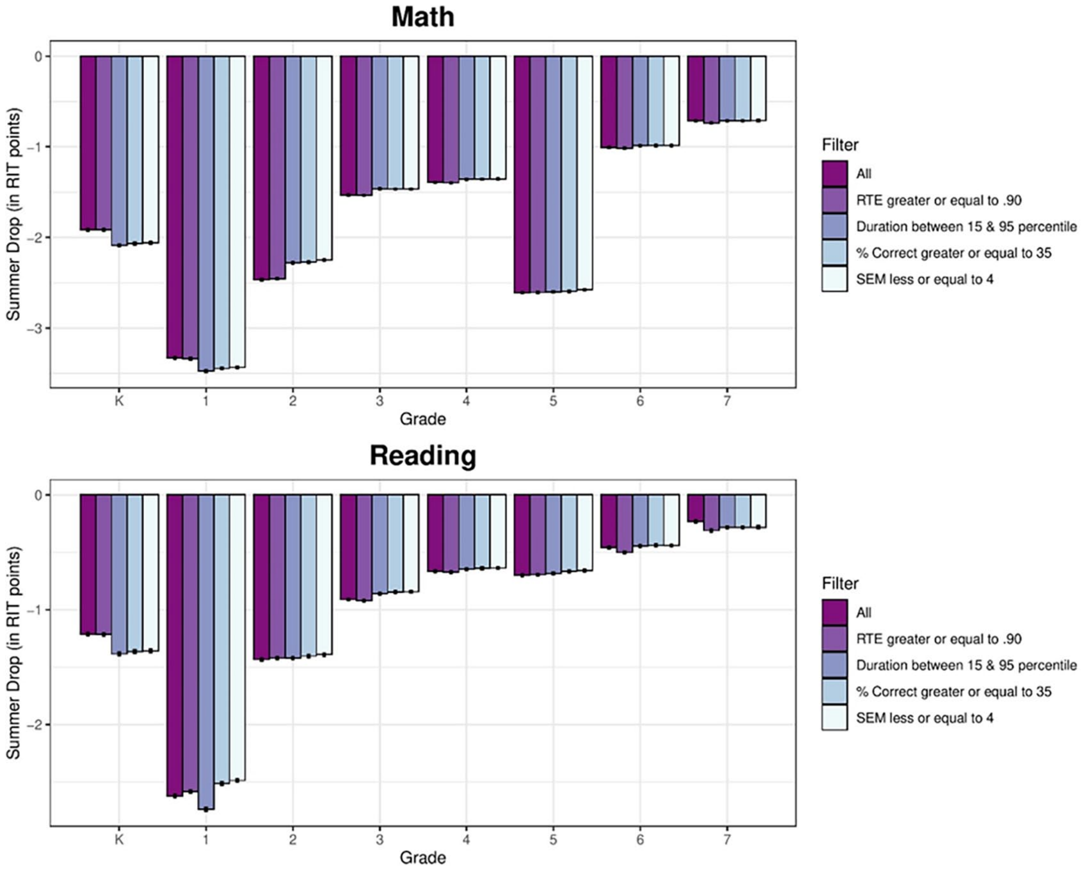

FIGURE 1. *Comparison of summer loss estimates from the growth model across various sample criteria.*
Note. Grade refers to the grade level students were in prior to the summer break. For further description of the effort filters employed, see Supplemental Appendix A (in the online version of the journal). RTE = response time effort; SEM = standard error of measurement; RIT = Rasch unIT.

likely does not capture other manifestations of low effort, such as when students respond in a typical amount of time, but are nonetheless providing suboptimal effort. We used RTE because it is supported by the most validity evidence for its intended use among effort metrics (Wise, 2015) and try to address its potential limitations using a range of metrics such as test duration, which avoids setting (conservative) response time thresholds separating effortful from noneffortful responses. While our analysis does not rule out all possible construct-irrelevant factors related to SLL, it does suggest that differential test effort between fall and spring is not likely one of them, at least when using MAP Growth assessments and using the particular measures of test effort available to us. Future research should examine whether findings vary in contexts in which assessments (including other interim assessments) are used for high-stakes decision-making and employing other effort metrics.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Megan Kuhfeld https://orcid.org/0000-0002-2231-5228
James Soland https://orcid.org/0000-0001-8895-2871
Andrew McEachin https://orcid.org/0000-0002-5113-6616

Supplemental Material

Supplemental material for this article is available online.

Note

1. We use the term "summer learning loss" in this article because this is widely used in the literature, but we acknowledge that it is often seen as deficit-oriented and does not fully capture the range of learning experiences that students have in the summer.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review*, *72*(2), 167–180.

Atteberry, A., & McEachin, A. (2021). School's out: The role of summers in understanding achievement disparities. *American Educational Research Journal*, *58*(2), 239–282.

Augustine, C. H., McCombs, J. S., Pane, J. F., Schwartz, H. L., Schweig, J., McEachin, A., & Siler-Evans, K. (2016). *Learning from summer: Effects of voluntary summer learning programs on low-income urban youth*. RAND Corporation.

Baird, M. D., & Pane, J. F. (2018). *Controlling for changes in test conditions when estimating education intervention effects*. RAND Cooperation. https://www.rand.org/pubs/working_papers/WR1245.html

Borman, G. D., & D'Agostino, J. V. (1996). Title I and student achievement: A meta- analysis of federal evaluation results. *Educational Evaluation and Policy Analysis*, *18*, 309–326.

Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education*, *77*(1), 1–31.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, *66*(3), 227–268.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2000). Summer learning and home environment. In R. D. Kahlenberg (Ed.), *A notion at risk: Preserving public education as an engine for social mobility* (pp. 9–30). Century Foundation Press.

Keesling, J. W. (1984). *Differences between fall-to-spring and annual gains Chapter I programs*. Advanced Technology.

Kim, J. S. (2006). The effects of a voluntary summer reading intervention on reading achievement: Results from a randomized field trial. *Educational Evaluation and Policy Analysis*, *28*(4), 335–355.

Kuhfeld, M., Condron, D., & Downey, D. (2021). When does inequality grow? A seasonal analysis of racial/ethnic disparities in learning in kindergarten through eighth grade. *Educational Researcher*, *50*(4), 225–238.

Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness*, *13*(1), 147–175.

National Summer Learning Association. (2019). *Scholastic summer learning tip sheet*. https://www.summerlearning.org/knowledge-center/12076/

NWEA. (2019). *MAP Growth technical report*. https://www.nwea.org/content/uploads/2021/11/MAP-Growth-Technical-Report-2019_NWEA.pdf

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104.

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

Schwartz, H., Augustine, C., & McCombs, J. S. (2021, April 15). Commit now to get summer programming right. *The RAND Blog*. https://www.rand.org/blog/2021/04/commit-now-to-get-summer-programming-right.html

Slavin, R. (2020). *The summer slide: Fact or fiction?* https://robertslavinsblog.wordpress.com/2020/08/20/the-summer-slide-fact-or-fiction/

Slavin, R., Karweit, N. L., & Madden, N. A. (1989). *Effective programs for students at-risk*. Allyn and Bacon.

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323.

Soland, J., Zamarro, G., Cheng, A., & Hitt, C. (2019). Identifying naturally occurring direct assessments of social-emotional competencies: The promise and limitations of survey and assessment disengagement metadata. *Educational Researcher*, *48*(7), 466–478.

von Hippel, P. T. (2016). Year-round school calendars: Effects on summer learning, achievement, families, and teachers. In K. Alexander, S. Pitcock, & M. Boulay (Eds.), *The summer slide: What we know and can do about summer learning loss*. Teachers College Press. https://www.google.com/books/edition/The_Summer_Slide/ZBu5DQAAQBAJ?

hl=en&gbpv=1&dq=The+summer+slide:+What+we+know+and+can+do+about+summer+learning+loss&printsec=frontcover; https://lbj.utexas.edu/sites/default/files/pdfs/PvonHippel-CV_2021.pdf https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2766106

von Hippel, P. T. (2019). *Is summer learning loss real? How I lost faith in one of education research's classic results*. Education Next. https://www.educationnext.org/is-summer-learning-loss-real-how-i-lost-faith-education-research-results/

von Hippel, P. T., & Hamrock, C. (2019). Do test score gaps grow before, during, or between the school years? Measurement artifacts and what we can know in spite of them. *Sociological Science*, *6*, 43–80.

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*, 323–357.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61.

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

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

Wise, S. L., & Kuhfeld, M. (2021). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130–149.

Authors

MEGAN KUHFELD, PhD, is a senior research scientist at NWEA. Her research seeks to understand students' academic and social-emotional learning trajectories and the school and neighborhood influences that promote optimal growth.

JAMES SOLAND, PhD, is an assistant professor of quantitative methods at Curry School of Education and Human Development at the University of Virginia. His research focuses on connections among measurement, estimating growth, and program evaluation, with applied interest in social-emotional learning.

BRENNAN REGISTER, MA, is a PhD student at the University of Maryland, College Park. Her research focuses on investigating the performance of multilevel and standard prediction algorithms on large-scale educational data sets.

ANDREW MCEACHIN, PhD in Education Policy, is the director of NWEA's Collaborative for Student Growth. His research focuses on understanding the causes and consequences of educational inequities.

Manuscript received May 23, 2022
Revision received October 14, 2022; January 3, 2023
Accepted February 16, 2023