DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Disrupted Data: Using Longitudinal Assessment Systems to Monitor Test Score Quality

Citation
An, Lily S., Andrew D. Ho, and Laurie Laughlin Davis. 2022. Disrupted data: Using longitudinal assessment systems to monitor test score quality. Educational Measurement: Issues and Practice.

Published Version
https://doi.org/10.1111/emip.12491

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37370969

Terms of Use
This article was downloaded from Harvard University's DASH repository, WARNING: No applicable access license found.

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Disrupted Data: Using Longitudinal Assessment Systems to Monitor Test Score Quality

Lily An[1], Andrew Ho[1], and Laurie Laughlin Davis[2]

[1]Harvard Graduate School of Education

[2]Curriculum Associates

Author Note

Correspondence concerning this article should be addressed to Lily An, Harvard Graduate School of Education, 14 Appian Way, Cambridge, MA 02138.

Email: lily_an@g.harvard.edu

Abstract

Technical documentation for educational tests focuses primarily on properties of individual scores at single points in time. Reliability, standard errors of measurement, item parameter estimates, fit statistics, and linking constants are standard technical features that external stakeholders use to evaluate items and individual scale scores. However, these cross-sectional, "point-in-time" features can mask threats to the validity of score interpretations, including those for aggregate scores and trends over time. We use test score data collected before and during the COVID-19 pandemic to show that longitudinal analyses, not just point-in-time analyses, are necessary to detect threats to desired inferences. We propose that educational agencies require and vendors include longitudinal data features, including "match rates" and correlations, as standard exhibits in technical documentation.

Keywords: longitudinal analysis, longitudinal data, COVID-19, reliability, technical reporting

Disrupted Data: Using Longitudinal Assessment Systems to Monitor Test Score Quality

The COVID-19 pandemic began affecting U.S. schools in March of 2020, nearly eliminating collection of spring standardized test scores. Flexible testing policies in spring 2021 led to variation in testing completion across the country (Data Quality Campaign, 2021). Such sparse and nonrandom missing data pose serious threats to desired interpretations about educational proficiency and progress. Furthermore, the limited data that have been collected from student testing since spring 2020 have included scores from potentially noncomparable conditions—for example, scores from tests taken at home as opposed to in school due to school building closures during the pandemic.

In this article, we show how these recent features of test data motivate the need for a longitudinal approach to examining data quality. Though multiple approaches exist to address missing data within a tested population (e.g., Peugh & Enders, 2004) and within individual item response vectors at a single test occasion (e.g., Sinharay, 2021), a key inference for test scores, especially in a pandemic, is about educational progress or decline for populations over time. The Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014) stress the importance of describing the target population and collecting information on testing conditions for assessing validity evidence. For example, standard 1.10 states: "Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance" (American Educational Research Association et al., 2014). Though the Standards do not comment on biases imparted due to missing data for aggregate analyses explicitly, this importance follows logically from standards like 1.10 given the target inferences.

Educational progress and declines represent information about score differences between time points. Standard 12.11 addresses this for individual score reporting: "when difference or growth scores are used for individual students, such scores should be clearly defined, and evidence of their validity, reliability/precision, and fairness should be reported" (p. 198). Combining the implications of these standards suggests a need to raise the standards on technical documentation for tests that measure the progress of populations over time. Documenting and maintaining high data quality is necessary for users to ensure their score inferences have sufficient validity evidence.

Traditionally, test developers and test score users attend to cross-sectional metrics, what we call single "point-in-time" metrics, to describe test data quality, including illustrative metrics we review in Table 1. We explain why these metrics are theoretically insufficient when changes in testing conditions, content, and populations occur over time. Then, we present additional statistics and graphical representations whose documentation will enable stakeholders to better understand threats to key inferences about aggregate-level progress.

Table 1

Examples of Cross-Sectional Quality Metrics Reported in K-12 Test Technical Documentation

Assessment	Reliability	Standard errors of measurement	Item fit statistics	Differential item functioning statistics
Smarter Balanced[a]	X	X		X
MCAS[b]	X	X		X
MAP Growth[c]	X	X	X	X
i-Ready[d]	X	X	X	X

Note. Based on authors' review of available technical documentation.
[a] Smarter Balanced Assessment Consortium (2020a, b).
[b] Cognia, Inc., and Massachusetts Department of Elementary and Secondary Education (2020).
[c] NWEA (2019).
[d] Curriculum Associates (2019).

Limiting features of "point-in-time" quality metrics

Point-in-time quality metrics like reliability and item fit indices may be insensitive to or even inflated by certain disruptions in an educational crisis like those caused by COVID-19. Consider two plausible scenarios: a high-proficiency remote assistant, or unequal opportunities to learn. In remote testing environments with reduced or no proctoring, parents, siblings, and others may be present during a remote test administration. If they provide consistent assistance, the student test taker's item responses will reflect a high proficiency that may be entirely consistent with an Item Response Theory (IRT) model. Similarly, if opportunities to learn are unequal and exacerbated by the pandemic, some students may show true gains in proficiency and others a true decline, resulting in lower year-to-year correlations. These may again be consistent with an IRT model, including all fit statistics (e.g., Engelhard, 2013) and differential item functioning (DIF) statistics. In short, tautologically, data that fit the model will show model fit.

These two scenarios differ in their threats to valid score inferences. The first scenario is clearly construct irrelevant. The desired inference is student proficiency, not student proficiency with varying degrees of parental or other assistance. The second scenario is construct relevant from the perspective of intended individual score interpretations, which are typically descriptive statements about what students know and can do. If students have variable opportunities to learn, they may indeed increase in their true variance of knowledge and skills. However, stark decreases in correlations over time caused by changes in opportunity to learn outside of school weaken the argument that test scores are indicators of the health of educational systems. In the case of COVID-19, when numerous nonschool factors had direct effects on learning, decreased

correlations suggest refocusing inferences on the effects of the pandemic, not the effects of schools (Bacher-Hicks & Goodman, 2021).

These scenarios also present a possibility that reliability may improve. It is well known that expanding true variance in the population will increase reliability coefficients (e.g., Gulliksen, 2013). This may occur if examinees receive assistance from a high-proficiency assistant or if opportunities to learn are correlated with initial proficiency. A score user who is unaware of the increased group variance would calculate a higher overall reliability coefficient in this year, and the test may not be flagged as anomalous.

Methods for diagnosing changes in test quality using longitudinal data

Given these undetected changes in test quality using common cross-sectional metrics, we demonstrate how descriptive statistics from longitudinal data can help score users detect potential threats to inferences about trends over time. We use data from Curriculum Associates' i-Ready Diagnostic mathematics and reading tests to illustrate how these statistics should be standard exhibits in technical reporting to improve test quality monitoring.

Measure

The i-Ready Diagnostic is a K–12 vertically scaled computer adaptive test that measures reading and mathematics skills and is administered by schools as an interim assessment to students three times a year, in the fall, winter, and spring. The i-Ready Diagnostic provides scale scores that place students into performance levels (e.g., below grade, early on-grade, mid on-grade, late on-grade, and above grade) corresponding to their mastery of reading or mathematics content. These results are also used to place students into personalized learning paths within the "i-Ready Online Instruction" system. Beginning in the fall of 2020, students were asked to

answer the following yes/no question at the beginning of each i-Ready Diagnostic testing session: “Are you working in your school building today?” Answers to this question were compiled across testing sessions within a test administration. Students were classified as “in school” if they answered yes to this question for all testing sessions within a test administration and were classified as “remote” if they answered no to this question for any testing session within a test administration.

Sample

To understand longitudinal patterns, we include students who have i-Ready scores at any time from the 2016–2017 academic year to the 2020–2021 academic year, in the fall, winter, or spring seasons. COVID-19 effects on U.S. testing began during the spring 2020 testing window. In our first analytic example, we restrict the data to students who followed standard grade progressions in three consecutive fall seasons, from fall 2018 through the fall of 2020. This means that all students who were in first grade in fall 2018 were also in second grade in fall 2019 and then third grade in fall 2020. As a percentage of all students who tested in each of these fall seasons, this resulted in only a small percentage of exclusions (1.8–2.1%). In our second analytic example, we use spring test takers from the spring of 2017 through the spring of 2021.

For the first analytic example, students are classified into two groups, “in school” or “remote,” based on their self-reported testing location in the fall of 2020. We apply these group labels retroactively to prior testing occasions so that differences in group performance can be understood longitudinally. In other words, the scores of the “remote” students in the fall of 2019 were students who tested in school in 2019 but remotely in 2020, whereas the “in school” students in the fall of 2019 were students who tested in school in both 2019 and 2020.

Longitudinal descriptive statistics: Means, standard deviations, and correlations

To illustrate the value of longitudinal descriptive statistics, we highlight the score properties of students who tested in kindergarten in 2018, first grade in 2019, and second grade in 2020. Table 2 presents means, standard deviations, and correlations of i-Ready scores for the kindergarten 2018 cohort in mathematics over this 3-year span. Rather than presenting a point-in-time report of 2020 results or a cross-sectional presentation of trends for a 2020 population that differs from the 2019 population, descriptive statistics like these enable us to track longitudinal progress for a stable population of students. Further longitudinal analyses can help us evaluate whether the reported results indicate a change in proficiency or could be confounded with a change in population or a change in the mode of administration.

Table 2

Longitudinal Descriptive Statistics for the Kindergarten 2018 Cohort in Mathematics by Location.

	Location	**2018 - K**	**2019 - 1st**	**2020 - 2nd**	**n**
Means	Remote	344.9	375.2	408.6	131,647
	In School	346.2	377.3	398.3	84,709
	Total	345.4	376.0	404.6	216,356
Standard Deviations	Remote	21.8	25.5	30.5	131,647
	In School	21.2	24.4	25.4	84,709
	Total	21.6	25.1	29.0	216,356
Correlations	Remote	0.64	0.51		131,647
	In School	0.64	0.73		84,709
	Total	0.64	0.57		216,356

Without the remote versus in-school test location flag, only the “Total” rows in Table 2 would be estimable. When splitting the test results by the students’ test location in fall 2020 and looking back at their prior results (when all students tested in school), we can discern whether 3-

year trends differ for remote versus in-school testers. For example, for the kindergarten cohort in fall of 2018, remote testers showed dramatic gains from 2019 to 2020 of 33.4 points on average, from 375.2 in 2019 to 408.6 in 2020. These gains exceed their gains from 2018 to 2019 of 30.3 points, and go far beyond the gains by in-school testers of 21.0 points from 2019 to 2020. Longitudinal links between the fall 2020 testing location and prior year scores enable detection of remote testing location as a potential threat to interpretations of early-grade gains. We can construct similar examples in other subjects and grades.

Figure 1

Pre-pandemic and During-pandemic Changes in Means, Standard Deviations, and Correlations for Remote vs. In-School Students in the Kindergarten 2018 Cohort

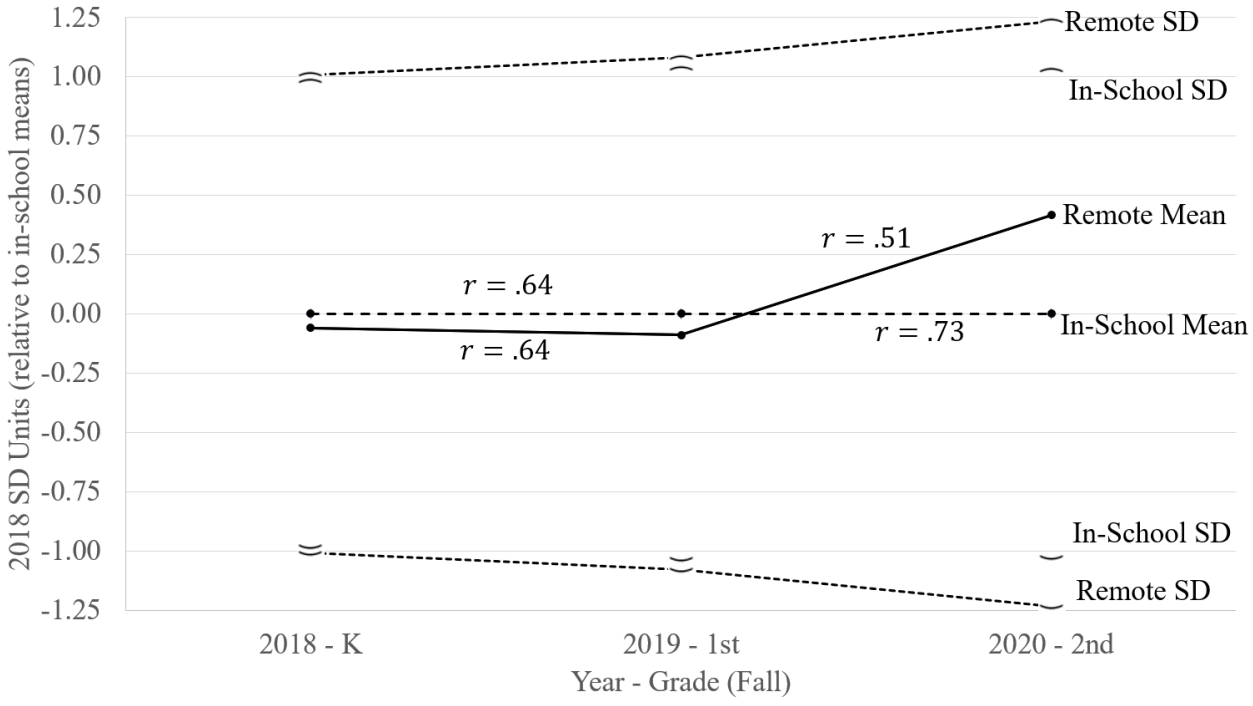


Figure 1 shows the differences between remote and in-school students by standardizing within-grade scores to the in-school mean of that year and the 2018 (prepandemic) standard

deviation for all students in each grade. (2018 standard deviations are 21.6, 23.6, and 24.7 scale points in grades K, 1, and 2, respectively.) Standardization helps to compare relative differences on an interpretable scale across grades. Remote students had slightly lower scores in prepandemic fall administrations, around 0.06 and 0.09 standard deviation units below in-school students in 2018 and 2019, respectively. During the pandemic, their scores were 0.42 standard deviation units higher. We also find that remote-testing students had more variable scores than their in-school counterparts in the fall of 2020, with standard deviations around 23% higher in that year, compared to around 1% and 8% higher standard deviations in prior years. Without longitudinal data, these findings in fall 2020 would not necessarily indicate an anomaly. Perhaps this is simply a more variable population. However, by comparing their standard deviations to those in previous years, when all these same students tested in school, we can establish that standard deviations are unusually large and that these changes are unexpected.

Longitudinal correlations also enable useful comparisons that can help users detect anomalous results. Table 2 shows that both populations had correlations of 0.64 from 2018 to 2019, from their kindergarten to first-grade years. From first to second grade, the correlation for remote students was 0.51, whereas the correlation for in-school students was 0.73. Reporting correlations for longitudinal data enables users to develop norms for these correlations over time. When they change dramatically in a particular year or for a particular subpopulation, users can then investigate threats to validity that point-in-time metrics alone could not detect.

Match rates and matching methods

In addition to reporting longitudinal diagnostics like means, standard deviations, and year-to-year correlations, we also recommend reporting changes in tested populations. This could take the form of the percent of a previous year's students who (a) remain in the data versus

those who (b) were not tested in the current year, as well as (c) the percent of students who are new in the current year. These percentages are "match rates," the percentages of students with available test scores in 1 year who also have available test scores in another year. Match rates may be particularly low in crises like these, where there may be "substantial and atypical numbers of students who are not in school who otherwise would be enrolled, above and beyond typical mobility rates" (Ho, 2021, p. 2). This attendance drop has been substantial during the COVID-19 pandemic in certain populations (Korman et al., 2020).

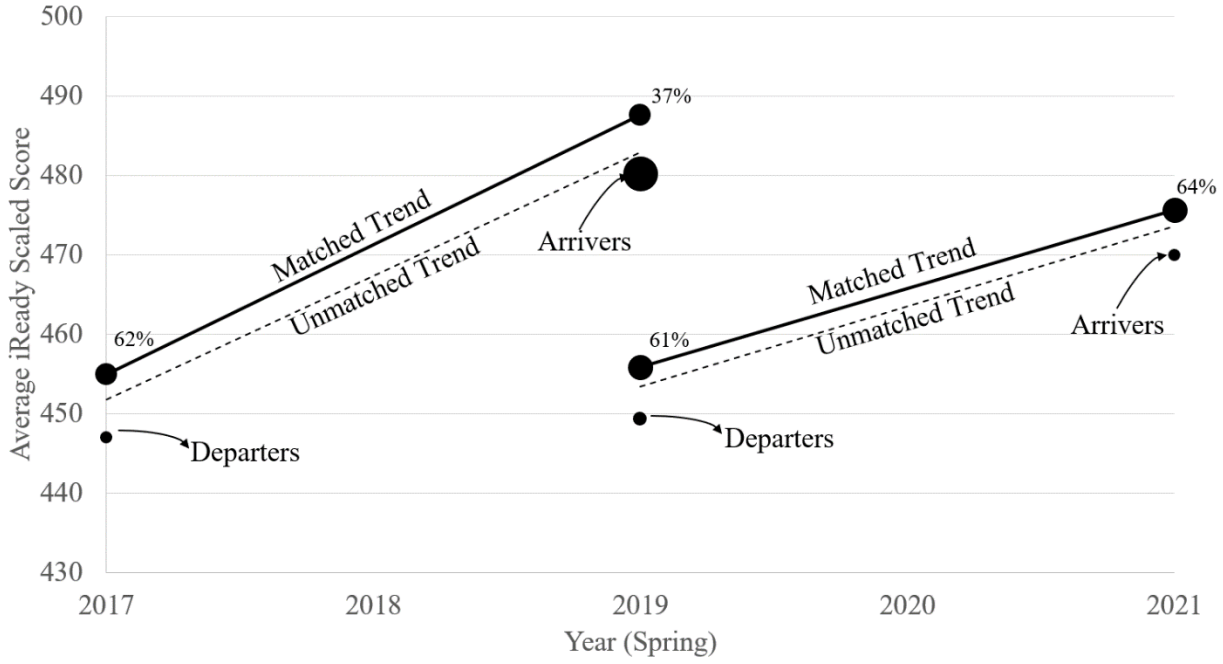
Match rates can also form the basis for matching methods that estimate the population proficiency and trends that would have occurred had the population remained constant. These methods make the often-untestable assumption that observed and included covariates, and no others, can explain which examinees were tested and which were not. Because of this speculation, matched results are typically the targets of research reports, not technical documentation, so we leave elaboration to others (e.g., Ho, 2021; Reardon et al., 2019; Stuart, 2010). Nonetheless, we maintain that match rates, a key ingredient to the integrity of matching methods, should be reported as an indicator of the coverage and representativeness of the population.

To illustrate the relevance of match rates for spring state test scores, we use a dataset of longitudinally linked spring i-Ready test scores from the spring of 2017 through the spring of 2021 to form our second analytic sample. Figure 2 shows mathematics results from two cohorts of third-grade students over a 3-year span: those in 2017 who were unaffected by the pandemic through 2019, and those in 2019 who were affected by the pandemic through 2021. The figure shows observed, that is, unmatched, trends as dotted lines. These show less average growth from

2019 to 2021 than from 2017 to 2019. This is a story consistent with the COVID-19 pandemic limiting student opportunities to learn mathematics.

Figure 2

Matched and Unmatched i-Ready Mathematics Score Trends for Two Cohorts of Third Grade Students



Note: Marker sizes are proportional to the size of the tested population. Match rates are labeled.

However, unmatched trends can be confounded by departing and arriving students. Figure 2 shows this by contrasting dotted lines with a solid line indicating the trend for students who have scores at both the beginning and ending time points. To explain the discrepancy between the solid and dotted lines, the figure also plots the average scores of departing students and arriving students. The first endpoint of each dotted line is the weighted average of the matched students and the departers. The second endpoint of each dotted line is the weighted

average of the matched students and the arrivers. In this case, the unmatched trend for the 2017 cohort is biased downward from the matched trend due to the relatively large influx of students with lower average scores in the spring of 2019. Both departers and arrivers were similarly nonrepresentative, but the arrivers formed a larger proportion. The unmatched trend for the 2019 cohort is similar to the matched trend, because the departers and arrivers were similar in proportion and similar in nonrepresentativeness (both were relatively low scoring on average).

The figure suggests that there are two key features to report for longitudinal comparisons. The first is the match rate, indicating the coverage of the matched group over the count of students involved in the endpoints of the unmatched trends. The second is representativeness, or the score differences between the matched students and those who are not in the matched group.

The match rate should be referenced to a specific year's total population. For example, the match rate for third graders in 2017 is 62%, and the match rate for fifth graders in 2019 is 37%. This means that 62% of third-grade test takers in 2017 also tested in 2019 as fifth graders, while only 37% of fifth-grade test takers in 2019 also tested in 2017 as third graders. Depending on which overall total test population is of interest, the match rate looks quite different (though the numerator, the count of students who tested in 2017 as third graders and in 2019 as fifth graders, remains the same). The match rate percentage drops substantially with a 2019 fifth-grade-based denominator due to expanding adoption of i-Ready testing over that period, meaning that there were many more students who took the i-Ready mathematics test in 2019 as fifth graders who had not taken it in 2017 as third graders compared to the number of students who had taken it in 2017 as third graders but did not test in fifth grade in 2019. In contrast, match rates for third graders in 2019 and fifth graders in 2021 are similar at 61% and 64%, respectively, with outflow more closely matching inflow.

The second key feature is the representativeness of the matched students, which is indicated by the proximity of the departer and arriver average scores to the matched student averages. As previously discussed, for the third-grade cohort in 2019, the matched students have low coverage and are nonrepresentative, but they are similarly nonrepresentative over time. The outflow of students who had low average scores largely balances an inflow of students with low average scores. Thus, the overall story of progress is relatively unchanged. In contrast, the 2017–2019 trend is biased downward by the disproportionately large influx of arriving students in 2019 with relatively low average scores.

This figure and relevant calculations show that, without checks on coverage and representativeness, test score trends cannot be distinguished from potential changes in populations. Whereas coverage is high and representativeness might be expected to be stable in typical years, the pandemic has caused both declines in coverage and changing representativeness. As tested populations are likely to continue to fluctuate through the recovery, and as other crises or mobility patterns may continue to affect tested populations in the future, we argue for standard reporting of match rates and the representativeness of matched students versus departing and arriving subpopulations.

Discussion

The COVID-19 pandemic has affected which students complete tests, where they test, and how frequently they test. We explained that standard quality checks like reliability, standard errors of measurement, model fit statistics, and DIF statistics cannot detect inflated true scores nor degraded correlations in true scores over time. In our sample of i-Ready fall test-takers, longitudinal statistics enabled us to detect unusually high means and unexpected increases in

variance for scores of students who tested remotely in the 2020–2021 school year by making comparisons to these students' prior data. Accurate interpretation of aggregate test score trends requires a clear reference population and a defensible score scale. i-Ready score users should therefore interpret results from remote testing at the studied grade levels with caution, particularly for interpretations and uses that have high stakes.

We have proposed here that test developers include specific longitudinal data features, including means, standard deviations, match rates, representativeness, and correlations, as standard exhibits in technical documentation. These will enable users to triangulate multiple sources of evidence on data quality and compare current metrics to historical metrics. Other test providers and users that maintain linked student-level information over time will similarly be able to use longitudinal statistics to monitor their test data quality and describe aggregate test score trends, even in nonpandemic testing conditions. The standard suite of metrics is well-suited to provide point-in-time descriptions of test data quality. However, disruptions to traditional testing require an analysis of changes over time to support valid interpretations of educational proficiency and progress.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bacher-Hicks, A., & Goodman, J. (2021, July 13). "The Covid-19 Pandemic Is a Lousy Natural Experiment for Studying the Effects of Online Learning." *Education Next, 21*(4). https://www.educationnext.org/covid-19-pandemic-lousy-natural-experiment-for-studying-the-effects-online-learning/

Cognia, Inc., & Massachusetts Department of Elementary and Secondary Education. (2020a). *2019 Legacy MCAS Technical Report.* MCAS Service Center. http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2019/Legacy%20ADA/2019_MCAS_Legacy_TechReport.pdf

Cognia, Inc., & Massachusetts Department of Elementary and Secondary Education. (2020b). *2019 Next-Generation MCAS and MCAS-Alt Technical Report.* MCAS Service Center. http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2019/NextGen%20ADA/2019_MCAS_NG_TechReport.pdf

Curriculum Associates. (2019). *i-Ready® Assessments Technical Manual.*

Data Quality Campaign. (2021). *In a Year Like No Other, Report Cards Remained the Same.* Retrieved September 8, 2021, from https://dataqualitycampaign.org/show-me-the-data-2021

Engelhard G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Gulliksen, H. (2013). *Theory of mental tests.* Routledge.

Ho, A. D. (2021). *Three test-score metrics that all states should report in the COVID-19-affected spring of 2021*. Retrieved September 8, 2021, from https://scholar.harvard.edu/files/andrewho/files/threemetrics.pdf

Korman, H.T.N., O'Keefe, B., & Repka, M. (2020). Missing in the Margins: Estimating the Scale of the COVID-19 Attendance Crisis. *Bellwether Education Partners.* Retrieved September 8, 2021, from https://bellwethereducation.org/publication/missing-margins-estimating-scale-covid-19-attendance-crisis

NWEA. (2019). *MAP® Growth™ Technical Report*.

Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research, 74*(4), 525–556. https://doi.org/10.3102/00346543074004525

Reardon, S. F., Papay, J. P., Kilbride, T., Strunk, K. O., Cowen, J., An, L., & Donohue, K. (2019). *Can Repeated Aggregate Cross-Sectional Data Be Used to Measure Average Student Learning Rates? A Validation Study of Learning Rate Measures in the Stanford Education Data Archive*. CEPA Working Paper No. 19-08. Stanford Center for Education Policy Analysis.

Sinharay, S. (2021). Score Reporting for Examinees with Incomplete Data on Large-Scale Educational Assessments. *Educational Measurement: Issues and Practice, 40*: 79-91. https://doi-org.ezp-prod1.hul.harvard.edu/10.1111/emip.12396

Smarter Balanced Assessment Consortium. (2020). *2018-19 Summative Technical Report.* Smarter Balanced. https://technicalreports.smarterbalanced.org/2018-19_summative-report/_book/index.html

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1-21. https://doi.org/10.1214/09-STS313