CADRE 2018

THE STATE OF DISTRICT-LEVEL INTERIM ASSESSMENTS

Amy Burkhardt & Derek C. Briggs

A Report Prepared by the Center for Assessment, Design, Research and Evaluation (CADRE), School of Education, University of Colorado Boulder

CADRE

Abstract

This study provides a descriptive overview of the state of district-level “interim assessment programs” across the country. What are typically called “assessments,” we term “assessment programs” which allows us to organize such into three distinct components: How they are designed, assembled into assessment events, and then delivered to students. We define eight categories of assessment programs, with respect to these three components. We survey the largest school districts in the country, and record all of the instances of the various assessment programs reportedly used. We then place these assessment programs into the eight distinct categories, and find that one-third of the 35 unique assessment programs are those where a single test vendor is responsible for all three components of an assessment. Such assessment programs have a reach of over 3.7 million students. Another one-third of the assessment programs fall into the category where the school district is responsible for all three components, with a reach of over 2.3 million students. A notable proportion of assessment programs (8%) are those where either the underlying items are licensed from a third-party item bank, or no information is found regarding the source of the items. Such assessment programs have a reach of over 1 million students. In this study, we also conduct a systematic review of the supporting documentation from the most popular assessment programs, assigning rubric scores based on the claims and evidence made regarding the four different steps in the item development process: design, review, piloting and statistics. Across all assessment programs, claims around the design and review aspects of item development are made more often than claims around piloting and statistics – and, while the item design claims are often supported, the item review process claims are oftentimes made without demonstrating standards alignment.

Introduction

If you visit a school district's website, and navigate to the page that is dedicated to assessments, it will likely offer information on the various assessments administered to students throughout the school year. This paper focuses on one particular subset of these assessments, that, unlike state-mandated summative tests, go by many different names. They are sometimes described as formative assessments[1], or interim assessments[2], or district assessments[3], or local assessments[4], depending on the school district. Other times the school district describes an assessment by its utility: predicting future success, benchmarking student ability, diagnosing learning needs, monitoring student progress and growth, or some other combination of these terms. Despite the variation in naming conventions, the common theme that connects these assessments is that they are intended to be used on a periodic basis (i.e., on more than one occasion during the school year), and their results are assumed to provide accurate information about student progress.

However, the outcome of a recent randomized controlled experiment conducted by Konstantopoulos, Miller, van der Ploeg, & Li (2016) gives pause to such an assumption, given the finding that the use of interim assessments had no significant effect on student achievement in grades 3-8, and a *negative* effect in grades K-2. One possible explanation of this phenomenon could be the questionable quality of the items underlying these assessments, or some combination of the quality of the items, along with teaching practices that go along with the heavy use of such instruments. As Perie, Marion, Gong & Wurzel (2007, p. 9) pointed out some time ago: "It sounds overly obvious to say that the quality of the interim assessment program is dependent upon the quality of the items included in such systems, but this point often gets overlooked."

[1] http://www.houstonisd.org/Page/133169
[2] https://www.sandi.net/staff/assessment-services/district-interim-assessments
[3] https://www.seattleschools.org/academics/assessments
[4] http://www.baltimorecityschools.org/Page/33027

As such, an argument we make in this paper is that the high confidence in these assessments appears to be disproportionate to the evidence readily available to evaluate their quality. To date, there isn't a widely known or accepted framework in place to this end; more specifically, there is no framework that intrinsically focuses on the quality of the underlying items that comprise these assessments. The purpose of this paper is help lay the groundwork for such a framework.

The first goal of this study is to better understand the range of district-level assessments currently in use across the country. In the process, we attempt to develop some standard terminology to use when characterizing different features of the assessment. The second goal is to design and use a rubric to evaluate the available documentation (e.g., technical reports) of the most popular assessments for evidence with respect to the quality of their item development. We first reviewed each piece of documentation, and then assigned a series of scores based on a four-dimension rubric that we created. The rubric scores reflect the degree to which the documentation makes claims and provides evidence regarding four aspects of the item development process: design, review, piloting and statistics.

Components of an Assessment Program

In this paper, the domain of assessments we wish to characterize are those that vary by school district and are offered periodically for the purpose of providing teachers with information about student progress in the subject areas of mathematics and English Language Arts. Each assessment within this domain can be characterized with respect to three components: the assessment design, the specific assessment events (e.g., a unique set of items that define a quiz or test) that need to be assembled, and the delivery mechanism of the assessment (e.g., the assessment platform). These three components are depicted visually in Figure 1. Taken together, these three components characterize not so much an "assessment" per se, but an assessment program. The best practice for the design of an assessment is through the process of first specifying a construct, developing a blueprint, and then selecting and ordering a series of items that ideally align with a test blueprint or some other conceptualization of an underlying construct of student ability. These items are selected from an item bank, and it is this particular aspect of the assessment design process (i.e., the item development process) to which we devote particular attention in this study. The item bank can either be proprietary, and used exclusively for a particular assessment or suite of assessments, or the item bank can be licensed and used by many different vendors to assemble an unlimited number of assessments. The second component of the assessment program represents the process of selecting some of the aligned items stored within the item bank, and assembling them together to form a test or a quiz. This assembled set of questions is then delivered to students, either through the use of an online platform, or by paper.

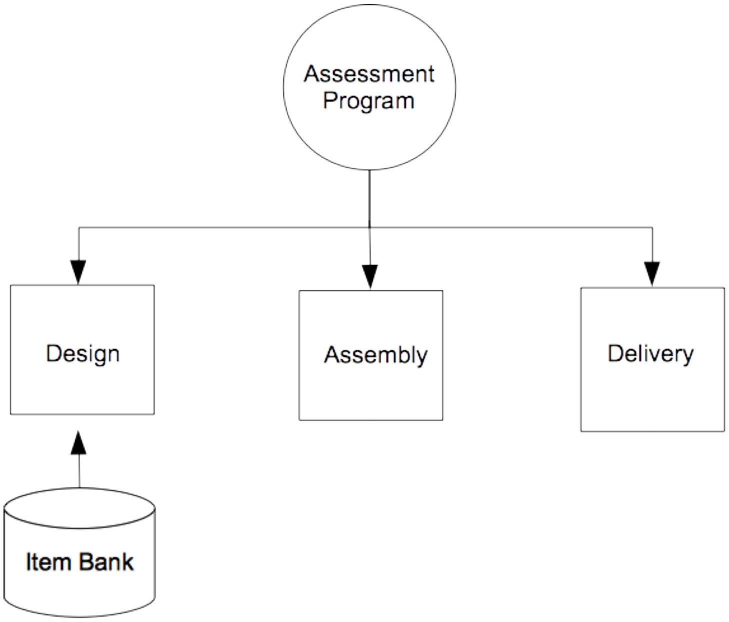

Figure 1 Anatomy of an Assessment

Research Questions

Research Question 1: Scope of Assessment programs

1. What is the national scope of assessment programs used periodically by school districts to provide information about student progress?

a. How can the different systems that school districts rely on be described and categorized?

b. What are the most popular assessment programs, and categories of assessment programs across the country?

c. What are the names of the assessments, the vendors of the assessment platforms, and (when available) the vendors of the item banks of these popular assessments?

Research Question 2: Documentation of Item Quality

2. To what extent does the publicly available documentation from popular assessment programs (1) describe the item development process, and (2) provide evidence to support claims made with respect to the quality of items?

a. Which aspects (design, review, piloting, and statistics) of the item development process are discussed in the documentation? Which aspects are ignored?

b. How often are unsupported claims being made about the different aspects item development process?

c. Within which aspects of the item development process, how often are claims substantiated by evidence? What type of evidence is being provided?

Background on Existing Frameworks for Evaluating Quality of Assessments

There are ongoing efforts to develop rubrics and checklists that examine the quality of K-12 assessments, and, while they all touch on the underlying items of the assessment, they tend to have a broader focus (which is a contrast to the dedicated focus of on items in this present study). The National Center on Intensive Intervention at American Institutes for Research focuses on students with severe and persistent learning or behavioral needs, and has developed rubrics for the “academic progress monitoring tools"[5]. This system, which rates many of the same assessments described in this present study, was developed to assist educators and families in being informed consumers of these assessments. The rating system focuses of criteria from three different aspects of the assessment tools: (1) psychometrics (e.g., disaggregated reliability and validity data – by subgroups), (2) progress monitoring (e.g., whether end-of-year benchmarks specify the level of performance expected at the end of the grade, by the grade level), and (3) data-based individualization (e.g., the tool's ability to help a teacher in planning for and adjusting their instruction to meet student needs). These rated progress monitoring tools are presented at a finer grain than in the present study, breaking down results of each assessment by the content area and the grade. For each rubric, the rating system relies on up to four different shapes to denote a rating: a dash, an empty bubble, a half bubble, or a full bubble.

Other frameworks (or checklists) have been developed to evaluate any K-12 assessment program, including those that feature state-administered standardized tests (e.g., Cizek, Schmid, Kosh, & Germuth, 2016; Shepard, 1977). Take, for example, the most recent checklist (Cizek, Schmid, Kosh, & Germuth, 2016) which was created to aid in the evaluation of K-12 assessments that are intended to measure student learning. The intended audience included educators, policy makers, and those who develop and administer such tests. The five aspects addressed in the checklist are: (1) test development, (2) test administration, (3) reliability evidence, (4) validity evidence, and (5) scoring and reporting. Instead of a rubric, the evaluation framework lists succinct statements covering these aspects, with the following codes: O = Observed, N = Not Observed, and NA = Not Applicable. For example, one evaluation element of the test development checklist is that the item development process is documented. With such a rating scale, the focus is on whether such evidence is available, but not whether such evidence verifies that the assessment tool is of high quality. The evaluation criteria were drawn from the following sources of best assessment practices: *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014), *Standards and Assessment Peer Review Guidance*

[5] https://intensiveintervention.org/chart/progress-monitoring

(USED, 2009), *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004), *Rights and Responsibilities of Examinees* (Joint Committee on Testing Practices, 1998).

Finally, the Buros Center for Testing is an independent, non-profit organization that produces reference materials, such as the Mental Measurement Yearbook, that serve to be "candidly critical evaluations of commercially available tests." The reports are not free to the public, but individual test reviews are available for a small fee. From an inspection of the sample reviews[6], a test undergoes evaluation by two independent reviewers, and the report includes a description of the test, a brief summary of the motivating nature of the development of the test, an overview of the technical aspects of the test, and commentary reflecting the overall quality of the assessment. In the sample reports, the technical documentation includes information on the reliability coefficients and other item statistics, as well as any available evidence and arguments supporting the validity of the test. Aside from the brief test development description in the report, there is no documentation for how the items of the test are designed and reviewed, which are two important considerations within this present study.

Methods

The Scope of Assessment Programs

The first goal of this paper is to describe the landscape of the assessment programs that are reportedly used by districts. This work was carried out by visiting the websites of school districts and documenting which assessments districts reported using for the purpose of monitoring, diagnose, predicting, or benchmarking the academic achievement and progress of their students in the content areas of math and ELA. The initial survey of school districts was performed by visiting the websites of the 50 largest school districts in the country[7] and searching for the page dedicated to the department responsible for providing information on which assessments are administered, and when during the school year they are administered. Twenty-six of the 50 largest school districts had this information

[6] http://buros.org/review-samples

[7]https://en.wikipedia.org/wiki/List_of_the_largest_school_districts_in_the_United_States_by_enrollment

readily accessible on their website. After completing this first phase, it was apparent that many regions of the country were not represented by this approach, and so we purposefully identified the two largest school districts in every state. When this information wasn't available for either of the top two school districts in a state, we continued on to the website of the next largest school district, until we could find assessments listed for two districts within each state (there are a few state exceptions, where no information was found to be available for compilation on any districts' websites). In total, we gathered assessment information from the websites of 70 school districts.

The information provided by the school district websites was then documented in a spreadsheet, which included the name of the assessment, the keywords that were used to describe it, and the url where this information was gathered. Latitude and longitude of the school district, as well as the size of the school district was also gathered to be used in visual displays of our findings.

For each assessment, we sought out the following details with respect to the three components of an assessment program:

1. Who designed the assessment?
    a. Are the items that comprise the assessment from a proprietary source or from a licensed item bank?
2. Who assembles assessment events to be delivered to students?
3. Who delivers the assessment?

From the answers to these questions, we defined eight program categories that characterized the 35 unique assessment programs reported from school districts. These eight categories are presented in Table 1.

Figure 2 presents the regional locations of the assessment programs within the eight categories; each circle represents an instance of an assessment program within a given category. The varying diameters of the circles represent the size of the school district; each color represents a different assessment program. In total, 106 instances of assessment programs were reported; 68% of school districts' (N=47) websites reference just one assessment program, but for 32% (N=22), multiple assessment programs were mentioned.

Table 1. *Seven Categories of Assessment programs*

Assessment Program Category			**Description**	**Example**
Item Bank	Assembly	Delivery		
District	District	District	Created by the school district or state, and delivered on a platform owned by the school district (or otherwise not disclosed)	STAAR interim assessments of Texas
District	District	Vendor	Created by the school district or state, but delivered through a third-party system.	Denver Public Schools and Illuminate
Vendor	Vendor	Vendor	Created by a vendor, using a proprietary item bank, and delivered by same vendor.	MAP (NWEA)
Teacher	Teacher	Vendor	Crated by teachers and delivered by a third-party vendor.	PM Nation
Purchased/Licensed	Vendor	Vendor	Assembled and delivered by a vendor, but it is often times unclear exactly which licensed item bank the items are from (the vendor's website may list many different possible item bank partners).	Synergy
No Information	Vendor	Vendor	Assembled and delivered by a vendor, but it is unclear where the items came from (that is, the product website doesn't discuss item development, and no technical report could be found online)	mClass: Math
Research Institution/ University	Research Institution/ University	Other	Designed and assembled by a research institution and delivered by another test vendor, or by paper.	DIBELS
Research Institution/ University	Research Institution/ University	Research Institution/ University	Designed and delivered by a research institution, and delivered by the same.	easyCBM

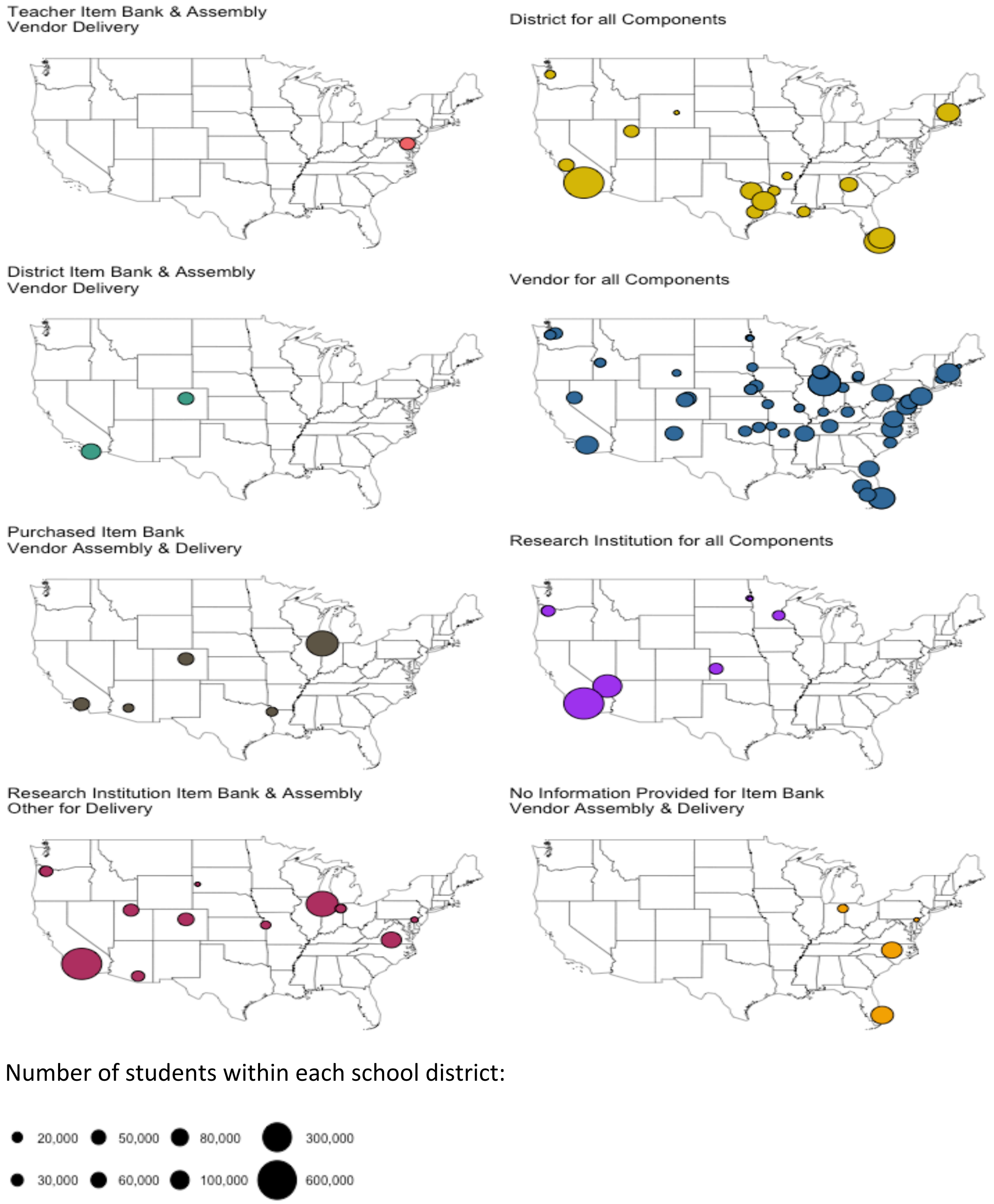
Teacher Item Bank & Assembly
Vendor Delivery
District for all Components
District Item Bank & Assembly
Vendor Delivery
Vendor for all Components
Purchased Item Bank
Vendor Assembly & Delivery
Research Institution for all Components
Research Institution Item Bank & Assembly
Other for Delivery
No Information Provided for Item Bank
Vendor Assembly & Delivery
Number of students within each school district:
20,000
50,000
80,000
300,000
30,000
60,000
100,000
600,000

Figure 2 Assessment Program Categories Across the Country

Table 3 presents the both the unique count of reported assessment programs within each of the eight different categories (Unique Assessments with Category) as well as the Number of Times School Districts Reported Assessment within Category. For example, the first assessment program category (Vendor for all Components) consists of ten unique assessments, and these assessments were mentioned a total of 58 times on school district websites. The table also provides the names of up to three of the most popular assessment programs in each category (in the event that there are fewer than three assessments for a given category, all assessments are presented). Fourteen of the 35 assessment programs were assembled and delivered by a vendor. For ten of the 14 assessments, it was clear that the vendor was directly responsible for the original test items that led to an operational item bank. For the remaining four assessments, the vendor was not directly responsible for the item design; instead three assessment programs relied on one or more preexisting item banks from another source, and one assessment provided no information about the item design at all (that is, we could not find any information about the item development process on the assessment program's webpage or through an internet search).

The other 21 assessment programs presented in Table 2 did not rely on a vendor for either the item development or assessment delivery process. Twelve of these assessments were characterized as district-related assessments (two of these were delivered by a third-party platform). This means that over 1/3 of the assessments were done "in house," by a school district.

Most of the remaining assessment programs belonged to the assessment programs characterized by researchers developing the items and assembling the assessments.

Table 3. *Assessment programs broken down by Category*

Assessment Program Category	**Unique Assessments within Category (%)**	**Number of Times School Districts Reported Assessment within Category (%)**	**Most Popular**	**Assembly Entity**	**Delivery Entity**
Vendor for all Components	10 (29%)	58 (55%)	MAP STAR 360 i-Ready Diagnostics	NWEA Renaissance Learning Curriculum Associates	NWEA Renaissance Learning Curriculum Associates
District for all Components	12 (34%)	15 (14%)	LEAP STAAR	Louisiana Texas	Not Disclosed Not Disclosed
Research Institution/University for Item Bank & Assembly; Other for Delivery	2 (6%)	14 (13%)	DIBELS SRI	DIBELS SRI	Paper or mCLASS Houghton Mifflin Harcourt
No information for Item Bank; Vendor for Assembly & Delivery	1 (3%)	3 (3%)	mClass: Math	Amplify	Amplify

Assessment Program Category	Unique Assessments within Category (%)	Number of Times School Districts Reported Assessment within Category (%)	Most Popular	Assembly Entity	Delivery Entity
Research Institution/University for all components.	4 (11%)	8 (8%)	easyCBM FastBridge Learning Smarter Balanced	Uni. of Oregon Uni. of Minnesota Smarter Balanced	easyCBM FastBridgeLearning Smarter Balanced
Licensed Item Bank*; Vendor for Assembly & Delivery	3 (9%)	5 (5%)	Illuminate Synergy CIM	Illuminate EduPoint CIM	Illuminate Edupoint CIM
District for Item Bank & Assembly; Vendor for Delivery	2 (6%)	2 (2%)	Denver Public Schools San Diego Unified SD	Denver Public Schools San Diego Unified SD	Illuminate Illuminate
Teacher for Item Bank & Assembly; Vendor for Delivery	1 (3%)	1 (1%)	PM Nation	Teachers	Performance Matters
Total	35	106			

*Note: The item banks include the following: CenterPoint, Fluence, Measured Progress, Inspect and Navigate

Evidence of Item Quality

We further explored the assessment programs that are listed as the “Most Popular” within each of the categories presented in Table 3 by collecting and conducting a systematic review of their supporting documentation. This review of the documentation focused on the claims and evidence made about the four different steps in the item development process: item design, item review, item piloting and item statistics. We applied a simple rubric to each of these four dimensions; the results of which serve as an overview of how often the different aspects of item development are discussed in the documentation, and to what extent claims about item quality are substantiated by evidence. Examining only three assessments within each category isn’t intended to be an exhaustive investigation into the documentation of item quality, but this evaluation serves to provide a glimpse into the variability of the documentation of item quality across different assessment programs and categories, as well as to provide an opportunity to demonstrate our coding scheme and rubric to evaluate such documentation.

As mentioned above, there are four aspects of the item development process that are the focus for reviewing the evidence of item quality. Below are exemplar expectations of documentation that provide supported claims for each aspect of item development process.

1. **Item Design.** Items design is guided by standards and/or a blueprint and/or a learning progression. Therefore, individual items can be traced back to specific standards, specifications, and/or locations on a learning progression.
2. **Item Review.** Items are reviewed by content experts, and criteria are set in place to ensure that students are presented with items that will accurately reflect learning. Items are reviewed periodically; there is some sort of feedback system has been implemented so that items can be continuously reviewed. In the event that an item bank has been inherited from some other source, all of these existing items (in addition to new items) undergo review.

3. **Item Field Testing.** Items are piloted prior to being administered to students. All items in the large item banks are field tested; not just a subset of them. The population that the items are field tested on should be representative of those intended to be presented with the items, rather than a convenience sample.
4. **Item Statistics**. Item statistics include classical test statistics of Cronbach's alpha (or some other measure of reliability), descriptive statistics, p-values and point-biserial correlations. The reliability estimate is indicative of a high proportion of true score variance, and the point-biserial correlations are at least moderately strong. There is a range of p-values, reflecting variability in the difficulty of the items. These statistics are not merely presented in the aggregate (an example of this is when p-values are reported as the average across all math items for a given grade). Instead, the item statistics are broken down to some subgroup of items, if not completely in the disaggregate of all items. In instances where the data have been fit to an IRT model, the discrimination parameters, item difficulty parameters, and standard errors are presented in the disaggregate. Furthermore, evidence is provided to show that data fit the model, and there is a range of values for the item difficulty estimates.

Reviewing and Scoring Assessment Programs for Item Quality

The process for evaluating the associated documentation of each assessment is as follows. We first identified the entities associated with the item banks and assembly components of the most popular assessment programs (in some cases these were from the same source). We then visited all associated websites to find information regarding the development of the items that comprise the assessment program. When technical reports or marketing collateral were available, they were downloaded. Screenshots of the entire website were also taken. Additional internet searches were made to gather as much information as possible regarding the documentation of the quality of items. We also sent an email to entities associated with the most popular assessments and requested a technical report.

After all of the available documents were gathered for each assessment program, every artifact was imported into the qualitative coding software MAXQDA, reviewed and coded (i.e., annotated), using the qualitative coding scheme provided in Appendix A. Coding the artifacts according to the coding scheme ensured that we captured the documents' claims and evidence surrounding the four different aspects of item development. Once all collateral and supporting documentation for each popular assessment had been annotated, the coded segments of the text were exported to Excel files[8]. The segments were then reviewed, and scored following the criteria described within the dimensions of the rubric (presented in Appendix B).

Although the details of the criteria for the scores of 0, 1, or 2 are specific to each dimension, the general rules hold across almost all aspects of item development: We assigned a score of 0 if no claim was made about one of the four aspects of item develop, we assigned a score of 1 if a claim was made, but there was no evidence substantiating it, and a score of 2 was assigned when the documentation included evidence supporting the claim. For example, in order to receive 2 points for Item Design, the documentation must state that all of the items are aligned to the Common Core State Standards, and it also must provide evidence, such as a table that crosswalks all of the items their aligned standards. The one departure from this general rule is for the statistics domain, where statistics reported in the aggregate receive 1 point, and disaggregated statistics receive 2 points.

Overall Findings from the Rubric

Table 4 presents the rubric scores for all of the popular assessment programs across all of the categories. The assessment programs within the category Research Institution/University for all Components, all received the highest score of 8 points. The assessment programs within the category No information for Item Bank; Vendor for Assembly and Delivery received zero points, because – although there was a website and marketing collateral for the assessment program – no information could be found related to the item development process for these items. The scores for the

[8] Coded segments available upon request.

assessment programs that rely on a licensed item bank were also low. Most often, the documentation of the item banks would provide some details of how the items are reviewed, but only stated that items were aligned, without any demonstration of this work. The most popular assessment programs within the most popular category (Vendor for all Components) had total scores that ranged from 3 to 7. Only one of the assessment programs within this category responded to the email request and provided a technical report (instances when a report was provided upon request is denoted with † in the table; § denotes that some item development information, other than a technical report was provided upon request; ‡ denotes that a request was made, but there was no response). Therefore, information regarding the item quality of the other two assessment programs within this category was restricted to publicly available information, such as marketing collateral and text on the websites.

Table 4. *Rubric Scores for all Assessment programs*

		Score: 0-2				
Assessment Program Category	**Name**	**Design**	**Review**	**Piloting**	**Statistics**	**Total**
Vendor for all components						
	MAP‡	1	2	0	0	3
	STAR 360†	2	2	2	1	7
	i-Ready‡	1	2	2	2	7
District for all Components						
	LEAP§	2	0	0	0	2
	STAAR‡	2	0	0	0	2
Research Institution/University for Item Bank & Assembly; Other for Delivery						
	DIBELS*	2	2	2	2	8
	SRI‡	0	1	2	0	3
No information for Item Bank; Vendor for Assembly and Delivery						
	Mclass: Math‡	0	0	0	0	0
Research Institution/University for all Components						
	easyCBM*	2	2	2	2	8
	FastBridge Learning*	2	2	2	2	8
	Smarter Balanced*	2	2	2	2	8
Licensed Item Bank; Vendor for Assembly and Delivery						
	Illuminate§	1	0	0	0	1

		Score: 0-2				
Assessment Program Category	**Name**	**Design**	**Review**	**Piloting**	**Statistics**	**Total**
	Synergy‡	1	0	0	0	1
	CIM*	1	0	0	0	1
	Fluence‡	1	0	0	0	1
	Measured Progress§	1	2	0	0	3
	Inspect†	1	2	1	1	6
	CenterPoint§	1	1	0	0	2
	Navigate§	1	2	0	0	3
District for Item Bank & Assembly; Vendor for Delivery**						
	Denver Public Schools	0	0	0	0	0
	San Diego Unified SD	0	0	0	0	0
Teacher for Item Bank Assembly; Vendor for Delivery**						
	PM Nation	0	0	0	0	0

*Denotes that technical report was found online without sending an email request.

** Denotes that no documentation at was available for this assessment program category (e.g., no website was dedicated to these assessments programs)

† Denotes that technical report (or online location) was provided upon request.

‡Denotes that a request was made, but there was no response.

§ Denotes that some item development information, other than a technical report, was provided upon request.

Figure 3 presents the frequency of the scores of 0, 1 and 2 for each of the four domains of item development rubric, across all 21 assessment program-related entities.

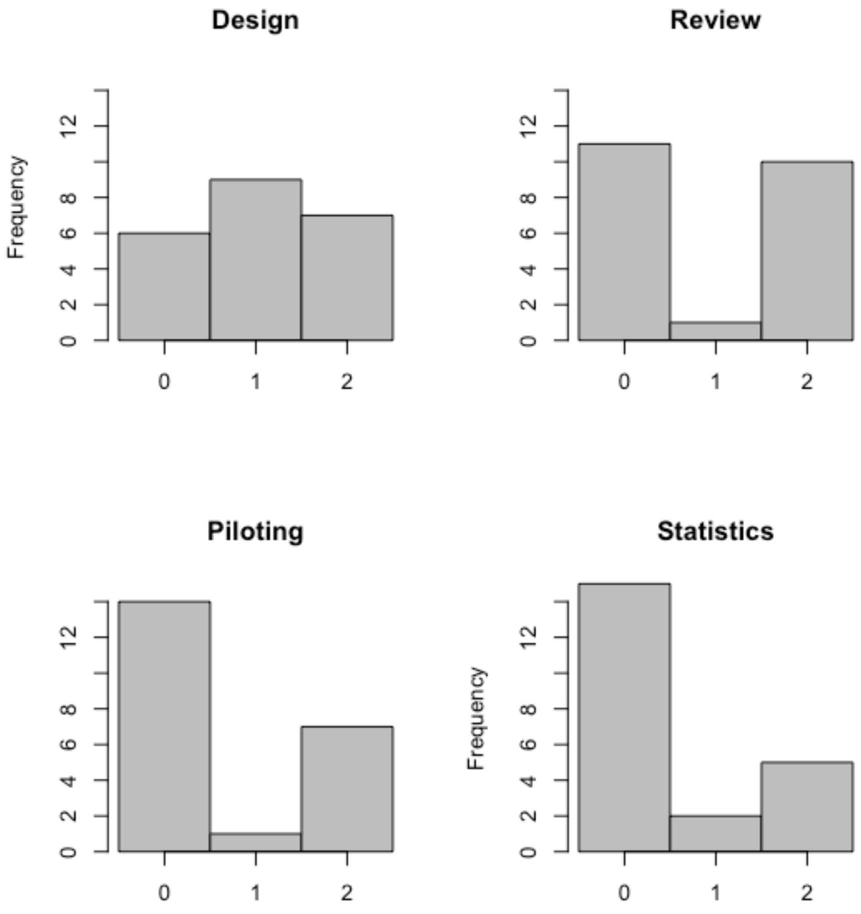

Figure 3 Rubric Scores (0-2) for Four Dimensions

Item Design

Findings from the documentation around the design step of item development revealed close-to-equal distributions across assessment programs that (1) made no claims about item design, (2) made unsubstantiated claims, and (3) demonstrated their claims about item alignment.

Below are some examples from assessment programs that received two points for this category. The technical reports of the assessment programs within the Research Institution/University detailed the item design process and provided a table to demonstrate an item-by-item crosswalk of how the items are aligned to some standards (e.g., DIBELS demonstrates alignment to both literacy standards and Common Core Standards). The Smarter Balanced Interim Assessment is comprised of previously administered items from their summative tests, and therefore provides blueprints and standards alignment for all of their items9. In addition to this item-to-standards alignment, a learning progression or learning model was sometimes also provided to demonstrate the underlying theory of the learning. For example, Renaissance's STAR (Vendor for all components) technical manual refers to a learning progression that guided the development of the items that are intended to measure progress in reading skills[10].

Assessment programs received a score of 1 point when the supporting documentation of the claims alignment wasn't presented. For example, in the document "A Comprehensive Guide to MAP K-12 Computer-Adaptive Interim Assessment," NWEA stated that "MAP is aligned to all state content standards, including the Common Core State Standards, so it provides educators with high-value comparative data and the ability to project proficiency on high-stakes tests." [11] However, these claims could not be substantiated by reviewing the publicly available documentation; instead other documents on NWEA's website simply reiterate the claim without providing access to the evidence:

A blog post states that an independent study concluded that more than 97% of MAP growth items align to the Common Core State Standards in math and English, but no link is provided to the study.[12]

[9] http://www.smarterbalanced.org/assessments/

[10] https://resources.renlearnrp.com/us/manuals/sr/srrptechnicalmanual.pdf

[11] https://www.nwea.org/content/uploads/2014/07/Comprehensive-Guide-to-MAP-K-12-Computer-Adaptive-Interim-Assessment

[12] https://www.nwea.org/blog/2018/study-concludes-map-growth-items-align-common-core-state-standards/

A video on the website describes how their interim assessment “Common Core MAP” aligns to the CCSS by saying, “NWEA’s content specialists first evaluate the standards to understand the cognitive complexity levels required by each standard and then use that information to inform our item development and innovation efforts. Every item we develop for the Common Core test is carefully reviewed by content specialists, rated with a DOK [Depth of Knowledge] category, and compared with a DOK range and target of the standard. This is standardized DOK process that was developed by Dr. Norman Webb.”[13] This video describes a design process but doesn’t adequately demonstrate how or to what extent individual MAP items are aligned to CCSS standards.

Item Review

For the item review aspect, almost half of the assessment programs supported their item review process with evidence (such as identifying content experts or providing a visual of the item review process), while the other half didn’t even mention that there was a review process for their items. The details of the review process varied across assessment programs. Perhaps the most demonstrative assessment program was DIBELS, in that the entire new technical manual for DIBELS Next was devoted to the describing how every item had been revamped since the 6th edition, based on evidence from various studies, and described how each measure (e.g., phoneme segmentation fluency) was changed[14] (It should be noted, of course, that DIBELS consists of a much smaller number of items that some other assessment programs, so the documentation of a thorough item review is a more manageable task than compared to a large item bank.)

Item Piloting

Over 60% of the assessment programs don’t mention item piloting in their documentation. However, the few programs that mentioned field testing supported this claim by providing details about the process. Smarter Balanced Interim Assessments and Renaissance’s STAR 360 received 2

[13] https://www.nwea.org/assessments/test-item-development/content-validity/
[14] https://dibels.org/pubs.html

points for this this dimension, because both assessment vendors described a process that suggested that all items were being field tested with a group of students which was representative of those who would be taking the assessment. Not all assessment programs had all of their items field tested, and some relied on a convenience sample and/or only field tested a subset of their items. EasyCBM, for example, reported that to field test some reading measures, they relied on a convenience sample from voluntary teacher signup, with schools located in the Northwest, Montana, Florida, Texas and Illinois.[15] Such documentation of the field testing process also received the maximum score of two, so as to not disparage or penalize documentation that provides details and transparency to their item development process.

Item Statistics

Of the four aspects of item development, item statistics were mentioned with the lowest frequency; only five of the 21 assessment programs providing disaggregated statistics in their reports and documentation. The documentation for Smarter Balanced interim assessments, for example, broke the item statistics down by the claim of the test; FastLearning disaggregated the results of the test by subtest. A score of 1 was assigned for documentation that presented on statistics, but only did so in the aggregate, such as KeyData System who reported classical item statistics, grouped by grade and content area.

Discussion

By re-defining an assessment in terms of its three components and calling it an “assessment program,” we were able to take all of the periodic assessments reportedly used within 70 of the large school districts and categorize them into these eight categories: (1) Vendor for all Components, (2) District for all Components, (3) Research Institution/University for Item bank & Assembly for all Components, (4) Research Institution/University for Item Bank & Assembly; other for Delivery, (5)

[15] https://files.eric.ed.gov/fulltext/ED547422.pdf

Licensed Item Bank; Vendor for Assembly & Delivery, (6) No Information for Item Bank; Vendor for Assembly & Delivery, (7) District for Item Bank & Assembly; Vendor for Delivery, and (8) Teacher for Item Bank & Assembly, Vendor for Delivery.

These categories not only serve as a way to standardize the language around assessment programs, but by classifying assessment programs according to this structure, we uncovered some interesting descriptions about the most commonly used assessment programs in this country.

For example, the ten assessment programs within the Vendor for all Components category make up 55% of the reported assessment programs across the country, which is a reach of over 3.7 million students within our sample of school districts. Undoubtedly, the reach of these assessment programs that are fully-owned by a single vendor are prevalent across the country. However, this categorization exercise revealed that these ten assessment programs comprise only one-third of all of the unique assessment programs across the 8 categories. This means that there are 25 other assessment programs that are being used in at least one school district across the country. Another interesting finding from this exercise is that we found that there are 12 assessments that are fully-owned by districts, which accounts for one-third of all of the unique assessments (a student reach of over 2.3 million). This categorization process also revealed that there are number of assessment programs where the item bank is licensed from some third-party vendor, or the source is unknown (a student reach of over 1 million students). These items are typically stored in very large item banks (upwards of 35,000 items), and – referring to findings from the second part of our investigation – the item development process isn't very well documented and transparent to the public (i.e., a lot of 0s were assigned to the four dimensions of item development). Whether or not such items are of high quality is still unknown at this point in our investigation, but such findings should prompt school districts to request additional evidence regarding item quality (if they don't already do so).

An additional source of utility for these categories is that they were useful in shaping the methodology of our documentation review, where we focused on the most popular assessment programs within each category.

The documentation review provided valuable insight across all of the assessment programs, as to what types of evidence seems to be lacking when item developers and associated entities write about the item development process. We found that the claims about design and development were mentioned more than that of piloting and statistics. However, claims around item design (e.g., "items are aligned to the Common Core State Standards") were often met without any evidence. Providing sufficient evidence for statistics was also missing from most documentation. In the same way that we found this documentation review helpful for illuminating quality evidence gaps in the programs we reviewed, we hope that the coding scheme and rubric can also serve to guide school districts in evaluating the documentation of their interim assessment programs – not as an end-all tool, but to help inform their conversations around the quality of their assessment program or as a starting point for weighing investments in assessment products.

In regard to the rubric dimension of piloting items, the scoring criteria didn't penalize documentation that reported a convenience sample for field testing. This piece of criteria reveals an important characteristic about the nature of this rubric: This rubric does not evaluate the quality of the evidence, but rather favors details and specificity in the item development process. In developing the rubric, we were careful to not undervalue transparency and honest documentation of item development practices. Transparency from test makers is important in establishing trust of their item development processes, and the provision of thorough the documentation signals that test makers might be willing to have their procedures reviewed and vetted.

This descriptive endeavor of understanding the breadth of the assessment programs and taking a first look at the quality of item development documentation is met with the following limitations. First, everything in this study is based on publicly available documentation. Some issues identified might be sins of omission (evidence regarding quality item design exists, it just isn't made public) vs. sins of commission (item quality is in fact poor). But sins of omission are still problematic. Another limitation is that the districts were a purposeful sample; our focus was on large districts, and our findings may not generalize to smaller districts with more limited resources. Finally, the rubric is new and has not yet been evaluated for inter-rater agreement.

Now that we have made some preliminary progress in developing a deeper understanding of the state of the assessment programs and their supporting documentation across the country, it is important to return to the overarching goal of this project, which is to ensure that the items that comprise “interim assessment programs” accurately reflect student learning. Therefore, even in these early stages, we are gathering data points that can help determine “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA/APA/NCME, 2014). However, the evidence that we have collected so far relies on artifacts from test makers, and -- it might go without saying -- that what has been done in this paper is just on one piece related to larger issue of the validity of interpretations and uses of scores from the assessment programs.

Bibliography

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

Cizek, G. J., Schmid, L. A. Kosh, A. E., & Germuth, A. A., (2016). A checklist for evaluating K-12 assessment programs. Kalamazoo: The Evaluation Center, Western Michigan University. Available from http://www.wmich.edu/evaluation/checklists

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations.* Washington, DC: American Psychological Association, Joint Committee on Testing Practices.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* Washington, DC: American Psychological Association, Joint Committee on Testing Practices.

Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness*, *9*(sup1), 188-208.

Perie, M., Marion, S., Gong, B. & Wurzel, J. (2007). The Role of Interim Assessments in a Comprehensive Assessment program. The Aspen Institute. https://www.achieve.org/files/TheRoleofInterimAssessments.pdf

Perie, M., Marion, S., & Gong, B. (2009). Moving Toward a Comprehensive Assessment program: A Framework for Considering Interim Assessments. *Educational Measurement: Issues and Practice*, Vol. 28, No. 3, pp. 5–13

Shepard, L. A. (1997). *A checklist for evaluating large-scale assessment programs* [Occasional Paper No. 9]. Kalamazoo, MI: Western Michigan University, The Evaluation Center.

Appendix A: Coding Scheme of Assessment Collateral

Below describes the coding scheme used to evaluate the quality of the collateral of the most popular assessments. The main codes are design, review, pilot, statistics, and validity. Each code then consists of a number of sub-codes describing the process.

1) **Item Design**
   a) Source of Items is explicitly stated: created in-house
   b) Source of items is explicitly stated: taken from a third-party item bank, but does not specify which one.
   c) Source of items is explicitly stated: taken from a third-party item bank, and specifies which item bank.
   d) Standards alignment is mentioned
   e) Standards alignment is demonstrated
   f) Test blueprint is mentioned
   g) Test blueprint is presented
2) **Item Review**
   1. Existence of item review process
   2. Criteria for reviewing existing items in item bank
   3. Criteria for reviewing newly developed items
3) **Item Field Testing**
   a) Items are piloted with a sample of the intended population.
   b) All items in the bank are piloted
   c) Only a subset of the items in the bank are piloted
4) **Item Statistics**
   a) p-values
   b) point-biserial
   c) Cronbach’s alpha
   d) Item statistics are presented in the aggregate (e.g., mean point-biserial correlation for all middle school math items)
   e) More detailed item statistic information is offered (not in the aggregate)
5) **Validity**
   a) Proposed use of items
   b) Evidence of theory supporting validity
   c) Evidence based on test content
   d) Evidence based on response processes
   e) Evidence based on relation to other variables

Appendix B: Rubric for Scoring Documentation of Item Quality

Score	0	1	2
Design	No mention of item alignment to blueprints, standards, or learning progression.	Claim of item alignment is made is mentioned, but there is no demonstration of alignment.	Claim is made and supported with details of process or by evidence (e.g., an alignment study or a table that presents a cross walk of items to standards).
Review	No mention of how items are reviewed.	Claim of item review process is made but there are no details of the process.	Claim is made and supported with details of the review process or by evidence
Piloting	No mention of how items are piloted or field tested.	General claims of piloting without any details.	Claims items within bank have been reportedly piloted with a sample of students representing the test taking population. In the event that only a subset of items have been field tested, the subset of items are described. In the event that the field testing relied on a convenience sample, the demographics of the sample are described.
Statistics	No mention of item-level statistics	Presents classical test statistics or IRT parameters on some of the items, but only in the aggregate (e.g., by content area, only broken down by grade).	Presents classical test statistics or IRT parameters, but in the disaggregate (broken down by subsets of items or individual items are presented).