International Journal of Testing

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hijt20

Designing classroom assessments with the end user in mind

Leanne R. Ketterlin-Geller, Jennifer McMurrer & Tina Barton

To cite this article: Leanne R. Ketterlin-Geller, Jennifer McMurrer & Tina Barton (2024) Designing classroom assessments with the end user in mind, International Journal of Testing, 24:2, 103-129, DOI: 10.1080/15305058.2024.2306549

To link to this article: https://doi.org/10.1080/15305058.2024.2306549

INTERNATIONAL JOURNAL OF TESTING
2024, VOL. 24, NO. 2, 103–129
https://doi.org/10.1080/15305058.2024.2306549

Routledge
Taylor & Francis Group

Designing classroom assessments with the end user in mind

Leanne R. Ketterlin-Geller[a], Jennifer McMurrer[b] and Tina Barton[c]

[a]Department of Education Policy & Leadership, Southern Methodist University, Dallas, TX, USA; [b]Research in Mathematics Education, Southern Methodist University, Dallas, TX, USA; [c]Duke Energy Corporation, Dallas, TX, USA

ABSTRACT

Teachers are key users of classroom assessment data; however, their needs and preferences are often overlooked in the design and development of the assessments themselves. We used principles of Human-Centered Design (HCD) to systematically solicit, study, and integrate teachers' needs in the design and development of a classroom-based early mathematics assessment. We recruited 18 teachers to participate in a series of research activities that align with HCD methodology. Data informed the articulation of the use case and test specifications, thereby elevating teachers' voices to improve test-development decisions that consider the end users from the start.

KEYWORDS

Designing classroom assessment; human-centered design; mathematics instruction; use case

Designing assessments is a complex process that involves multiple components and is informed by various perspectives and constraints. Essential components that drive assessment design and development decisions include alignment with the intended purpose and uses, accordance with content representation and relevance, assurances of accessibility and fairness across the tested population, and adherence to technically rigorous standards (American Educational Research Association et al., 2014). Also important to the decision-making process is the involvement of stakeholders who will use and interpret the results (Gulikers et al., 2009; Ryan, 2002). For large-scale summative assessments, these stakeholders may include policy makers at the state or local level who will use the results

CONTACT Leanne R. Ketterlin-Geller lkgeller@smu.edu Department of Education Policy & Leadership, Southern Methodist University, Dallas, TX, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/15305058.2024.2306549.

to inform decisions related to evaluating programs and allocating resources (Kane, 2013). In the case of classroom-based assessments, important voices include local stakeholders, such as teachers, students, and parents, who may use or interpret the results to guide instruction, target effort, or provide supplemental support, respectively.

International guidelines on test development often reference the critical role stakeholders' input plays at various points in the test development process. For example, the International Test Commission's Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores (2012) explicitly call attention to collecting data from stakeholders via focus groups, think-aloud or individual interviews, or other means about the comprehensibility and interpretability of data conveyed on score reports (see Standard 2.5.1.1). Seeking stakeholders' input during other phases of test development is not emphasized in these guidelines, and is often not seen in practice. For example, McDonald et al. (2021) report on a 3-year collaboration between researchers and stakeholders at a local school district to evaluate the teachers' perceptions of feasibility and usability of two existing early grades mathematics assessments. This example illustrates the importance of including teachers' voices when selecting and implementing existing assessments within a novel context; however, teachers' input was not considered from the initiation phases of instrument development.

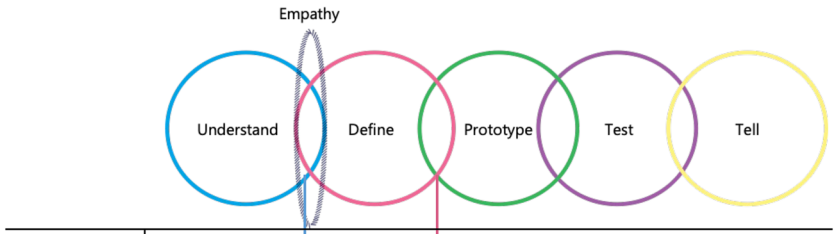
Without considering teachers' perspectives during all phases of instrument development, assessment design and development decisions may inadvertently lead to results that are less useful to the end user. At least two plausible unintended consequences of this situation are that (a) stakeholders deem the assessments ineffective for addressing their needs, and thus discontinue their use, or (b) stakeholders use the assessment results in unintended ways, thereby compromising validity. Given the expense (e.g., fiscal, human capital, time) involved in test development and implementation, both outcomes are untenable. In this manuscript, we describe a novel application of Human-Centered Design (HCD) to test development. We applied HCD principles as a way of systematically integrating teachers' perspectives in the design and development phases of classroom-based assessments for two early mathematics constructs, with an emphasis on specifying the use case and delineating design specifications.

HCD: a process for integrating teachers' voices in test development

HCD, also referred to as user-centered design, is a process to understand the needs and wants of the end users of an object, product, system, or service, and is often used in product design and technology or interface

design. A series of phases is strategically used to gather data to allow developers to better understand users' experiences within a given context with the goal of improving a process, product, or experience (Nielsen, 1993). HCD draws from both qualitative and quantitative methods of research practiced in ethnographic and social sciences to approach problem solving in a way that centers on the intended users' experience to inform design solutions and outcomes. According to Giacomin (2014), HCD is "based on the use of techniques which communicate, interact, empathize and stimulate the people involved, obtaining an understanding of their needs, desires and experiences which often transcends that which the people themselves actually realised" (p. 609). The resulting insights are them applied to define the core problem being solved for the user and address key aspects of their experience to improve usability and adoption of the product or solution.

HCD can be conceptualized as involving five general phases of implementation: understand, define, prototype, test, and tell (see Figure 1). Within the first phase of the HCD methodology, researchers seek to understand deeply a given context by engaging with participants and environments to generate empathy. Secondary and primary research is conducted to collect data that can help define a design opportunity within the given context. By including primary sources of research such as interviews, focus group data, and observations, the end users remain central to the process, and the development of future prototypes and products is grounded in the actual users' perspectives and needs.



Research Questions	What instructional decisions in early mathematics do K-2 teachers need and/or want to make that can be guided by assessment data?	How do K-2 teachers currently perceive of assessments and the role of data in their decisions?	Phase 1 What are teachers' perceptions of the usefulness of learning progressions-based classroom assessments? What test specifications support teachers' use and engagement with the assessments? Phase 2 Which aspects of the initial prototype are useful and usable, and which aspects are not?
Methodological Approach	•Two focus groups •Prioritization activity	•Love Letter/ Break-up Letter •Affinity diagramming •KJ Technique	Phase 1 •Focus groups •Design charettes Phase 2 •Focus group •Participatory co-design

Figure 1. Human-centered design process and methods used by phase.
Note. Adapted from Professors Kate Canales and Gray Garmon's instruction in the Southern Methodist University Master of Arts in Design and Hasso Plattner Institute of Design at Stanford (2020).

Once data are collected and externalized, researchers analyze data to identify themes. These themes then can be synthesized to harvest insights related to the needs of the end user and identify opportunities for prototyping (or testing to learn with a low fidelity version of the product) to answer remaining questions. Synthesis of the data and themes identified involves both the application of analytical thought and developing holistic understanding of the needs of the users. By developing a deeper understanding of the context and needs, designers can begin to develop prototypes that will directly or indirectly address the users' given needs.

After prototypes have been developed (often through rapid, low-resolution iteration), they are tested to find answers to further questions and iterated on to solve for the greater design challenge. Once testing of prototypes is complete and findings are determined, solutions and possible recommendations can be made through the tell or dissemination phase.

As illustrated above, HCD offers an approach to integrating end users' perspectives throughout the design and development of a process, product, or experience; yet, it is not commonly applied in educational contexts. In this manuscript, we describe the application of HCD to test development with the aim of designing classroom-based assessments that are more reflective and responsive of the end users' needs. Beginning with establishing a use case, we integrated teachers' voices throughout the test development process. Although we follow established guidelines for test development (c.f., Lane et al., 2016), the application of HCD in this context is novel. Our goal is to illustrate how this methodology can be rigorously applied to improve test development decisions.

A use case for classroom-based assessments

Use cases are commonly used in software development or for business processing systems as a mechanism for incorporating the voices of key stakeholders and users in the documentation of the goals for the system. During the process of articulating a use case, end users work with developers to create a shared vision for and understanding of how the end users will interact with the system and how the system will respond to these users' actions (Cockburn, 2000). A useful analogy is a wheel, in which the use case is the hub from which the system-level requirements radiate like spokes. The concept of a use case has been sparingly applied in the field of educational assessment but shows promise for supporting research–practice partnerships (Penuel & Watkins, 2019) and interdisciplinary development teams (Penuel et al., 2014).

When applied to test development, a use case specifies how the end-users want to use and interact with the test results. Within the process of using classroom-based assessment data to guide instruction, teachers are

the end users and want to make specific interpretations (e.g., monitoring student progress) and resultant actions (e.g., planning next steps in instruction, forming heterogeneous or homogeneous groups). Moreover, temporal considerations relate to their actions (e.g., reteach concepts prior to introducing new information). Test design and development decisions can support the end users' intended actions and interpretations of the test scores. It follows that precisely delineating the use case for an assessment system may enhance the validity of the uses and interpretations of the results.

A priori consideration of the intended uses and interpretations of test scores is the first step when designing an assessment. The Test Standards (American Educational Research Association et al., 2014) and notable test development resources (c.f., Lane et al., 2016) clearly state that test development begins with specifying the test's purpose, and subsequent test development decisions must align with the purpose. However, a communication gap between test developers and end users often compromises the validity of test-score uses and interpretations (Chatterji, 2013). Use cases bring to this process the intentional systemic integration of the end users' voices. By meaningfully incorporating test-score users throughout the test design and development process, the resulting test scores are more likely to be usable and useful, thereby avoiding the unintended consequences resulting from underuse or misuse of test results.

In this manuscript, we describe how we applied HCD in an iterative process to establish the use case and subsequent test design decisions for a classroom-based assessment system focused on two early mathematics constructs. Before reporting on the methods and findings from the research activities we conducted with teachers as the end users of classroom-based assessments, we describe the importance of these early mathematics constructs in kindergarten through Grade 2 (K-2). Within this discussion, we introduce the concept of learning progressions, which forms the basis of instruction and assessment for these constructs.

Classroom-based assessment of early mathematics

Early mathematics skills are strong predictors of future mathematics (Watts et al., 2014) as well as science (Claessens & Engel, 2013), reading (Duncan et al., 2007), and future socioeconomic status (Ritchie & Bates, 2013). However, much of the curriculum, instruction, and assessment in the early grades focuses on reading (Clements & Sarama, 2016). This disconnect places teachers in a precarious position: knowing the importance of early mathematics, yet having few resources to provide students with high-quality learning experiences. Although teachers may be proficient in designing instructional resources, they often have little to no experience creating high-quality assessments (Bennett, 2011). Without trustworthy

data to guide instructional decisions, teachers may be underprepared to develop student proficiency in early grades mathematics concepts.

Two early mathematics constructs have emerged as being particularly important yet underrepresented in curriculum, instruction, and assessment: numeric relational reasoning (NRR) and spatial reasoning (SR). NRR is closely related to number sense and is defined as a child's ability to analyze relationships between numbers or expressions using knowledge of properties of operations, number decomposition, and known facts (Baroody et al., 2016; Carpenter et al., 2003; Farrington-Flint et al., 2007). NRR is predictive of mathematics achievement in the short and long term (Aunio & Niemivirta, 2010; Nunes et al., 2012) and supports the development of other mathematics concepts, such as algebraic reasoning. Demonstrating NRR is not akin to reproducing a sequential set of procedures to execute an algorithm, but instead requires reasoning through a "strategic" decision-making process based on the relationships between numbers (Whitacre et al., 2016). As an illustration, the child in the following scenario is demonstrating NRR: When presented with an addition problem of 6 + 5, the child notes that as 5 + 5 is 10 and 6 is 1 more than 5, then 6 + 5 must be 1 more than 5 + 5.

SR is a child's ability to interact with, navigate in, and understand their environment (National Research Council [NRC], 2001, 2009). SR is predictive of future science, technology, engineering, and mathematics (STEM) performance. In a meta-analysis examining the relationship between spatial skills and mathematics, Atit et al. (2021) confirmed previous findings documenting a direct positive relationship between SR and mathematics skills. Moreover, students with strong SR abilities are more likely than students with weaker SR skills to pursue STEM college degrees and careers (Uttal & Cohen, 2012; Wai et al., 2009). SR includes two primary components: spatial orientation and spatial visualization (NRC, 2009). Spatial orientation refers to the ability to picture objects or settings from different perspectives, such as mentally reorienting oneself to see something from above or below. Spatial visualization is used when a child imagines an object and transforms it mentally. Evidence suggests that SR skills support students' overall mathematics knowledge and specific concepts such as place value, relationships between numbers (including representations on the number line), and operations (Battista, 1990; Cheng & Mix, 2013; Newcombe & Frick, 2010; NRC, 2001).

To support students' development of early mathematics concepts, including NRR and SR, researchers have begun investigating the developmental sequence of knowledge and skills that underlie proficiency in these constructs. These are most commonly referred to as learning trajectories or learning progressions; however, because of the overlap in their meaning (Confrey, 2018), we use the term learning progressions. Grounded

in research on student learning, learning progressions are theories that describe the development of sophistication in students' thinking. They illustrate the progression of understanding from novice or naïve conceptualizations to more advanced or deeper understandings (Alonzo, 2018; Bennett, 2015; Corcoran et al., 2009; Pellegrino, 2014). Learning progressions are not intended to be deterministic; instead, they represent a hypothetical progression through which individual student's own pathways may vary. Learning progressions hold great utility for designing instruction that builds on students' prior knowledge, supports conceptual growth, addresses misconceptions, and integrates multiple ways of knowing. As research continues to expand in this area (c.f., Confrey, 2018; Sarama & Clements, 2019), learning progressions may be a valuable resource for teachers.

Although instructional resources to teach early mathematics concepts are emerging, the availability of classroom assessments are lagging behind. The Measures of Early Mathematics Reasoning Skills (MMaRS; NSF 1721100) was designed to fill this need. The initial solution framework focused on creating a universal screening assessment system to provide teachers with information about students' risk status for attaining proficiency in these constructs (c.f. Kettler et al., 2014, for a description of universal screening assessment practices). However, after initial discussions with involved stakeholders, this use case was questioned, and the utility of such an assessment was scrutinized. So began our quest to specify a use case and subsequently design classroom assessment resources that would better align with teachers' needs and preferences. We implemented HCD as the methodological approach so as to systematically integrate teachers' voices, while simultaneously help us understand the barriers and challenges to implementing and using tests in their classrooms. This process enabled us to design the MMaRS classroom assessment resources for maximum usability and utility.

Instantiated example of HCD in test development

For the remainder of this manuscript, we describe the research methods and results associated with applying HCD to design and develop the MMaRS classroom-based assessment system for NRR and SR. For each HCD phase, we present the research questions we investigated; the methods and procedures followed, including participants and data collection approach; and the findings. All participants provided their written informed consent to participate; each phase of the research was approved by the Institutional Review Board at Southern Methodist University and ethical standards for conducting research were followed. Our goal is not to provide a thorough treatment of each study; instead, we summarize how each phase contributed to the overall

test development process (for detailed findings for each phase, see Barton et al., 2019; McMurrer et al., 2020, 2021). The methodological approach associated with each phase is summarized in Figure 1.

Understand phase

The primary research question guiding this phase focused on understanding K-2 teachers' needs and wants from an assessment system to guide early mathematics instruction.

Participants

Eight K-2 teachers from five diverse school districts in the southern region of the United States provided written consent to participated in this research. As displayed in Table 1, all participants were white females with at least 3 years of teaching experience. Participants were nominated by school or district leaders to participate based on their interest in early mathematics education.

Procedures

We conducted two focus groups and a prioritization exercise. Focus groups—or facilitated semi-structured group discussions—were selected to address the research questions because they provide a forum for exploring a variety of perspectives based on a shared experience (Carey & Asbury, 2012). Rich data are generated when the facilitator encourages interaction among group members and discussion of varied experiences; consensus is not expected or needed.

Both focus groups were facilitated by an experienced member of the research team using standardized protocols (Krueger & Casey, 2009). Each lasted approximately 1.5 h and was conducted in-person at the research office. At least two other members of the research team attended and took field notes. Each meeting was structured to include an introductory activity to allow participants to get to know each other and the facilitator, followed

Table 1. Summary of participants from the understanding phase of the research.

Teacher No.	Gender	Years of Experience	Age	Race/Ethnicity
1	Female	3	DNS	White/European American
2	Female	6	30-39	White/European American
3	Female	10	30-39	White/European American
4	Female	18	40-49	White/European American
5	Female	20	40-49	White/European American
6	Female	20	50-59	White/European American
7	Female	20	50-59	White/European American
8	Female	20	50-59	White/European American

DNS: Did not state.

by an overview of the purpose of the meeting. Every effort was made to ensure that participants fully understood the expectations of their participation in the session and all ethical considerations for the use of inclusive, non-biased language and the social-emotional safety of each participant was accounted for throughout the planning and execution of each activity.

In the first focus group, teachers were prompted with questions about how they make instructional decisions and their experiences with assessments and data in reading and mathematics. The facilitator guided the discussion and encouraged all participants to engage, providing ample wait time and using open-ended questioning techniques to elicit as much participation as possible.

Six of the original eight teachers participated in a second follow-up focus group that was intended to seek confirmation (via member-checking) of our previous findings and allow participants to prioritize the instructional decisions. At the start of the meeting, participants received a set of green and red stickers (9 of each). Green stickers indicated high-priority decisions, and red stickers indicated low-priority decisions. Around the room were posters with instructional decisions that emerged from the first focus group. Teachers individually applied the stickers to the statements according to their level of priority. Following this activity, the trained facilitator led a discussion to understand the rationale for the teachers' placement of stickers.

Findings and conclusions

An important outcome of the Understand phase of the HCD process is developing a deep understanding of the end users' preferences and needs for a given product or system (Bowie & Cassim, 2016; Hanington, 2003). By seeking to understand, researchers develop empathy for the end users as well as gather insights that can lead to innovation and creative understanding that will inform the ideation and prototype phases of development.

During the first focus groups, teachers shared their experiences and opinions related to their use of data, score reports, and other tools for assessment and data collection. Ultimately, participants identified 27 fine-grained instructional decisions they want to make using data from classroom assessments.

Quantitative data from the second focus group (number of stickers for each statement) were summarized to identify high priority instructional decisions. Twelve of the 27 statements received two or fewer stickers of either color, indicating that a majority of the participants had neutral perceptions of these decisions. Eight statements received four or more high priority stickers, and six received four or more low priority stickers. One statement received two of each color sticker. Transcripts from the facilitated discussion were open coded (Creswell, 2013) by two members of

the research team using NVivo software to better understand teachers' rationale for their selections. These codes helped us refine and cluster the statements into two meaningful types of decisions that form the use case for the classroom-based assessment system for NRR and SR:

1. Determine how to form groups for learning stations or centers and intervention (9 high priority stickers; 0 low priority stickers)

    - Know students' current level of skills and knowledge (6 high priority stickers; 0 low priority stickers)
    - Understand gaps in students' knowledge based on errors or misconceptions (5 high priority stickers; 0 low priority stickers)

2. Determine the instructional sequencing that builds on students' prior learning (4 high priority stickers; 0 low priority stickers) and identifies subsequent learning targets that could form the content of the groups

    - Determine which content should be weighted more heavily in instruction (7 high priority stickers; 0 low priority stickers)
    - Know how to increase the sophistication of content (4 high priority stickers; 1 low priority stickers)
    - Understand which content needs to be spiraled (e.g., reviewed and practiced multiple times) (4 high priority stickers; 0 low priority stickers)

Instructional decisions that received the lowest priority based on the quantitative and qualitative analyses were: evaluating instructional strategies based on their effectiveness (0 high priority stickers; 12 low priority stickers), creating differentiated problem sets or number talks (2 high priority stickers; 13 low priority stickers), grouping students based on affective characteristics such as motivation or interest (0 high priority stickers; 8 low priority stickers), designing assessments (0 high priority stickers; 4 low priority stickers), and projecting growth toward end-of-the-year benchmarks (0 high priority stickers; 4 low priority stickers).

Define phase

In this phase, designers begin to define the context in which the new process, product, or experience will be used. Developing empathy for the end user is central to this phase as it may improve the desirability and utility of the end result (Fisher et al., 2021). In our research, this phase focused on defining the context and constraints teachers face when using assessment data to make instructional decisions. Our research question was: How do K-2 teachers perceive of assessments and the role of data in their practices?

Participants and procedures

Participants included six of the original eight teachers. Research activities included collecting teachers' affective expression of their personal experiences and conducting a consensus-building activity.

To collect teachers' affective expressions, we implemented an HCD strategy called Love Letter/Breakup Letter (Martin & Hanington, 2012) to explore the complicated and nuanced relationship that teachers have with assessment data. A "love letter" and a "breakup letter" are typically written expressions of the emotions one feels with regard to an intimate relationship with a person, whether at the beginning, middle, or end of the relationship. This approach helps designers "understand the less tangible aspects of the things they create; specifically, the social, human values and meanings conveyed through the things and experiences we design, as well as their understandability and usability" (Shedroff, 2003, p. 159).

At the start of the activity, teachers were prompted with a "Dear Assessment Data" letter format as both a Love Letter and a Breakup Letter. An example of the format was shared verbally with unrelated content, so as not to influence the participants. The example letters modeled what it sounded like to personify an inanimate object or system (in our example, the postal system and another well-known delivery system). These examples were developed to encourage creativity and foster a level of comfort with the type of writing being requested. Teachers were provided with approximately 10 min to write the letters and given the opportunity to share with the group when finished.

The consensus-building activity involved two HCD strategies: the KJ Technique and affinity diagraming. The KJ Technique is an approach to brainstorming and building consensus among large groups of people within a limited amount of time (Kawakita, 1982). This technique originated in Japan as a method employed during business meetings to work through difficult challenges in a large group, while providing every participant with an opportunity to participate. The method is traditionally completed silently until the participants are asked to discuss. Each participant has an opportunity to write their ideas on sticky notes, and all are encouraged to write as many ideas as possible, even building on one another's ideas and spurring further brainstorming. The next phase involves associating ideas into meaningful groups. Affinity diagraming—a collaborative method to organize ideas, themes, and priorities by "creating a visual representation of a team's observations, knowledge, concerns, and ideas" (Martin & Hanington, 2012, p. 104)—can be employed at this stage.

A member of the research team who was skilled in implementing HCD methods began the KJ Technique by prompting teachers with the question, "What makes data useful and useable to you?" Two additional members of the research team observed and took field notes, but did not engage in the activities. Teachers silently wrote ideas on sticky notes and

posted them on the wall. The facilitator guided the teachers through the process of grouping and regrouping ideas silently, then through the step of identifying themes and creating names or labels for each group, first individually, then collectively.

An affinity diagram was created to illustrate the associations between groups. Teachers were asked to share verbally their thinking related to the label for each cluster of ideas. Collectively, the teachers decided to combine groups, remove groups, change the labels, or leave them the same, depending on unanimous agreement with all participants. Once the labels were finalized for each group of ideas, participants voted for up to three themes that represented the highest priority for their practices. These themes were then ranked based on the highest number of votes and listed in order of importance to represent the group consensus.

Findings and conclusions

Two researchers used an open coding method (Creswell, 2013) followed by thematic analysis to analyze the artifacts from the HCD activities. Responses were organized into four aspects of instrument development: (a) design of assessments (e.g., purpose, test and item format, content blueprint), (b) test administration or implementation (e.g., administration window, duration of testing, standardization procedures), (c) presentation of test results (including score reports), and (d) the instructional utility or action based on the results.

The Love Letters and Breakup Letters were analyzed based on affective expressions related to the four aspects of instrument development. Expressions were interpreted as positive, negative, or mixed. Across both letters, teachers' references to "presentation of test results" were coded as expressing positive emotions 1.2 times as frequently as negative emotions. By contrast, teachers' references to "test administration or implementation" were coded as expressing negative emotions 4.25 times more frequently than positive emotions. These coding frequencies are aggregated and displayed in Table 2 and then disaggregated by source in Tables 3 and 4. In Table 3, for the Love Letters, the majority of comments focused

Table 2. Aggregated affect data from love and breakup letters.

	Total references coded		Affective references coded					
			+		Mixed		–	
Theme	No.	% of total	No.	%	No.	%	No.	%
Design of assessments	19	15	4	21	2	11	13	68
Test administration or implementation	43	33	8	19	1	2	34	79
Presentation of test results	47	36	21	45	8	17	18	38
Instructional utility or action	22	17	9	41	3	14	10	45

Table 3. Affect data from love letter.

	Total references coded		Affective references coded							
			Positive			Mixed			Negative	
Theme	No.	% of total	No.	%	Sample Quotes	No.	%	Sample Quotes	No.	%
Design of assessments	6	11	4	66	I can depend on you.	2	33	You say you've changed. Have you?	0	0
Test administration or implementation	9	17	8	89	I love that you show up 3 times each year, come rain or shine.	1	11	Dare I hope that you won't turn against me. Let's try one more time.	0	0
Presentation of test results	27	51	21	78	You are so great when you give me detailed information that shows how my students are thinking. I am then able to compare your previous visits with your real-time visit so I can see where my students stand. Also, it is lovely when I see patterns with groups of students and therefore can give a teaching point that is effective to a small group of students. I really needed you to remind me that Doris doesn't always remember how to subtract and she sometimes needs reminders to help get her on track.	6	22	Will kids be served by the information you hold? I love the detailed information some of you provide.	0	0
Instructional utility or action	11	21	9	82	It makes it easy for me to group my kids and also to know what skills I need to reteach.	2	18	Here you stand … promising to show me the path forward. How I wish I believed this prosaic story.	0	0

Table 4. Affect data from breakup letter.

	Total references coded		Affective references coded						
			Positive		Mixed		Negative		
Theme	No.	% of total	No.	%	No.	%	No.	%	Sample Quotes
Design of assessments	13	17	0	0	0	0	13	100	Beating down my students with you long windedness and strange vocabulary so it appears they do not know things they do know. Just leave me alone and let me ask what I want and need to know. You say one thing and do another.
Test administration or implementation	34	44	0	0	0	0	34	100	Why did you pick today to come by? Didn't you know that Mason didn't take his meds this morning? … What about Felicity? Did you not hear her fighting with Gabby at recess and then cry in the corner of the room? You are no good to me today—GO AWAY! Just when I get done with you, you show up again! I always want to break up with you when you drag on and on. I can see the frustration with my students and yet sometimes I have to keep pushing. Often, I feel like all of my time is devoted to assessing, and I never can get to the point where I can TEACH. I'm just so exhausted from testing 22 kids and it taking so long.
Presentation of test results	20	26	0	0	2	10	18	90	It is even more frustrating when the data obtained seemingly has nothing to do with what I teach or need to teach. Sick of you wasting my time and energy only to give me vague, unspecific information (Imagine Math) The emotional damage you bring to kids is unforgiveable. We are more than a test score! So many times your information is not accurate. My kids know so much more than your info shows. I'm done!
Instructional utility or action	11	14	0	0	1	9	10	91	Data should guide me, but so often it doesn't. I'm tired of pouring all of my time and energy into you so I have nothing left for what matters—actually teaching my students.

on "presentation of test results." In Table 4, for the Breakup Letters, the majority of the comments focused on "test implementation or administration."

The KJ Technique and affinity diagraming were designed to bring cohesion to the participants' perspectives. A wide range of ideas emerged during the initial stages of this activity. As the facilitator supported the participants' clustering and labeling of groups, three distinct groups were formed. First, teachers want data that align with the content expectations. Second, teachers want flexibility in test implementation that is focused on individualization. Third, and finally, teachers want data that are actionable and (1) support instructional decisions, (2) are communicated in ways that are clear and actionable, and (3) provide meaningful information for parents.

Analyzing the data across research activities led to four insights that helped us define the context and constraints associated with teachers' use of classroom assessment data. First, teachers in our sample value data that are directly linked to their students' learning and areas of growth. They want fine-grained information that pinpoints students' skills and gaps in their understanding. They want data that directly connects to instructional practice. Teachers also noted that they feel greater ownership when they have data that are actionable and interpretable.

Second, teachers expressed frustration with assessments that were mandated by external agencies (e.g., local or state education agency) and were perceived as having limited value. Value was diminished when the content of the assessments did not reflect their instructional goals and when the data did not inform their instruction. Moreover, teachers expressed a lack of agency when mandated assessments took considerable time away from their instructional priorities. Relatedly, the third insight points to teachers' appreciation for district policies that allow for flexibility in classroom assessment practices. Teachers expressed a sense of empowerment when district leaders allowed them to create their own formative assessments or contribute multiple sources of data.

Fourth, participants emphasized the importance of conveying assessment results in a way that is clear and actionable. Across data sources, participants made multiple references to data that were confusing and reports that were not easy to communicate with students or parent. Teachers expressed a desire for assessment reports or results that translate into action effectively and conveniently in the classroom setting.

As a primary outcome of this phase, we decided to operationalize the constructs as learning progressions to more closely align with the development of students' understanding in the domains. This shift was made to account for teachers' request for data that accurately reflect the

fine-grained processes of learning and allow them to take strategic action. Reporting data associated with the learning process may also facilitate communication with students and parents.

Prototype phase 1: exploring the learning progressions

The next phase of the HCD process is to develop prototypes based on the end users' needs and preferences, as identified in the Understand and Define Phases. The primary research questions guiding this phase were: (1) What are teachers' perceptions of the usefulness of learning progressions-based classroom assessments? (2) What test specifications support teachers' use and engagement with the assessments?

Participants

Participants were recruited from the same school districts as the previous research activities and were nominated by their school or district leaders. Ten female teachers with direct experience teaching K-2 were recruited to participate. As displayed in Table 5, teachers were roughly evenly distributed across grades and had 4–20 years of experience.

Procedures

Two research activities were conducted to address the research questions: focus groups and design charettes. Prior to participating in these activities, participants completed a 1-h online module that we created to better understand learning progressions. The module included a basic definition of learning progressions, an overview of how learning progressions align with the learning process, and how assessments can be designed based on learning progressions. All research activities were virtual and synchronous, using Zoom video conferencing due to school closures caused by the COVID-19 pandemic.

Table 5. Summary of participants from the prototyping phases of the research.

Teacher No.	Gender	Years of Experience	Age	Race/Ethnicity	Grade Level	District Size & Locale
1	Female	4	20–29	White/European American	K	Large City
2	Female	14	30–39	White/European American	K	Large Suburb
3	Female	7	30–39	Black/African American	K	Large City
4	Female	20	50–59	White/European American	1st	Large City
5	Female	12	40–49	White/European American	1st	Large Suburb
6	Female	6	20–29	White/European American	1st	Rural Fringe
7	Female	10	30–39	White/European American	2nd	Large City
8	Female	4	20–29	White/European American	2nd	Large City
9	Female	13	30–39	Hispanic/Lantio American	2nd	Large Suburb
10	Female	20	40–49	White/European American	2nd	Large Suburb

The focus group was designed to solicit teachers' perceptions about the usefulness of learning progressions-based classroom assessments and lasted approximately 1 h. Questions drew out participants' reactions to learning progressions and how they could use them in their practice. To allow participants more opportunities to share their perspectives during the focus group, the participants were divided into two groups of five. Each group was facilitated by a member of the research team and two additional researchers attended to take field notes.

To address the second research questions, we implemented design charettes, an HCD strategy adapted from the importance–difficulty matrix technique outlined by Hanington and Martin (2019), which are "a workshop-style technique that provides a collaborative space that allows for [the] creation and cross-pollination of design ideas to occur. Designers and non-designers—including project stakeholders, engineers, and users—can participate in a design charette" (Martin & Hanington, 2012, p. 58). Participants move between groups at systematic intervals so ideas can be shared across a range of participants.

To implement the design charettes, we held a 1.5-h meeting that was facilitated by three members of the research team. Participants were arranged into three groups of three or four. An interactive whiteboard space (e.g., Miro.com) was introduced, and participants were instructed to draw or sketch, use icons, and use virtual sticky notes to collaboratively design their ideal assessment. After 20 min, two members of the original group were moved to a new group. This rotation repeated twice for a total of three sessions of 20 min each.

The group reconvened and a lead facilitator from each group shared their final design prototypes. Individual participants reflections were solicited.

Findings and conclusions

The focus groups elicited teachers' perceptions of the usefulness of learning progressions-based assessments. Following an open coding method (Creswell, 2013) and substantive categorization process (Corbin & Strauss, 2015; Maxwell, 2005) alongside visual, sense-making tools such as the concept map (Miles & Huberman, 1994; Strauss, 1987), the following themes emerged:

- Road map: Teachers referred to learning progressions as a road map that identifies the direction of children's learning. They also recognized that it can be adaptable for individuals, because not everyone will follow a linear pathway. Instructional decisions they referenced included how to sequence content with the end in mind.
- Signpost: Teachers noted that learning progressions could indicate where children are at in their learning. Instructional decisions include

planning their instruction to identify the "next steps" in the learning and designing practice activities to build on students' current level of knowledge and skills.
- Gap detector: Teachers noted that learning progressions could help them identify gaps in students' learning or misconceptions. Instructional decisions include planning backwards to fill the gaps, such as going to previous topics and correcting misunderstandings.
- Inform groupings: Teachers could use learning progressions for any of the above-mentioned purposes in a whole group or small groups.
- Communicate with parents: Because of the intuitive nature of learning progressions, teachers could communicate in a manner accessible to parents the message of where a child is in the learning process and where they are going.

These themes aligned with the needs and preferences identified in the Understand Phase and the context and constraints that emerged in the Define Phase. We concluded that the learning progressions may be a viable solution.

Design charettes were conducted to identify the test specifications that would support teachers' use and engagement with classroom assessments. Design elements emerged from this prototyping session included:

Item design
- Assess using hands-on materials that do not feel like an assessment to the children (e.g., game-like).
- Items could include a mix of observations and performance tasks.
- Materials should be accessible so all students have equitable opportunities.
- Items should allow teachers to "see" students' thinking.

Administration
- Facilitated by technology to aid in administration and scoring and to reduce the paperwork burden
- One-to-one administration or in a small group and should require little preparation
- Administration should take 5–7 minutes per testing session and can be extended over multiple small chunks
- Starting place can vary based on differences in students' prior knowledge

Prototype phase 2: exploring test design

Using the results from the Prototype Phase 1, we developed a prototype of the MMaRS assessment interface for both teachers and students. We

sought to address the research question: Which aspects of the initial prototype are useful and usable, and which aspects are not? For those aspects that were not useful or usable, we conducted a rapid prototyping session to improve the design.

Participants and procedures

Six of the 10 teachers who participated in the prototyping Phase 1 sessions also participated in Phase 2. We conducted a focus group via Zoom that lasted about 2 h. A member of the research team facilitated the focus group, and two additional members took field notes. We shared a low-fidelity prototype sequential process (e.g., one page at a time). Participants received a list of words and were instructed to select three to five that best described their immediate reactions to the prototype. The word list was adapted from Benedek and Miner (2002) and Rohrer (2008). The trained facilitator led a discussion about their responses.

Following the discussion, participants were divided into two groups of three and engaged in a participatory co-design activity to redesign aspects of the prototype that participants labeled as being problematic. Co-design involves eliciting and valuing participants' voices by actively engaging end users in the design process (Martin & Hanington, 2012; McKercher, 2020). Because of the virtual environment, participants printed the materials and redesigned the prototype. After working independently, the research team facilitated a collaborative discussion using a virtual whiteboard to allow each participant to share their design.

Findings and conclusions

First independently and then in a collaborative discussion, two researchers coded the data for organizational categories and possible substantive categories (Maxwell, 2005). The substantive categorizations included:

- Information architecture: The positive words most often used by teachers to describe the prototype included "organized," "clear," and "accessible." The most often used negative words were "confusing" and "complex." Teachers emphasized the need to use concise and understandable language that is familiar to teachers.
- Design elements: Teachers overwhelmingly preferred information presented in lists and in a vertical flowchart format. Teachers reacted positively to many of the design features, including the use of circles and bolding for important terms alongside the squared text boxes with columns of narrative. Teachers were mixed in their perception of the length of the prototype (some thought it was too long, and some thought it was just right).

- Level of detail: A majority of the teachers perceived the prototype as overwhelming, wordy, and too technical. Teachers wanted sufficient detail to make decisions but noted that the text needs to be comprehensible.
- Pathways for decision-making: Teachers noted the usefulness of the flowcharts and the simplicity of the decision trees. Most teachers also liked seeing examples of student thinking in the flowcharts to help interpret the content; however, one teacher was confused as to the purpose of the examples.

The positive aspects of the initial prototype were retained in the revised prototype. The recommendations were discussed and incorporated into the co-designed prototype. Together, the participants created the next iteration of the prototype. In Appendix A, we display an annotated example page of a teacher interface prototype with excerpts of teacher feedback listed for reference and the participants' co-design recommendations. The purpose of the resource prototype is to help teachers select the constructs they would like to assess within the learning progression for their students.

Discussion

When designing a process, product, or experience, Johnson et al. (2005) points out that a mismatch between the end users' perceptions and those of the designers can lead to complications from both perspectives: The end users may be less likely to get what they want and need, and developers may spend additional time and resources trying to address problems later in the development (e.g., providing technical support, redesign). When designing classroom-based assessments that are intended to support teachers' instructional decisions, the wants and needs of teachers should drive many of the design decisions. The process of involving stakeholders in the design of assessments is not new (Ryan, 2002). As the primary user of classroom-based assessment results, teachers are key stakeholders whose input should be considered throughout the design and development phases; however, teachers' voices are often not systematically incorporated. What may follow are assessments that are not useful or usable as tools to support teachers' instructional decisions. Consequently, teachers may disengage from the assessment process (e.g., not administer the assessment or administer the assessment but not use the results) or use the results in unintended ways. These outcomes compromise the validity of the uses and interpretations of the test results.

In this manuscript, we described a novel application of HCD to the design of classroom-based assessments. Although various methods exist to incorporate stakeholders' perspectives in test development (e.g., focus

groups, interviews), we opted to apply HCD because it offered a thorough and systematic approach. HCD concurrently emphasizes developing empathy for the end users and co-designing plausible solutions. By thoroughly understanding teachers' wants and needs from an empathetic perspective and using this understanding in the co-design process, we were better able to identify design elements that improved the usefulness and usability of the final assessment system.

In the Understand Phase, the guiding research question focused on understanding teachers' needs and wants from an early mathematics assessment system. Focus group discussions provided deeper insights into the teachers' current assessment practices and experiences. Teachers identified multiple decisions they want to make to support their mathematics instruction; many of which they do not currently have data. Narrowing in, they prioritized two key decisions that would most impact their instruction, and thus students' opportunities to learn NRR and SR. These activities supported the specification of the use case for the MMaRS assessment system.

Data collected in the Define Phase addressed the research question of how teachers perceive of assessments and the role data play in their current practices. Our analyses illustrated the disengagement teachers experience when assessment systems do not meet their needs. The Love Letters and Breakup Letters illuminated teachers' perceptions about multiple aspects of test design, administration, and use. Most frequently, teachers expressed negative emotions associated with test administration. They were frustrated by mandated assessments that they perceived as having little value in their instruction. Conversely, positive emotions were expressed when discussing the presentation of test results. Teachers sought out results that would provide actionable information to guide their instruction. With these data, we defined the context and constraints teachers face when using classroom assessment data and used this information to build the initial prototypes of the MMaRS assessment.

During the Prototype Phase, we used HCD strategies to iteratively design the testing interface to maximize usability. Our research questions primarily focused on identifying, understanding, and ameliorating teachers' perceptions of the usefulness and usability of the assessments. We engaged users in a two-phase prototyping exercise. Through this process, we created a shared vision for and understanding of how teachers—as the end users—will engage with the MMaRS assessment system (Cockburn, 2000). As a result of improving the communication between the end users and the test developers, HCD facilitated a closer alignment between the intended and enacted uses and interpretations of the test results. Once the MMaRS assessment system is operational, additional data about the actual classroom-based uses can be collected to verify this alignment.

Implications for future research and practice

Further research is needed on uses of and approaches for integrating stakeholders' perceptions into the design and development of assessments. First, given the context of the MMaRS assessment, which focuses on two early mathematics constructs that are not mandated to teach within the school curriculum, it was necessary for us to engage with stakeholders to articulate the use case. However, in content areas with high accountability expectations, teachers may have less choice about the uses of assessments. Additional research is needed to determine how to meaningfully integrate stakeholders' voices in these settings.

Second, given the importance of stakeholders' perceptions on the validity of the uses and interpretations of assessment results, additional research is needed to examine how to integrate stakeholders' voices into existing assessment systems. For example, many international assessments (e.g., Trends in Mathematics and Science Survey; TIMSS) incorporate context questionnaires as part of test administration. Teachers respond to questions about a variety of topics including their school's culture and environment, perceptions and attitudes toward teaching, professional learning experiences, and instructional practices. For instance, on the TIMSS questionnaire, teachers are asked to rate their perception of the importance of specific assessment practices in their mathematics and/or science classes. They are also asked to state whether or not they have received professional development on mathematics and/or science assessment, and if they need professional development on mathematics and/or science assessment. Test developers could use these data to determine teachers' current assessment practices and make inferences about their need for professional learning experiences. Although these data may not directly contribute to future iterations of the assessment, they could meaningfully inform how policy makers engage with teachers about the role of assessment in their practice. Future research could explore how findings from these types of questionnaires can proactively inform iterative test development decisions.

Third and finally, the research described in this manuscript illustrates the use of specific HCD approaches. Further research is needed to explore the applicability of other HCD approaches and/or the integration of other qualitative and quantitative research designs. Expanding on the range of approaches available to test developers seeking to integrate stakeholders' voices may support future test development efforts.

Limitations

Several key limitations of the research reported in this manuscript impact the generalizability of the findings. First, the sample sizes used in each

phase were small; however, they were appropriate for the HCD activities used to collaborate with the participants. Although care was taken to select participants from various contexts (e.g., school type, geographic locale, diversity of student population), all participants were recruited from one southern state in the United States. As such, all participants experienced similar education policies and practices associated with testing. This similarity in context may impact the transferability of the findings to other regions in the United States or elsewhere. Second, the collaborative nature of some research activities (e.g., focus groups, KJ Technique, design charettes) may have led participants to respond in socially desirable or expected ways, thereby causing a false sense of consensus. Because the outcomes from the research activities impacted the design of the MMaRS assessments, these limitations may impact the usefulness and usability of these assessments in other contexts.

Conclusions

This manuscript presented an approach for meaningfully incorporating stakeholders' perspectives into the design and development of a classroom-based assessment system. HCD offers a systematic way of integrating users' voices into the design of products, services, and systems, but is not often applied in educational contexts. We instantiated the use of HCD methods in the design and development of classroom-based assessments for each mathematics construct. By focusing on the end users' needs and wants in an assessment system, our goal is to improve the utility and usability of the test, thereby supporting valid interpretations and uses of the results.

Acknowledgments

The authors would like to thank the Research in Mathematics Education team for their contributions in planning and facilitating this research as well as their assistance with the analyses. Finally, we extend our gratitude to the teachers who partnered with us on this journey.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This material is based upon work supported by the National Science Foundation under Grant No. 1721100. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Alonzo, A. C. (2018). An argument for formative assessment with science learning progressions. *Applied Measurement in Education*, *31*(2), 104–112. https://doi.org/10.1080/08957347.2017.1408630

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

Atit, K., Power, J. R., Pigott, T., Lee, J., Geer, E. A., Uttal, D. H., Ganley, C. M., & Sorby, S. A. (2021). Examining the relations between spatial skills and mathematical performance: A meta-analysis. *Psychonomic Bulletin & Review*, *29*(3), 699–720. https://doi.org/10.3758/s13423-021-02012-w

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, *20*(5), 427–435. https://doi.org/10.1016/j.lindif.2010.06.003

Baroody, A. J., Purpura, D. J., Eiland, M. D., Reid, E. E., & Paliwal, V. (2016). Does fostering reasoning strategies for relatively difficult basic combinations promote transfer by K-3 students? *Journal of Educational Psychology*, *108*(4), 576–591. https://doi.org/10.1037/edu0000067

Barton, T., Knight, K., Hatfield, C., Perry, L., & Ketterlin-Geller, L. R. (2019). *Teacher advisory panel technical report: Fall 2018 – summer 2019* (Technical Report No. 19-25) [Technical Report]. Southern Methodist University, Research in Mathematics Education.

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education*, *21*(1), 47–60. https://doi.org/10.2307/749456

Benedek, J., & Miner, T. (2002, July 8–12). Measuring desirability: New methods for evaluating desirability in a usability lab setting. In *Proceedings of UPA 2002 Conference*. https://nanopdf.com/download/measuring-desirability-new-methods-for-evaluating_pdf

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Bowie, A., & Cassim, F. (2016). Linking classroom and community: A theoretical alignment of service learning and a human-centered design methodology in contemporary communication design education. *Education as Change*, *20*(1), 126–148. https://doi.org/10.17159/19479417/2016/556

Carey, M. A., & Asbury, J. (2012). *Focus group research*. Taylor & Francis.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school*. Heinemann.

Chatterji, M. (2013). Bad tests or bad test use: A case of SAT use to examine why we need stakeholder conversations on validity. *Teachers College Record: The Voice of Scholarship in Education*, *115*(9), 1–10. https://doi.org/10.1177/016146811311500901

Cheng, Y.-L., & Mix, K. S. (2013). Spatial training improves children's mathematics ability. *Journal of Cognition and Development*, *15*(1), 2–11. https://doi.org/10.1080/15248372.2012.725186

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record: The Voice of Scholarship in Education*, *115*(6), 1–29. https://doi.org/10.1177/016146811311500603

Clements, D., & Sarama, J. (2016). Math, science, and technology in the early grades. *The Future of Children*, *26*(2), 75–94. http://www.jstor.org/stable/43940582 https://doi.org/10.1353/foc.2016.0013

Cockburn, A. (2000). *Writing effective use cases*. Addison-Wesley Professional.

Confrey, J. (2018). *Future of education and skills 2030: Curriculum analysis – a synthesis of research on learning trajectories/progressions in mathematics* (EDU/EDPC[2018]44/ANN3). Organisation for Economic Co-operation and Development.

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.) SAGE.

Corcoran, T., Mogat, F. A., & Rosher, A. (2009). *Learning progressions in science: An evidence-based approach to reform (CPRE Research Report #RR-63)*. Consortium for Policy Research in Education.

Creswell, J. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). SAGE.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Farrington-Flint, L., Canobi, K. H., Wood, C., & Faulkner, D. (2007). The role of relational reasoning in children's addition concepts. *British Journal of Developmental Psychology*, *25*(2), 227–246. https://doi.org/10.1348/026151006X108406

Fisher, W., Oon, E. P.-T., & Benson, S. (2021). Rethinking educational assessment from the perspective of design thinking. *EDeR. Educational Design Research*, *5*(1), 1–33. https://doi.org/10.15460/eder.5.1.1537

Giacomin, J. (2014). What is human centered design? *The Design Journal*, *17*(4), 606–623. https://doi.org/10.2752/175630614X14056185480186

Gulikers, J., Biemans, H., & Mulder, M. (2009). Developer, teacher, student and employer evaluations of competency-based assessment quality. *Studies in Educational Evaluation*, *35*(2–3), 110–119. https://doi.org/10.1016/j.stueduc.2009.05.002

Hanington, B. (2003). Methods in the making: A perspective on the state of human research in design. *Design Issues*, *19*(4), 9–18. https://doi.org/10.1162/074793603322545019

Hanington, B., & Martin, B. (2019). *Universal methods of design expanded and revised: 125 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Publishers.

Hasso Plattner Institute of Design at Stanford. (2010). *An introduction to design thinking: Process guide*. Stanford University. https://web.stanford.edu/~mshanks/MichaelShanks/files/509554.pdf

International Test Commission. (2012). *International guidelines on quality control in scoring, test analysis, and reporting of test scores*. www.intestcom.org

Johnson, C. M., Johnson, T. R., & Zhang, J. (2005). A user-centered framework for redesigning health care interfaces. *Journal of Biomedical Informatics*, *38*(1), 75–87. https://doi.org/10.1016/j.jbi.2004.11.005

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kawakita, J. (1982). *The original KJ method*. Kawakita Research Institute.

Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (2014). An introduction to universal screening in educational settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings* (pp. 3–16). American Psychological Association.

Krueger, R. A., & Casey, M. A. (2009). *Focus groups: A practical guide for applied research* (4th ed.). SAGE.

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2016). Test development process. In S Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–18). Routledge/Taylor & Francis Group.

Martin, B., & Hanington, B. (2012). *Universal methods of design: 100 ways to explore complex problems, develop innovative strategies, and deliver effective design solutions*. Quarto Publishing Group USA.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). SAGE.

McDonald, J. A., Merkley, R., Mickle, J., Collimore, L.-M., Hawes, Z., & Ansari, D. (2021). Exploring the implementation of early math assessments in Kindergarten classrooms: A research-practice collaboration. *Mind, Brain, and Education*, *15*(4), 311–321. https://doi.org/10.1111/mbe.12293

McKercher, K. A. (2020). *Beyond sticky notes. Doing co-design for real: Mindsets, methods and movements*. Beyond Sticky Notes.

McMurrer, J., Barton, T., Hatfield, C., & Ketterlin-Geller, L. R. (2021). *MMaRS Teacher Advisory Panel: Teacher resource development* (Tech. Rep. No. 21-06). Southern Methodist University, Research in Mathematics Education.

McMurrer, J., Mota, A., Pinilla, R., Hatfield, C., & Ketterlin-Geller, L. R. (2020). *Teacher advisory panel: Summer 2020* (Technical Report No. 20-22) [Technical Report]. Southern Methodist University, Research in Mathematics Education.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). SAGE.

National Research Council (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. National Academies Press. https://doi.org/10.17226/12519

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. National Academies Press. https://doi.org/10.17226/9822

Newcombe, N. S., & Frick, A. (2010). Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education*, *4*(3), 102–111. https://doi.org/10.1111/j.1751-228X.2010.01089.x

Nielsen, J. (1993). *Usability engineering*. Academic Press. https://doi.org/10.1016/B978-0-08-052029-2.50005-X

Nunes, T., Bryant, P., Barros, R., & Sylva, K. (2012). The relative importance of two different mathematical abilities to mathematics achievement. *The British Journal of Educational Psychology*, *82*(Pt 1), 136–156. https://doi.org/10.1111/j.2044-8279.2011.02033.x

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, *20*(2), 65–77. https://doi.org/10.1016/j.pse.2014.11.002

Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science educa-

tion. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 201–216. https://doi.org/10.1177/0002716219843249

Penuel, W. R., Confrey, J., Maloney, A., & Rupp, A. A. (2014). Design decisions in developing learning trajectories-based assessments in mathematics: A case study. *Journal of the Learning Sciences*, *23*(1), 47–95. https://doi.org/10.1080/10508406.2013.866118

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301–1308. https://doi.org/10.1177/0956797612466268

Rohrer, C. P. (2008, October 28). Desirability studies: Measuring aesthetic response to visual designs. *XD Strategy*. https://www.xdstrategy.com/desirability-studies/

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice*, *21*(1), 7–15. https://doi.org/10.1111/j.1745-3992.2002.tb00080.x

Sarama, J., & Clements, D. H. (2019). Learning trajectories in early mathematics education. In D. Siemon (Ed.), *Researching and using progressions (trajectories) in mathematics education* (pp. 32–55). Brill.

Shedroff, N. (2003). Research methods for designing effective experiences. In B. Laurel (Ed.), *Design research: Methods and perspectives* (pp. 156–159). Massachusetts Institute of Technology.

Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge University Press.

Uttal, D. H., & Cohen, C. A. (2012). Spatial thinking and STEM education: When, why, and how? In B. Ross (Ed.), *Psychology of learning and motivation* (pp. 147–181). Academic Press.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over fifty years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817–835. https://doi.org/10.1037/a0016127

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher (Washington, D.C.: 1972)*, *43*(7), 352–360. https://doi.org/10.3102/0013189X14553660

Whitacre, I., Schoen, R. C., Champagne, Z., & Goddard, A. (2016). Relational thinking: What's the difference? *Teaching Children Mathematics*, *23*(5), 302–308. https://doi.org/10.5951/teacchilmath.23.5.0302