Paper ID #25044

Validating a Measure of Problem Framing Ability to Support Evidence-Based Teaching Practice

Lindsey D White, University of New Mexico

Lindsey is a PhD student in the Organization, Information, & Learning Science program at the University of New Mexico. Her research interests include instructional design, technology, and adult learning as they relate to clinical educators and clinical education in healthcare. Lindsey, a Nebraska native, completed her Bachelor of Science in Education at the University of Nebraska-Omaha. She earned her Master of Science in Exercise Science from the University of Georgia. Prior to beginning her PhD, she worked for almost 7 years at Stanford University as a Certified Athletic Trainer.

Dr. Vanessa Svihla, University of New Mexico

Dr. Vanessa Svihla is a learning scientist and associate professor at the University of New Mexico in the Organization, Information & Learning Sciences program and in the Chemical & Biological Engineering Department. She served as Co-PI on an NSF RET Grant and a USDA NIFA grant, and is currently co-PI on three NSF-funded projects in engineering and computer science education, including a Revolutionizing Engineering Departments project. She was selected as a National Academy of Education / Spencer Postdoctoral Fellow and a 2018 NSF CAREER awardee in engineering education research. Dr. Svihla studies learning in authentic, real world conditions; this includes a two-strand research program focused on (1) authentic assessment, often aided by interactive technology, and (2) design learning, in which she studies engineers designing devices, scientists designing investigations, teachers designing learning experiences and students designing to learn.

Dr. Yan Chen, University of New Mexico

Yan Chen is a Postdoctoral Fellow in the Departments of Organization, Information & Learning Sciences and Chemical & Biological Engineering at the University of New Mexico. Her research interests focus on computer supported collaborative learning, learning sciences, online learning and teaching, and educational equity for multicultural/multiethnic education.

Mr. Todd Hynson, University of New Mexico

With 16 years of experience working in the registrar/student services environment, Todd Hynson, who initially began in front-line customer service, now serves as the registrar for the University of New Mexico Health Sciences Center. He currently is pursing a PhD in the OILS program at the University of New Mexico.

Ian A Drackert, University of New Mexico

I am currently a Sr. Academic Advisor for Liberal Arts and Integrative Studies at the University of New Mexico. I am also an OILS graduate student working with a team designing an instructional training program to help increase the consistency of Design Skills Test coders for better data analysis as it pertains to FACETS research.

Austin C. Megli, University of New Mexico

Austin C. Megli is an Organization, Information, and Learning Sciences PhD Student at the University of New Mexico. He also holds a Juris Doctor and Masters of Business Administration from the University of New Mexico. His research interests are in distance learning and interaction analysis.

Jordan Orion James, University of New Mexico

Jordan O. James is a Native American Ph.D. student in the Organization, Information, and Learning Sciences (OILS) program as well as a lecturer at the University of New Mexico's School of Architecture and

Planning in the Community & Regional Planning program. He has served as a graduate research assistant on an NSF-funded project, Revolutionizing Engineering Departments, and has been recognized as a Graduate Studies student spotlight recipient and teaching scholar. Jordan studies learning in authentic, real-world conditions utilizing Design-Based Research methodologies to investigate design learning and social engineering, in which he studies urban planners who design real-world interventions for communities and students who use design to learn. A member of the Grand Portage Band of the Lake Superior Chippewa Jordan obtained both his Masters of Community & Regional Planning and Bachelor of Media Arts from the University of New Mexico in Albuquerque where he lives with his wife and three daughters.

Claire Yvonne Saul

Validating a Measure of Problem Framing Ability to Support Evidence-Based Teaching Practice

Purpose. In this evidence-based teaching practice paper, we report on development, implementation, and validation of the Design Skills Test (DST), an assessment of design problem framing ability. **Methodology.** The DST includes an authentic design scenario and a coding scheme to characterize 1) *factual and conceptual information* used to frame the problem in terms of needs/constraints; 2) *design practices* used (e.g., generating ideas, considering multiple stakeholders, remaining tentative); and 3) *stylistic choices* (e.g., organizing their response, depicting context). We developed three DST scenarios and tested them in a chemical engineering program over a three-year period (n=580). To make data analysis feasible, two undergraduate peer-learning facilitators analyzed each DST independently (14 PLFs contributed), following minimal training. **Results.** Using a validity-as-argument approach (Linn, 1994), we argue that the DST provides valid information about design problem-framing ability, provided the information is used for course improvement purposes. Inter-rater reliability for *factual/conceptual* codes was 65% to 83%; for *practice* codes 52% to 77%; and for *stylistic* codes 68% to 80%). **Conclusions.** Our findings indicate that the DST sheds light on students' design problem framing ability and provides valid evidence to help faculty evaluate the impact of incorporating design challenges, as not all design challenges support students to learn how to design. Given that professional engineering design practice relies on knowing how to frame problems, it is important for students to have opportunities to develop problem framing ability. **Implications.** While reliability with minimal training was lower than would be acceptable for research purposes, for instructional purposes, this represents a significant reduction of faculty time. To enhance reliability, we worked with instructional designers to develop an online, self-paced training.

Introduction and research purpose

The idea of using evidence to inform instruction undergirds faculty development and departmental change initiatives, many of which include threading team design challenges through core courses. While there are assessments that measure conceptual understanding and surveys that measure perceptions (e.g., design beliefs, engineering identity, design self-efficacy, team skills, etc.), these provide an incomplete understanding of student individual progress on design problem framing ability. Students typically get a lot of practice solving problems, but comparatively little practice framing problems. Yet, the ability to frame design problems appears to be one of the most important predictors of creative design outcomes [1-4], making it an important skill to develop. We present initial efforts to develop, implement, and validate a measure of design problem framing ability that can be feasibly used to inform instruction.

Background

We ground our effort to design an assessment of problem framing ability in research on problem framing, performance-based assessment, and reliability and validity as arguments.

What is design problem framing ability?

Design problems are typified as ill-structured and ill-defined, meaning the design problem does not contain all the information required to solve it at the outset and that there may be many solution paths and satisficing solutions [5-8]. As a result, design involves framing the problem in order to solve it. To characterize design problem framing, we consider research on how experts carry out this process, which, broadly, involves three dimensions: (1) factual and conceptual information, both already known and gathered in the process, (2) design practices, (3) and design judgments and style.

In the first dimension, experienced designers assess what they know and do not know and gather factual and conceptual information about the problem [9-11]. Factual and conceptual information typically includes needs, design requirements, constraints, insufficiency of existing solutions [2, 4, 10, 12]. With more experience, designers appear to spend more time on problem framing [13] and to consider broad contextual issues as they frame problems [14]. They gather more and more varied information to understand the problem [12]. They seek divergent stakeholder perspectives, analyze design requirements and constraints, research shortcomings of existing solutions, and identify resources available to them [15].

The second dimension includes multiple design practices, including knowing ways to gather needed information, bounding the problem, considering the problem and needs from multiple points of view, generating initial ideas, staying tentative about and evaluating possible solutions in light of needs [9-11]. Experienced designers also navigate the ambiguity and uncertainty involved in problem framing. They deliberately resolve ambiguity and rule out untenable solution paths by gathering information [16]. They remain tentative in this process, exploring alternative solutions and conjecturing about the impacts of these solutions [17].

The third dimension encompasses individual and disciplinary styles as designers make judgments about the problem frame. Designers make many decisions as the frame problems, and because they fill in gaps in knowledge abductively (rather than deductively or inductively), these decisions are subject to disciplinary norms and individual styles [18-22].

Why use performance-based assessment to measure design problem framing ability?

While it is clear that students need multiple opportunities to develop problem framing ability [23], because of the complexity of problem framing as an engineering practice, determining whether educational experiences foster problem framing ability remains inadequately addressed in the research literature. Past efforts to assess design skills broadly and problem framing specifically have included direct and indirect measures. Indirect measures include self-evaluations, surveys and peer evaluation [24]. In the case of a complex skillset like problem framing ability, multiple-choice and constructed response exams fall into this category because they provide proxy data about the skill, but do not directly measure problem framing ability. Indirect measures have the advantage of being easy to analyze, often even by someone without expertise provided they have a scoring guide, rubric, or answer key. For instance, Osgood and Johnston [25] developed a measure of design ability, which they operationalized as problem framing, evaluating alternatives, and communicating their design ideas. Their measure included three scenario-based multiple-choice items and five Likert-scale questions related to problem framing. The multiple-choice scenario involved designing a chair for someone over six feet tall and posed questions such as "You just finished the first meeting with the client to discuss the problem, which lasted 15 minutes. Of the following, the first task you should complete is: (a) Develop a schedule of all tasks to be completed. (b) Find out more about chair design and background information. (c) Brainstorm ideas based on what the client said was important. (d) Write requirements to define the problem." An example Likert item is "Research is not necessary to develop product requirements." However, because of the ill-structured nature of design, and because students are not always effective judges of what they do and do not know, especially related to complex and ill-structured tasks [26], such assessments are not typically a valid means to measure problem framing ability. They are unlikely to effectively predict actual problem framing behavior.

In contrast, performance-based assessments (PBA) are direct measures, and can include assessment of actual behaviors (process) or the results of behaviors (products) on a realistic or authentic task. PBAs typically provide a better prediction of actual problem framing behavior. While an instructor may make and act on many formative assessments of student behaviors, it is challenging to observe all teams for the duration of their design process; likewise, video- or audio-recorded design team meetings are laborious for instructors to analyze, and require expertise to make sense of. Evaluating the products for evidence of how students used the three dimensions is a more feasible approach that may still shed light on students framed a problem. For instance, Shah [27] created a design brief to assess problem framing ability: "A new activity set for children (1 to 4 years old) is desired, to be produced from easily cleanable and durable materials. It should provide for many imaginative activities. It should be expandable for use by groups of children. It should be easily erectable and transportable. Cost should not exceed $40 for the base set."

PBA—whether focused on process or product—typically include standardized yet realistic tasks evaluated using a set of criteria [28]. This necessitates training and calibration for those scoring PBA [29]. For PBA to serve as a direct measure of skills like problem framing ability, it is essential that the tasks be authentic [30], but finding the balance between authenticity and feasibility continues to be a tension.

How should the reliability and validity of performance-based assessments be established?
Rather than being a fundamental property of the measure, validity is fundamentally tied to how an assessment is used [31]. To evaluate the validity of our measure of problem framing ability for making instructional decisions, we consider validity and reliability as *arguments* [32-34]. Reliability framed as an argument allows us to "expand the sources of evidence available for demonstrating the social and scientific value of reliability" [4, p. 2]. For validity it is "the uses and interpretations of the results of an assessment that are validated rather than the assessment itself" such that it is "a matter of degree rather than an all-or-none judgment" supported by "multiple types of evidence" [5, p. 6]. See Table 1 for suggested evidences.

A common approach to establishing reliability in PBAs is to seek inter-rater reliability, with a subset scored by multiple individuals [49, 50]; generally, two raters are sufficient [23]. Inter-rater reliability tends to be highest when the tasks are constrained [21], but for open-ended design tasks, this presents a challenge. However, by grounding the coding scheme in the language and expectations of the task, the reliability is higher compared to when generalized coding schemes are used [23].

When considering validity (Table 1), Moss [24] highlights the importance of incorporating stakeholder/participant views, asking, "Is it more valid to evaluate performances isolated from the everyday context in which they were produced?" [51]. Because constructs "are value-laden and socially dependant *[sic]*" [34] validity should also take this into account. DeLuca [2] maintains the importance of multiple perspectives in validity arguments: "more sensitive, complex and multi-perspective validity arguments [...] serve to move the field of validity forward in ways that respond to the contemporary purposes and multiple uses of educational assessments" [34]. Transparency consequences, fairness, transfer, cognitive complexity, content quality, content coverage, meaningfulness, and cost should also be considered as part of arguing for validity [41, 52]. In terms of feasibility, time-cost is particularly salient [26].

Table 1. Validity Dimensions relevant to assessment in general and performance assessment in particular, summarized from the literature

Traditional dimensions of validity *[35, 36]*	**Dimensions of validity as argument**
Construct validity: assessment measures the construct it says it measures, has internal correlations/structure	**Credibility**: Confidence in the assessment that it is congruent with established/desired practices **Directness**: Skills are directly measured [37-39]. **Cognitive complexity**: Task analysis shows assessment requires problem solving, metacognition [39-41].
Content validity: degree to which the assessment tasks and their format represent and measure relevant content/practice	**Scope/Coverage**: Set of knowledge and skills required for performing the activity is included in the assessment [39]. **Transparency**: Terms of evaluation of performance (coding scheme) are available to students and public, library of exemplars available [39, 40, 42, 43] **Authenticity**: Reflects real world content in context [30, 39, 44] **Meaningfulness**: Includes worthwhile educational activities, includes stakeholder voices [39, 40, 45, 46] **Quality**: Content reflects field, as judged by content experts [39]
Criterion validity: degree to which the assessment tasks are systematically related to an outcome; correlations with other assessments **Generalizability/ external validity**: Extensibility outside local context	**Systematic validity**: Assessment induces changes in educational system that enhance its ability to foster learning [37, 41, 46, 47] **Fairness/Bias**: Equitable access to resources, opportunities to learn, prior experiences [39-41] **Consequences**: (Un)intended effects on teaching and learning [37, 40-42, 46, 47] **Transferability/Particularizability**: Applicability in other settings yet includes particulars embedded in contexts, thick description [41, 42]. **Ecological validity**: Approximates real world performance [48]

Methodology

Study design and research questions

In this paper, we detail the initial and recent development of the design skills test (DST), a measure of problem framing ability. We describe the coding scheme we developed and implementation of the DST to assess guide and assess the impact of curricular changes. We then share the approaches we have taken to making coding feasible, from assessing the reliability to developing a new self-directed training for coders. We address the following research questions:

- To what extent are DSTs valid for informing faculty of the development of problem framing skills, using validity-as-argument dimensions?
- To what extent is the coding scheme, which measures factual/conceptual design problem representation, design practices, and design style, able to be applied in a feasible yet reliable manner across coders?

Developing an assessment of design problem framing ability

The DST and coding scheme were originally developed in 2005 to assess the impact of changes to a capstone biomedical engineering (BME) design course [3, 4, 53]. That version included an authentic design scenario written by an expert—an engineer with industry experience who was called upon to serve as an expert witness in trials that involved engineering. The scenario involved designing a blood-rewarming device that could be dropped from a helicopter and powered by the human heart. We recruited engineering faculty who also had industry experience to complete the task as a think-aloud protocol. This guided development of the coding scheme and established ecological validity. The coding scheme was refined based on the literature on design process and novice versus expert approaches to design. We grounded the initial codes in expert and student responses. This allowed us to characterize (1) the kinds of factual and conceptual information used to frame the problem in terms of needs/constraints[1]; (2) the design practices used (e.g., generating ideas, considering multiple stakeholders, remaining tentative about ideas); and (3) stylistic choices, such as how they organized their response and how much context they depicted in representations. We established that two coders with minimal training and limited expertise in engineering could reliably apply the coding scheme [4].

In 2016, we defined principles for creating new versions of the DST for chemical engineering (ChemE). In addition to needing to be accessible to all students who might complete it (which in this case, included incoming first-year students through capstone students), the DST should be based on "an authentic, real-world design problem that has yet to be solved, and that would require significant effort, time, and expertise to solve" [54]. To create a set of DSTs for use in ChemE, we identified two problems in an email from the Deutscher Technologiedienst GmbH (used with permission), an interdisciplinary technology problem solving company. A consulting engineer drafted several other options, from which we chose one that best met the criteria detailed above, resulting a set of three DSTs (Table 2).

Table 2. Description of versions of ChemE Design Skills Tests (DSTs) and the number coded by two or more people

Design skills test version	***Description***	***When used (number)***
Dishwasher	When dirty dishes sit in a dishwasher prior to running it, bacteria grow on them and produce unpleasant odors. This challenge asks students to consider how they would begin designing and prompts them to consider: • Reduction of odor to a barely perceptible level • Maintenance-free • Approx. 10 year service life (like the dishwasher itself) • No residues or other effects on the washing up in the machine • No inherent smell • Easy to integrate • Autonomous system – no need to switch on or add cleaning agents etc.	First year, Senior year (824 DSTs coded by 15 people)

[1] Design requirements encompass both needs and constraints. Depending on how a design problem is framed, a constraint can *flip* and become a need, or vice versa. For instance, the *maximum cost* of a designed solution can be considered a constraint created by the market, but a *low cost solution* can be considered a need for a specific group of stakeholders.

	• Cheap • Not sensitive to water and steam	
Diaper	Adult patients who are cared for due to dementia or similar and who wear incontinence products are inconvenienced when changed too early. Even checking whether changing is needed requires waking the patient and undressing them, a time-intensive task. This challenge asks students to consider how they would begin designing and prompts them to consider: • It must be as simple as possible to integrate the sensor into the incontinence product, i.e. the sensor will not be fixed onto the body facing surface of the product when the product is put on the patient • Simple application/integration into the manufacturing process (printing, coating, etc.) • Cordless transmission of the degree of moisture to a receiver (e.g. traffic light system) • Flexible sensor material without metal components • Check of level of product saturation has to be possible without undressing/awakening of residents • Low additional costs compared to modern incontinence products	Sophomore year (80 DSTs coded by 4 people)
Paint	When painting a wall, preparations to avoid painting on undesired surfaces from over-painting and paint splatter is the most time-consuming aspect of the process (e.g., using painter's tape and drop clothes to cover surfaces). This challenge asks students to consider how they would begin designing and prompts them to consider: • As much as possible, minimize the amount of paint splatter when applying paint to surfaces. • Paint distribution methods can include rollers and brushes or any other painting method. • Should result in at least a 25% reduction in the amount of time it usually takes to prep and paint a 1000 ft^2 home or apartment. • Should not add more than $300 to the cost of painting a 1000 ft^2 home or apartment.	Junior year (144 DSTs coded by 7 people)

Experts completed all versions and we pilot tested the assessment with a small group of students. We adapted the coding scheme for each problem and two experts applied the coding scheme to the student work with no training. Based on this, we refined the code descriptions where they were ambiguous (Table 3).

Table 3. Sample factual/conceptual, practice, and stylistic codes. Some codes were specific to the version noted; in no version is noted, it was identical across versions

Code	***Description***	***Value 1***	***Value 0***	***Value -1***
Sample factual and conceptual codes used in framing the problem including attending to constraints and considering needs outlined in the design brief.				
Constraint_ Alarm *(Diaper)*	Alarm must be triggered automatically	Mentioned clearly	Not violated, but not mentioned clearly	Constraint is violated
Constraint_ Cost *(Paint)*	The cost of the solution is considered	Affordability/cost is directly mentioned, or $300 is mentioned	Not violated, but not mentioned clearly	Cost is high, involving robotics, nanoparticles, etc.
Needs_ Reduce odor *(Dishwasher)*	Reduce odor to barely perceptible level	Mentioned odor, possibly suggests a need to measure	Mentioned vaguely	No mention directly of odor

Needs_ Splatter *(Paint)*	Prevent paint splatter.	Mentioned clearly	Mentioned vaguely or indirectly	Not mentioned
Design practice codes				
Roles *(Dishwasher)*	People who use the dishwasher, service it, manufacture it are mentioned.	At least one person is mentioned	No direct mention of specific people, but "you" is mentioned.	No mention of people, directly or indirectly other than self ("I")
Roles *(Paint)*	Painters, manufacturers, clients are mentioned	At least one person is mentioned	No direct mention of specific people, but "you" is mentioned.	No mention of people, directly or indirectly other than self ("I")
Use-Case	Describes how the design is used, envisions use	Vivid or clear description with details, even if constraints are violated	A vague description of use, hard to picture	No sense of how design would be used
Scaffolding to solution	Experienced designers plan steps toward solution	Puts forth steps toward solution	No steps, but no firm solution put forth	Solution put forth
Ideation	New ideas presented about the problem or possible solutions. Not restating.	More than one idea present	One idea present	No ideas present
Tentative language	Uses tentative language to discuss design ideas	Could be/ might be, maybe	Mix of both	Should, must, need to be, have to be
Diagram_ Function	Diagram depicts function of design or how system works	Diagram depicts function or system	Diagram depicts simple or partial function	No diagram
Stylistic codes				
Diagram_ Context	Diagram depicts design context	Diagram depicts detailed context with several elements of the design	Diagram depicts iconic or simple context, or labels context	No diagram
Organized response	Student response is organized, includes headings, sections	Includes headings and sections, such as the problem, constraints.	Includes a list, but not clearly organized a priori	No clear markers of organization, though writing may be organized

Data sources and analysis

Data were collected over a 4-year period in BME and a 3-year period in ChemE to track the impact of introducing design challenges two research universities. The DST is intended to assess how students get started in design problem framing. Students are therefore given only 15 minutes in class to work on the problem. Students are reassured that we are interested in how they get started, and that it would take a team many months to reach a viable solution. Typically, students fill one page of text with writing and drawings in that time. The DST is given at the beginning, a midpoint, and end of each academic year that includes design challenges. In BME, this included only capstone design, but in ChemE, this included first-year through capstone. Thus, in the latter, we collected hundreds of completed tests each year. To make coding feasible, we recruited multiple undergraduates who had completed at least two ChemE courses. We assigned two coders to each completed DST and cross-classified coding assignments (Figure 1, however, some did not complete their assignments; these were assigned to other coders). Coders received minimal training as follows: all coders received 20 *dishwasher DSTs* and applied the coding scheme independently. As a group, we discussed two in detail, then solicited questions and clarifications. A key correction included that "credit" should not be given, but rather, coders should use positive and negative scores to *characterize* the response. During five coding sessions, the primary researcher responded to their questions, which were few. Coders were encouraged to mark their uncertainly using X999 rather than making a choice.

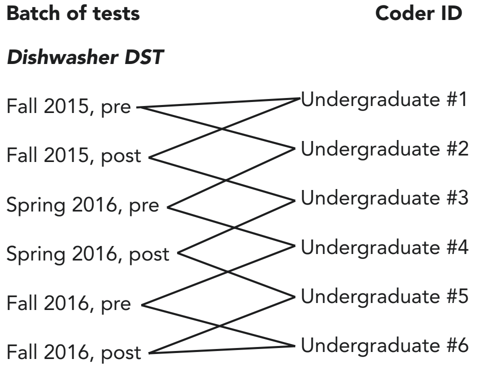


Figure 1. Simplified depiction of cross-classified coder assignment.

We analyzed the reliability of coded data in Excel. We organized the data by DST, participant, code, and coder. We calculated the percentage match of each code and pair of coders respectively, and organized these by code type (factual/conceptual, practice, style), omitting codes marked by coders as uncertain. Because of the cross-classified nature of data, typical measures such as Cohen's Kappa are not appropriate because these assume two identical coders scored all data [55]. Likewise, while Fleiss' Kappa allows calculation with multiple coders and does not require the coders to be identical [56], it does require the same total number of coders for each datapoint [57]. Our data violate this assumption, and thus we rely on the practical calculation of percent agreement; we acknowledge this represents an inflated estimate.

Results

Using the validity-as-argument dimensions (Table 1), we found the DST provided valid information on design problem framing ability for course improvement purposes (Table 4). This is grounded in findings that experts viewed the tests as requiring problem framing skills, that expert performance on the DSTs displays problem framing ability at an expert level, and that expert review of the DST coding scheme established that it was immediately recognizable as distinguishing between work that was or was not characteristic of design problem framing. The DST was sensitive to and directed changes made in the BME course [3, 4, 53]. The course included a mini-design challenge in the first two months. Students completed a kit-based challenge of assembling a digital stethoscope. While the instructor was not pleased with how difficult the task was for students to successfully complete, she was convinced to replace the challenge when she saw the precipitous drop from a pre-course average of 42% to a post-mini-design average of 20% on representing stakeholder needs (though students showed growth in areas coded as stylistic, suggesting they gained increased understanding of disciplinary norms). In following years, after replacing the kit-based challenge with a redesign challenge that emphasized identifying stakeholder needs, the DST showed evidence of growth in representing stakeholder needs [58].

In terms of reliability, when using coders with relatively little training (as was the case in our study), we conjectured that factual/conceptual codes would be most reliable because they tend to be low inference and require very low background knowledge. We conjectured that the practice codes would have lower reliability than factual codes because these require higher inference and knowledge of design practice can influence the judgment one makes. We conjectured that training could improve scoring of practice codes. We conjectured that the stylistic codes would have similar reliability to factual/conceptual codes because these require a similar level of inference. These codes do not rely on specific domain knowledge, even when reflecting disciplinary norms[2], and therefore are less likely to be improved by training. We conjectured that there would be similar rates of agreement across versions of the DST. In general, we conjectured that there would be variability across codes, with tentative language being less reliable than other codes overall.

Table 4. Validity as argument for design skills tests.

Dimensions of validity-as-argument	**As assessed in design skills tests**
Credibility: Confidence in the assessment that it is congruent with established/desired practices **Directness**: Skills are directly measured [37-39]. **Cognitive complexity**: Task analysis shows assessment requires problem solving, metacognition [39-41].	DSTs are congruent with design problem framing practice and directly measure the skills. The scenarios and skills require students to engage problem solving and metacognitive skills.
Scope/Coverage: Set of knowledge and skills required for performing the activity is included in the assessment [39]. **Transparency**: Terms of evaluation of performance (coding scheme) are available to students and public, library of exemplars available [39, 40, 42, 43] **Authenticity**: Reflects real world content in context [30, 39, 44] **Meaningfulness**: Includes worthwhile educational activities, includes stakeholder voices [39, 40, 45, 46] **Quality**: Content reflects field, as judged by content experts [39]	Each DST provides contextual detail and design requirements required to frame the problem. The DST is not currently transparent, but because it is not used to make high stakes decisions, this is appropriate. Each DST task reflects real world contexts by using actual, unsolved design problems. DST scenarios were pulled from stakeholder materials. Experts judged the DSTs and coding scheme to reflect the expectations of design problem framing.
Systematic validity: Assessment induces changes in educational system that enhance its ability to foster learning [37, 41, 46, 47] **Fairness/Bias**: Equitable access to resources, opportunities to learn, prior experiences [39-41] **Consequences**: (Un)intended effects on teaching and learning [37, 40-42, 46, 47] **Transferability/Particularizability**: Applicability in other settings yet includes particulars embedded in contexts, thick description [41, 42]. **Ecological validity**: Approximates real world performance [48]	The DSTs are used to assess the impact of curricular changes. Only in making their analysis feasible will this lead to systemic change. While full solution of a DST would reveal inequities, *framing* the problem does not. A few students had noted they had never painted, never owned a dishwasher, or never interacted with diapers, yet their responses were not qualitatively different from their peers and their responses were sensible, suggesting the problem contexts were accessible to them. No unintended effects on instruction have been observed to date. The DSTs approximate real world performance, including as assessed by industry experts.

[2] In this case, this is because disciplinary norms included only whether or not to represent problem context in the diagram.

On the *dishwasher DST*, we found the overall agreement between coders for factual/conceptual and stylistic codes was higher than the practices codes, as conjectured (Figure 2). The average overall percent agreement by code type between any two coders ranged from 65% on factual/conceptual codes, 52% on practice codes, to 68% on stylistic codes.

On the *diaper DST*, we again found the overall agreement between coders for factual/conceptual and stylistic codes was higher than the practices codes, as conjectured (Figure 3). The average overall percent agreement by code type between any two coders ranged from 70% on factual/conceptual codes, 57% on practice codes, to 78% on stylistic codes.

On the *paint DST*, we again found the overall agreement between coders for factual/conceptual and stylistic codes was higher than the practices codes, as conjectured (Figure 4). The average overall percent agreement by code type between any two coders ranged from 83% on factual/conceptual codes, 77% on practice codes, to 80% on stylistic codes.

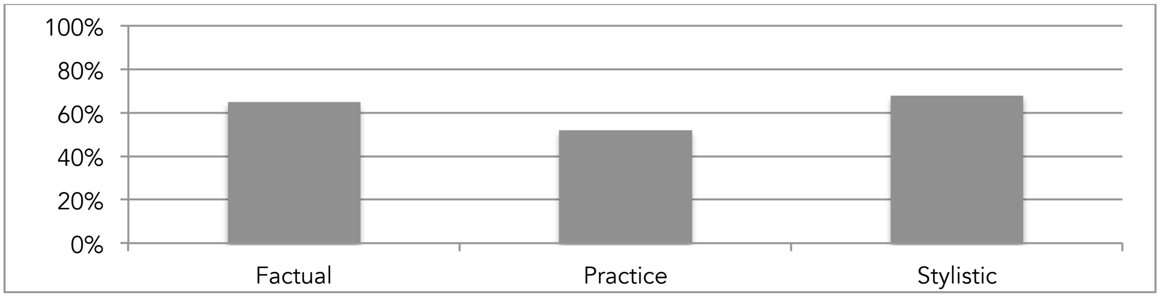


Figure 2. Average percent agreement between any two coders, by code type, *Dishwasher DST*

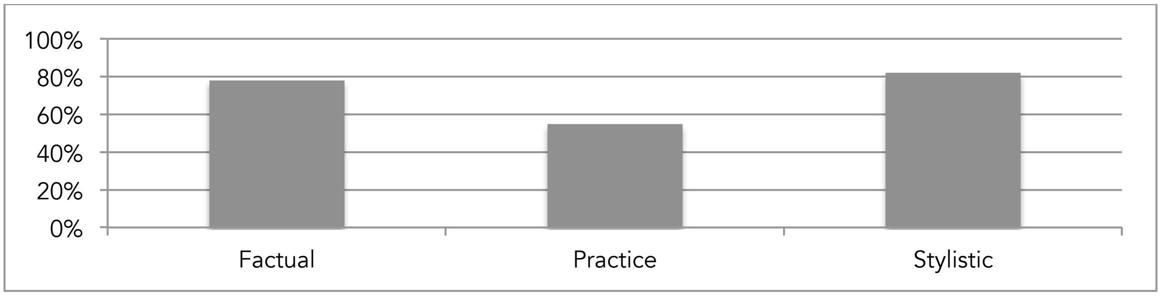


Figure 3. Average percent agreement between any two coders, by code type, *Diaper DST*

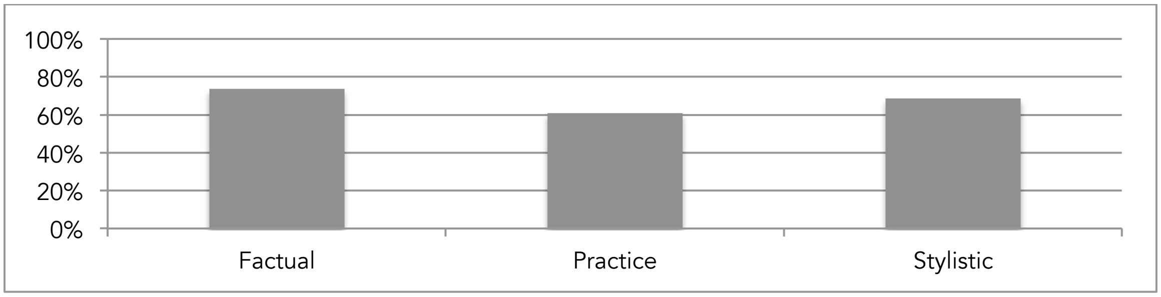


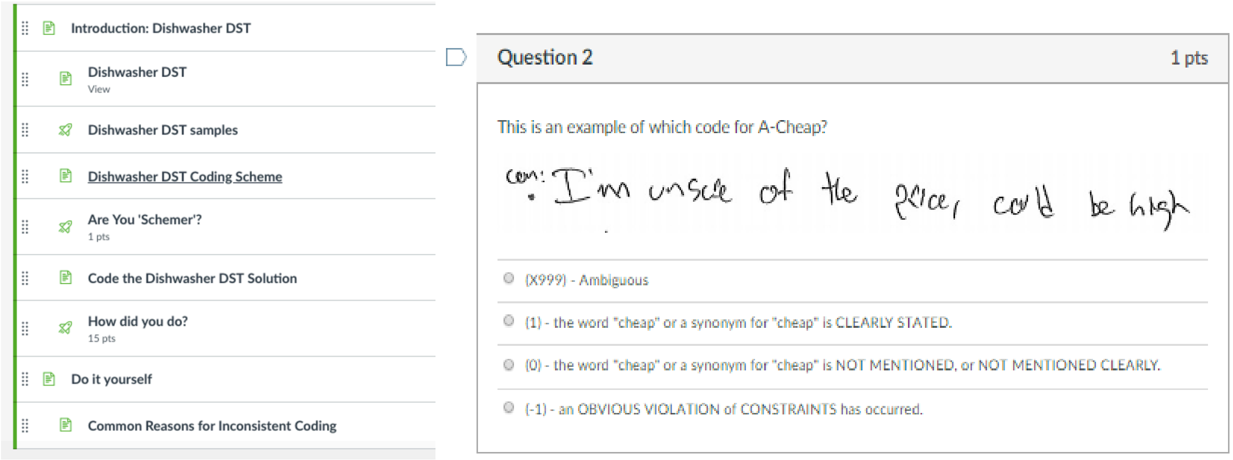
Figure 4. Average percent agreement between any two coders, by code type, *Paint DST*

Although the profile of reliability across code types (factual/conceptual, practice, style) aligned to our conjectures, we did find variability across the versions. This may be explained in part by the number of tests coded and the number of coders involved, in part because of the cross-classified coding assignment approach (Table 2). There were fewer *diaper DSTs* coded by few coders. As we continue this work and expand the number of coders, we anticipate the overall reliability scores for this version would align somewhat more with the other versions.

Overall, we found that the two codes with the lowest reliability were use-case and tentative language. By removing the use-case and tentative language codes, the practice codes reach a similar level of reliability to other codes. This suggests having coders with limited training requires restricting the code set somewhat to reach minimal levels of reliability. However, it is important to note that because we were not able to report a kappa statistic, these rates are likely overestimates. In line with validity-as-argument, we would not see these scores as valid assessments of individual progress at this point; however, we do think the scores could be used to make decisions about overall trends. To enhance this capacity, we developed a self-paced training program in coordination with instructional designers.

Designing a DST self-paced training

We developed a self-paced, online training to provide step-by-step instruction on how to code DSTs. The training includes a 10-minute introductory module (10 minutes to complete) that situates the importance of accurate coding, explains that coding is neither computer programming nor grading, and explains that it is akin to characterizing material properties (a metaphor we hope is accessible to our engineering undergraduate students). For each DST (diaper, dishwasher, paint), there is a specific training module, and coders complete only the module for the DST they will be coding (Figure 5, left). Each DST-specific module introduces the DST scenario and coding scheme using a worked example and brief formative assessments (Figure 5, right). The DST-specific module takes approximately 30 minutes to complete, not including the time required for coding. Learners are directed to complete an initial coding assignment of 20, compare codes with a partner, discussing and resolving disagreements. Learners watch a short video that models how to resolve coding disagreements. After completing their coding assignment, learners can return to the training to compare their reliability with another coder using an embedded interrater reliability tool. Completing the course training through either path will lead to a printable certificate of completion.


Introduction: Dishwasher DST
Dishwasher DST
View
Dishwasher DST samples
Dishwasher DST Coding Scheme
Are You 'Schemer'?
1 pts
Code the Dishwasher DST Solution
How did you do?
15 pts
Do it yourself
Common Reasons for Inconsistent Coding

Question 2 1 pts

This is an example of which code for A-Cheap?

Con: • I'm unsure of the price, could be high

○ (X999) - Ambiguous

○ (1) - the word "cheap" or a synonym for "cheap" is CLEARLY STATED.

○ (0) - the word "cheap" or a synonym for "cheap" is NOT MENTIONED, or NOT MENTIONED CLEARLY.

○ (-1) - an OBVIOUS VIOLATION of CONSTRAINTS has occurred.

Figure 5. Overview of the Dishwasher DST module and sample formative assessment

Scholarly significance

Overall, we found the DST provided valid information to instructors about whether or not their design challenge supported students to develop problem framing ability. We presented a feasible approach to dealing with the time-consuming coding process—involving undergraduate students with minimal training. While the reliability with minimal training was lower than would be acceptable for research purposes, we argue that for instructional purposes—assigning two undergraduates to code only the factual and stylistic codes (which represent 60% of the codes) that had consistently higher levels of agreement between any two coders—this represents a significant reduction of instructor effort.

Our decision to cross-classify coder assignment provided a means to evaluate each coder, a focus of our ongoing research. We also note that careful documentation of metadata was critical to our process, as the DSTs were given each semester and many individuals have worked with the data over several years. However, a limitation of the current study is that we did not end up with a dataset that lent itself to use of a kappa statistic to account for chance agreement. Our future work will investigate ways to ensure that there are always the same number of coders involved. To support this, we have created an interrater reliability calculation tool that provides guidance on the number of coders needed per item and calculates the Fleiss kappa [56].

Future studies will evaluate the impact of a self-paced online training on reliability. For smaller classes, minimal training paired with faculty review of non-consensus codes may be more feasible than providing training, but for larger classes, we conjecture that the training will improve reliability. Retaining one well-trained undergraduate student from one year to the next could also enhance the process by ensuring that there is at least one experienced coder in the team. If it is successful, this approach could be expanded to other areas, from grading complex assignments to introducing new researchers to qualitative analysis.

Ongoing studies are investigating variability across DST versions used in ChemE. This involves within-coder and cross-version comparisons, as well as comparing instances in which a student was enrolled in two levels of courses at the same time (common for transfer students and students). This effort will provide further evidence for validity of these, as well as our principles, for using DSTs to guide instructors to develop design challenges that support students to develop problem framing skills.

While we view our approach as feasible and transferrable, our context differs from others. As a research institution, we have access to funds to hire undergraduates, many of whom desire opportunities to work on grant-funded research projects like this one. Like a number of engineering departments (as evidenced by the NSF program, *Revolutionizing Engineering and Computer Science Departments*), ours is in the midst of a major effort to improve our ability to support diverse student success. These factors are relevant when considering ways to transfer our approach to other institutional contexts.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. #1623105. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also thank the students who participated in initial coding efforts.

References

[1] J. W. Getzels, "The problem of the problem," *New directions for methodology of social and behavioral science: Question framing and response consistency,* vol. 11, pp. 37-49, 1982.

[2] J. W. Getzels, "Problem ‐ finding and the inventiveness of solutions," *The Journal of Creative Behavior,* vol. 9, pp. 12-18, 1975.

[3] V. Svihla, A. J. Petrosino, T. Martin, and K. R. Diller, "Learning to design: Interactions that promote innovation," in *Innovations 2009: World Innovations in Engineering Education and Research*, W. Aung, K.-S. Kim, J. Mecsi, J. Moscinski, and I. Rouse, Eds., ed Arlington, VA: International Network for Engineering Education and Research, 2009, pp. 375-391.

[4] V. Svihla, "Collaboration as a dimension of design innovation," *CoDesign: International Journal of CoCreation in Design and the Arts,* vol. 6, pp. 245-262, 2010.

[5] D. H. Jonassen, "Toward a Design Theory of Problem Solving," *Educational Technology Research and Development,* vol. 48, pp. 63-85, 2000.

[6] K. Dorst, "The Design Problem and its Structure," in *Analysing Design Activity*, N. Cross, H. H. C. M. Christiaans, and K. Dorst, Eds., ed Chichester: Wiley, 1996, pp. 17-34.

[7] K. Dorst, "The Problem of Design Problems," *Design Thinking Research Symposium, Sydney,* vol. 17, 2003.

[8] V. Goel and P. Pirolli, "The Structure of Design Problem Spaces," *Cognitive Science,* vol. 16, pp. 395-429, 1992.

[9] J. Restrepo and H. Christiaans, "Problem Structuring and Information Access in Design," *Expertise in design,* 2003.

[10] N. Cross, "Design cognition: Results from protocol and other empirical studies of design activity," in *Design knowing and learning: Cognition in design education*. vol. 5, C. M. Eastman, W. M. McCracken, and W. C. Newstetter, Eds., ed Oxford, UK: Elsevier Science, 2001, pp. 79-103.

[11] M. Mehalik and C. Schunn, "What constitutes good design? A review of empirical studies of design processes," *International Journal of Engineering Education,* vol. 22, p. 519, 2007.

[12] K. M. Bursic and C. J. Atman, "Information gathering: a critical step for quality in the design process," *Quality Management Journal,* vol. 4, 1997.

[13] C. J. Atman, R. S. Adams, M. E. Cardella, J. Turns, S. Mosborg, and J. Saleem, "Engineering Design Processes: A Comparison of Students and Expert Practitioners," *Journal of Engineering Education,* vol. 96, p. 359, 2007.

[14] C. J. Atman, K. Yasuhara, R. S. Adams, T. J. Barker, J. Turns, and E. Rhone, "Breadth in Problem Scoping: a Comparison of Freshman and Senior Engineering Students," *International Journal of Engineering Education,* vol. 24, pp. 234-245, 2008.

[15] P. G. Dominick, *Tools and tactics of design*: Wiley, 2001.

[16] M. Basadur, G. B. Graen, and S. G. Green, "Training in creative problem solving: Effects on ideation and problem finding and solving in an industrial research organization," *Organizational Behavior and Human Performance,* vol. 30, pp. 41-70, 1982.

[17] A. Morozov, D. Kilgore, and C. Atman, "Breadth in design problem scoping: Using insights from experts to investigate student processes," in *American Society for Engineering Education Annual Conference and Exposition*, 2007.

[18] W. A. Nelson, "Problem Solving Through Design," *New Directions for Teaching and Learning,* vol. 2003, pp. 39-44, 2003.

[19] J. Kolko, "Abductive Thinking and Sensemaking: The Drivers of Design Synthesis," *Design Issues,* vol. 26, pp. 15-28, Winter2010 2010.

[20] K. Dorst, "The core of 'design thinking' and its application," *Design Studies,* vol. 32, pp. 521-532, 2011.

[21] W. Visser, "Design: one, but in different forms," *Design Studies,* vol. 30, pp. 187-223, 2009.

[22] A. Rourke and J. Sweller, "The worked-example effect using ill-defined problems: Learning to recognise designers' styles," *Learning and Instruction,* vol. 19, pp. 185-199, 2009.

[23] E. P. Douglas, M. Koro-Ljungberg, N. J. McNeill, Z. T. Malcolm, and D. J. Therriault, "Moving beyond formulas and fixations: solving open-ended engineering problems," *European Journal of Engineering Education,* vol. 37, pp. 627-651, 2012.

[24] T. W. Banta and C. A. Palomba, *Assessment essentials: Planning, implementing, and improving assessment in higher education*: John Wiley & Sons, 2014.

[25] L. Osgood and C. R. Johnston, "Design Ability Assessment Technique " in *ASEE Annual Conference & Exposition*, Indianapolis, Indiana, 2014.

[26] P. A. Kirschner and J. J. G. van Merriënboer, "Do Learners Really Know Best? Urban Legends in Education," *Educational Psychologist,* vol. 48, pp. 169-183, 2013.

[27] J. J. Shah, "Identification, measurement and development of design skills in engineering education," in *DS 35: Proceedings ICED 05, the 15th International Conference on Engineering Design, Melbourne, Australia, 15.-18.08. 2005*, 2005.

[28] B. Stecher, *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford Center for Opportunity Policy in Education, 2010.

[29] Research for Action, "Using Common Assignments to Strengthen Teaching and Learning: Research on the 1st Year of Implementation," 2014.

[30] J. T. M. Gulikers, L. Kester, P. A. Kirschner, and T. J. Bastiaens, "The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes," *Learning and Instruction,* vol. 18, pp. 172-186, 2008.

[31] S. Messick, "Validity," in *Educational measurement* R. L. Linn, Ed., ed New York, NY: Macmillan Publishing, 1989.

[32] L. A. Shepard, "Evaluating test validity," *Review of Research in Education,* vol. 19, pp. 405-450, 1993.

[33] L. J. Cronbach, "Five perspectives on validity argument," *Test validity,* pp. 3-17, 1988.

[34] C. DeLuca, "Interpretive validity theory: mapping a methodology for validating educational assessments," *Educational Research,* vol. 53, pp. 303-320, 2011.

[35] L. J. Cronbach, *Essentials of psychological testing*, 5th ed. New York: Harper & Row, 1990.

[36] American Psychological Association, "Standards for psychological and educational testing," *American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Washington, DC,* 1985.

[37] S. Dierick and F. Dochy, "New lines in edumetrics: New forms of assessment lead to new assessment criteria," *Studies in educational evaluation,* vol. 27, pp. 307-329, 2001.

[38] A. M. Uhlenbeck, "The development of an assessment procedure for beginning teachers of English as a foreign language," 2002.

[39] L. Baartman, T. Bastiaens, P. A. Kirschner, and C. Van der Vleuten, "The wheel of competency assessment: Presenting quality criteria for competency assessment programs," *Studies in Educational Evaluation,* vol. 32, pp. 153-170, 2006.

[40] R. K. Hambleton, "Advances in assessment models, methods, and practices," in *Handbook of educational psychology*, D. C. Berliner and R. C. Calfee, Eds., ed New York, NY: MacMillan, 1996, pp. 899-925.

[41] R. L. Linn, E. L. Baker, and S. B. Dunbar, "Complex, performance-based assessment: Expectations and validation criteria," *Educational Researcher,* vol. 20, pp. 15-21, 1991.

[42] G. S. Maxwell, "Dealing with inconsistency and uncertainty in assessment," 2009.

[43] R. Donmoyer, "Generalizability and the single-case study," *Qualitative inquiry in education: The continuing debate,* pp. 175-200, 1990.

[44] E. G. Guba and Y. S. Lincoln, "Competing paradigms in qualitative research," *Handbook of qualitative research,* pp. 105-117, 1994.

[45] L. McDowell, "The impact of innovative assessment on student learning," *Innovations in Education and Training International,* vol. 32, pp. 302-313, 1995.

[46] S. Messick, "The Psychology of Educational Measurement," *Journal of Educational Measurement,* vol. 21, pp. 215-237, 1984.

[47] L. W. Schuwirth and C. P. Van Der Vleuten, "Changing education, changing assessment, changing research?," *Medical education,* vol. 38, pp. 805-812, 2004.

[48] G. E. McKechnie, "Simulation techniques in environmental psychology," in *Perspectives on environment and behavior*, ed: Springer, 1977, pp. 169-189.

[49] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational Research Review,* vol. 2, pp. 130-144, 2007.

[50] S. J. Meisels, F. Liaw, A. Dorfman, and R. F. Nelson, "The Work Sampling System: Reliability and validity of a performance assessment for young children," *Early Childhood Research Quarterly,* vol. 10, pp. 277-296, 1995.

[51] P. A. Moss, "Enlarging the dialogue in educational measurement: Voices from interpretive research traditions," *Educational Researcher,* vol. 25, pp. 20-29, 1996.

[52] P. A. Moss, "Shifting conceptions of validity in educational measurement: Implications for performance assessment," *Review of Educational Research,* vol. 62, pp. 229-258, 1992.

[53] V. Svihla, A. J. Petrosino, and K. R. Diller, "Learning to design: authenticity, negotiation, and innovation," *International Journal of Engineering Education,* vol. 28, pp. 782-298, 2012.

[54] V. Svihla, A. Dayte, J. R. Gomez, V. Law, and S. Bowers, "Mapping Assets of Diverse Groups for Chemical Engineering Design Problem Framing Ability," *Proceedings of ASEE's 123 Annual Conference & Exposition,* pp. 1-25, 2016.

[55] M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha, "Beyond kappa: A review of interrater agreement measures," *Canadian journal of statistics,* vol. 27, pp. 3-23, 1999.

[56] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin,* vol. 76, p. 378, 1971.

[57] A. A. Mitani, P. E. Freer, and K. P. Nelson, "Summary measures of agreement and association between many raters' ordinal classifications," *Annals of epidemiology,* vol. 27, pp. 677-685. e4, 2017.

[58] V. Svihla, A. J. Petrosino, T. Martin, K. Rayne, S. R. Rivale, and K. R. Diller, "Learning to design: the role of authenticity and the distribution of cognition in student design teams," in *AERA*, New York, NY, 2008.