Abstract Title Page

Title: Validity as Process: A Construct Driven Measure of Fidelity of Implementation

Authors and Affiliations: Ryan Seth Jones, Vanderbilt University

Background / Context:

Estimates of fidelity of implementation are essential to interpret the effects of educational interventions in randomized controlled trials (RCTs). While random assignment protects against many threats to validity, and therefore provides the best approximation to a true counterfactual condition, it does not ensure that the treatment condition faithfully reflects researchers' intentions (Lipsey, 1993). If program success relies upon changes in practices, routines, or behaviors of participants, which is the case in most K-12 education interventions, then causal inferences about success or failure rely on a valid account of the various states of the program in treatment conditions.

Although fidelity measures are widely considered a necessary component of field experiments there is no consensus on how one should go about this important work (e.g. Cordray & Pion, 2006; O'Donnell, 2008). Many agree that measures should be grounded in an a-priori program theory (O'Donnell, 2008), but there is very little direction on what constitutes a sufficient theory (Donalson & Lipsey, 2006). As a result, researchers rarely articulate the theories that guide the construction of their fidelity measures, and the few that do provide very general descriptions that are open to multiple interpretations (e.g. Nelson et. al, 2012).

Fidelity measures without clearly articulated program theories are at risk for the two biggest threats to validity: construct underrepresentation and construct irrelevant variance (Messick, 1994). It is no surprise then that validity evidence for fidelity measures is rarely reported (Hagermoser-Sanetti & Kratochwill, 2009), and when validity evidence is reported it is typically done so only at the instrument level.

I argue that fidelity measures should be construct driven to address these threats to validity (Messick, 1994). Construct driven measures use visible performance to make inferences about invisible theoretical constructs. In contrast, many fidelity measures are designed as task driven assessments. Although task driven measures are sometimes appropriate in other contexts, such as judging an athlete's performance in the Olympic games, the goal of measuring fidelity is not to simply describe behavior during a specific moment in time. Instead, fidelity measures use a sample of observational moments to make inferences about overall implementation. So, to establish a measure's validity researchers should clearly articulate three things: latent program theory, observable evidence of program theory, and the nature of the correspondence between them.

Purpose / Objective / Research Question / Focus of Study:

I used Wilson's (2005) four building blocks of measurement to guide the development of a fidelity measure in a randomized controlled trial of the middle school statistics curriculum, *Data Modeling*. I found this framework useful due to its emphasis on all three components of construct driven measurement. Guided by the four building blocks (Figure 1), this presentation illustrates the use of this framework by addressing the following questions.

1. What are the latent constructs that make up the *Data Modeling* program theory?
2. What would constitute an "item" in the context of *Data Modeling* instruction?
3. What are the likely outcomes from our observations, and how do they relate to our construct?
4. How do we model our measurements to provide inference about the construct?

Significance / Novelty of study:

This fidelity measure is novel in two regards. First, it frames validity as establishing a relationship between latent theory and observation, which is always uncertain. In contrast, validity is sometimes treated as a once-and-for-all event (Wilson, 2006). Wilson's (2006) framework, however, treats validity as an ongoing, iterative process of construct conceptualization, operationalization, and measurement modeling. It is through this process that the validity of inferences made from observable data is maintained.

Second, this framework has yet to be used for a fidelity measure. Many of the terms and concepts within the framework have a shared meaning in the context of cognitive assessments, but the meaning is less clear for fidelity measures. We discuss the meaning of these concepts for our fidelity measure, and discuss challenges we faced in our novel use of the framework.

Usefulness / Applicability of Method:

Construct. The first step in Wilson's framework is to articulate latent constructs of interest in a construct map. The construct map consists of descriptions of the different states of the construct, and orders the states in relation to the values of the researcher. For fidelity measures the construct map serves as an initial conceptualization of a fidelity scale since it orders the different states in terms of increasing fidelity. For example, the description of the highest level of the construct map describes the state of the theory under ideal implementation. The construct map is a useful grain size because it allows for descriptions of the constructs making up ones program theory that are specific enough to have meaning, but is still general enough to contribute to knowledge beyond a specific intervention.

We drew from multiple iterations of design studies (Cobb, et al. 2003) and qualitative analyses to inform our initial theoretical commitments. Through this design research the curriculum stabilized into a sequence of 7 units addressing data display, center statistics, variability statistics, general use of statistics, chance, modeling, and inference. Within these units students experience a recurring activity structure in which they consider a question, invent a method for addressing the question, and then compare and critique their methods during a whole class discussion called a "method review." For example, in the first unit students independently measure the length of a common object and are often surprised to find great variability in the measurements. It is in the face of this variability they are asked to look for patterns or trends in the data that would help them estimate the "true" length of the object. Next, they invent ways to display the data that would lead others to see "at-a-glance" the trend they noticed. Last, the students discuss the different choices made to create the displays, and the effect of these choices on what one could see (or not see) about the data.

Through this sequence we have found that although student invented methods are often different than conventional ones, they contain conceptual similarities that can be cultivated into more conventional concepts. Through previous qualitative research we developed theories about the usefulness of specific types of inventions and about productive teacher strategies for making the mathematical concepts in the inventions salient to students. For example, many students invent data displays without a consistent scale because they think it is not useful to represent numbers for which there is no data (Lehrer, Kim, & Schauble, 2007). Teachers can juxtapose these methods with displays that have a consistent scale and ask students what each shows and hides about he data. Students then have an opportunity to develop an understanding of scale that is motivated by the goal of communicating important trends in the data (e.g. scale allows students to see "gaps" in the data). These theories led us to focus our conceptualization of

fidelity of the practice of using the variability in student-invented methods to amplify mathematical concepts and practices. We articulated five categories of likely teacher practices, and ordered them in terms of increasing fidelity. Figure 2 is a simplified version of our construct map. There are three important things to notice about our map. First, the language is general enough to be applied to all of our units. However, the specific concepts we want to see emphasized are different for each unit, so we also articulated unit specific construct maps to communicate these differences. Second, the descriptions of our theory are in latent terms, that is, they cannot be directly observed. So, the map itself does not constitute a measure. Third, we are measuring teacher practice, which is inherently interactional and must be observed to measure.

Item. The second step in Wilson's framework directs researchers to develop an operationalization of their theory. The term "item" can often evoke images of test forms for cognitive assessment, but more generally an item refers to any real world situation where one's construct can be manifested in observable action (Wilson, 2005). Since we wanted to observe the instructional role of student invention we could not arbitrarily schedule observations. Instead, we needed to observe at the moments where their use was most salient. This led us to focus on the instructional moments where students were discussing their invented methods, the "method reviews." We then developed three types of observable data for each method review. The first type was a set of summary variables that described broad characterizations of the whole class. We also collected images of each student invention to account for the types produced in each class. However, teacher practice is best observed in classroom interactions. To characterize this we developed a set of binary variables to be scored in five-minute adjacent segments. These variables looked for unit general student and teacher participation, but also unit specific mathematical concepts.

Each variable was designed to index a particular level of the construct map. This rationale provided initial validity information. Through iterative piloting and discussions with intervention experts we refined these variables to better index each level. In addition, it challenged our initial conceptions of our theory. For example, our original construct map had only four levels, but piloting the observable variables revealed that we were not accounting for a common use of student-invented methods. So, we added a new level describing the use of inventions to promote only procedural understanding of conventional methods. This highlights that developing valid measures is not a linear process. Rather, it is iterative. Theory motivates operationalization, but operationalization can often challenge ones initial theory.

The critical role of instrumentation in developing fidelity measures is rarely addressed in intervention research. Occasionally researchers will reference vague notions of using multiple instruments to "triangulate" data (Nelson, et. al 2012), but they do not address the ways in which our measures are both supported and constrained by material instruments. For example, the types of data one can collect during live observations using with pencil and paper are very different than the possibilities when computers are available to aid the observer. We capitalized on new technologies to create an instrument to aid in the collection, organization, and transmission of observation data. This instrument allowed observers to mark variables, collect images, and it also automated the segmentation of the class into five-minute sections. However, it also influenced the construction of our variables since some were not feasible given the available instruments. For these reasons the role of instrumentation in validity should be more clearly articulated in fidelity measures.

Outcome Space. When researchers use the observable variables above to index a class there are a large number of possible data profiles. For example, consider a fictional, and much

simpler observation system consisting of 2 variables that could be marked "high," "medium," or "low." There are nine possible outcomes in this simple system, and each of them should inform the construct of interest in some way. However, in our system there are many more possible outcomes. The third step in Wilson's framework requires researchers to consider the outcome space of possible data profiles, and how they relate to the construct map. We began by looking for qualitative distinctions in the data profiles that related to our construct. For example, profiles indicating that student-inventions were used to discuss all relevant concepts were mapped to a higher level of the construct map than classes that only discussed one relevant concept.

We used these qualitative distinctions to generate scoring algorithms for data profiles. We considered two types of outcome spaces to generate these scoring rules. First, each of our segment variables was binary, and since they were designed to index particular levels of the construct map we could relate each possible outcome ("0" or "1") to the level of the map it corresponds to. Table 1 is a list of the segment variables and the level a score of "1" is related to. However, we also considered the outcome space of co-occurrences of particular variables. For example, if one of the mathematical concept variables is scored then we can look to the other variables to better understand how the concept was discussed.

Measurement Model. A measurement model describes theoretically grounded expectations for influences on item responses. Modeling the item responses is important to further interrogate the validity of the items and to make inferences about ones latent construct. We are still collecting data, so I do not yet have empirical illustrations of our measurement model, but we plan on using a Rasch model to examine the relationship between observed data and our construct map. The Rasch model is a member of a family of generalized linear mixed models (Wilson, 2005). We are committed to this model for three reasons. First, our data are binary so the logit link in the model allows us to estimate the probability of observing a variable with a linear combination of item and teacher practice parameters. Second, the model estimates a fixed effect for each item that describes the influence of an item on the probability of an occurrence during a segment. This provides information about the "behavior" of the items. Third, this model estimates teacher practice parameters on the same logit scale as the items, which provides information about the fidelity of implementation. For example, there should be a higher probability of observing mathematical concepts in classes with high fidelity than in those with low fidelity.

Conclusions:

Theoretically grounded measures of fidelity are desperately needed in RCTs of educational interventions. However, establishing the validity of such measures is challenging and uncertain work. We have found much needed guidance from Wilson's (2005) framework in developing measures with ongoing and iterative attention to construct validity. Deliberately thinking about construct articulation, item design, outcome space, and measurement model produced meaningful validity evidence, and allowed us to make principled refinements to increase the validity of our measure. However, as previously mentioned, we do not see validity as something that is established once-and-for-all. In this spirit, we continue to interrogate our measure, and will continue to refine it.

Appendices

Not included in page count.

Appendix A. References

Cobb, P., Confrey, J., diSessa, A., Lehrer, R. & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Cordray, D. S. & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E., McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation (pp 103-124).* Washington, DG: American Psychological Association.

Donalson, S.I & Lipsey, M.W. (2006). Roles for theory in contemporary evaluation practice: developing practical knowledge. In I. Shaw, J. Greene, & M. Mark (Eds.). *The Sage handbook of evaluation* (pp. 56-75). Thousand Oaks, CA: SAGE.

Hagermoser-Sanetti, L.M. & Kratochwill, T.R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. School Psychology Review 38(4) p. 445-459

Lehrer, R., Kim, M.J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. International Journal of Computers for Mathematical Learning, 12(3), 195 216. Lehrer, R., & R

Lipsey, M.W. (1993). Theory as method: small theories of treatments. *New directions for program evaluation,* 57, p. 5-38

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 23(2) p. 13-23

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, 78(1), 33-84.

Wilson, M. (2005). *Constructing measures*. Mahwah (NJ): Lawrence Erlbaum.

Appendix B. Tables and Figures
Not included in page count.

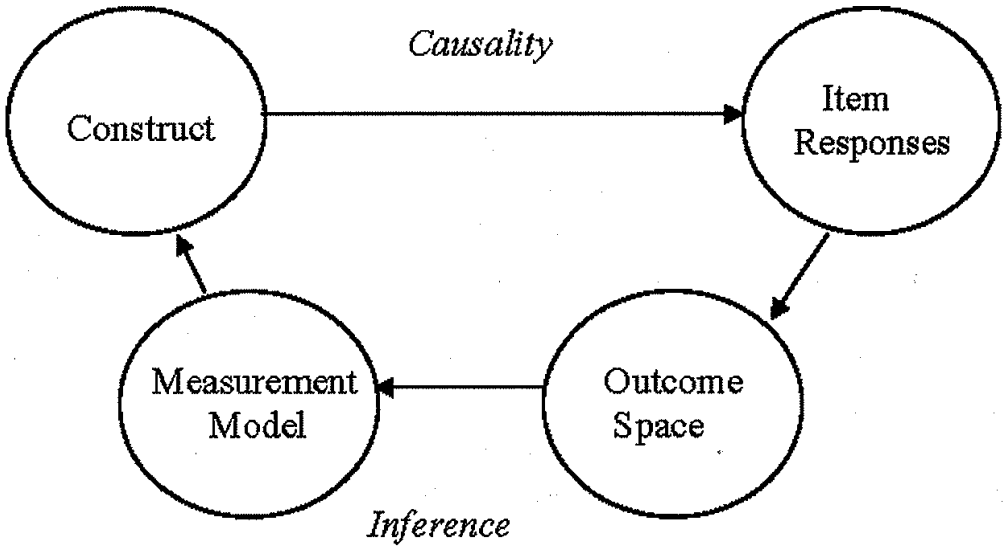

Figure 1: Wilson's four building blocks of measure development.

Level	Use of Variability in Student Invention
5	Student invented methods are seen as a resource to communicate different mathematical strategies in order to synthesize specific mathematical ideas into systems of meaning.
4	Student invented methods are seen as a resource to communicate different mathematical strategies in order to promote specific mathematical ideas.
3	Student invented methods are seen as an instructional resource to promote a right/wrong orientation towards mathematical procedures.
2	Student invented method are seen as an instructional resource to primarily increase engagement.
1	Student invented methods are not valued as an instructional resource.

Figure 2: Construct Map describing theoretical grounding for fidelity measure

Table 1: Relationship between unit 1 segment variables and construct map

	Observable Variable	**Construct MAP Level**				
		1	**2**	**3**	**4**	**5**
Student Contributions	Did students share invented displays?		X			
	Did students make comments or ask questions about the **conceptual** elements of the invented displays?				X	
	Did students make comments or ask questions about the **procedural or calculational** elements of the invented displays?			X		
Teacher Practice	Did the teacher select student-invented displays to be shared?				X	
	Did the teacher compare different ways to display data?				X	
	Did the teacher use questions similar to the ones in the curriculum to support students to think about and discuss displaying data?		X			
	Did the teacher make connections between different students' thinking?					X
	Did the teacher connect student thinking to the big ideas?				X	
	Did the teacher press students to explain their thinking?				X	
	Did the teacher **only** focus on conventional displays?	X				
	Did the teacher focus **only** on the procedural elements of creating displays?	X				
	Did the teacher **primarily** lecture on displays without letting students discuss their thinking?	X				
	Did the teacher treat displays as "right" or "wrong"?			X		
Unit Specific Mathematical Concepts	Was the order of a display talked about?				X	
	Was the scale of a display talked about?				X	
	Was the grouping in a display talked about?				X	
	Was the effect of design decisions on the shape of a display talked about?					X
	Did the teacher and/or students talk about what one or more displays show about the data?				X	
	Did the teacher and/or students talk about what one or more displays hide about the data?				X	