The Role of Design Experiments and Invariant Measurement Scales in the Development of Domain Theories

C. Victor Bunderson
Brigham Young University & The EduMetrics Institute

Van A. Newby
The EduMetrics Institute

OVERVIEW

In order to explain the concepts in the title, this paper refers to four other papers in this volume. It builds on the introduction to domain theory and validity-centered design in Bunderson(2002), and adds to that paper information about the need for and nature of principled design experiments and theory-based calibrations. These tools make it possible to construct, over time and over repeated cycles, the validity argument for a domain theory and its associated construct-linked scales. This article refers to the paper reporting work-in-progress on the development of a domain theory in English as a second language (ESL) by Diane Strong-Krause (2002). It provides background for that paper by discussing the need for theory-based calibrations in the ESL speaking domain. This need is great, because of the multi-semester nature of the domain, and the ever-incomplete nature of the sample available to the investigator in the environment of the Brigham Young University (BYU) English Language Center.

This paper also provides some context for the paper by Newby, Conner, Grant, and Bunderson (2002). Principled Design Experiments as described herein require several types of measurement scale invariance. IRT models in general offer invariance across different samples of people and sets of items. Beyond that, it has been argued that the Rasch model provides equal interval scales through its link to Additive Conjoint Measurement (ACM). The analysis in the Newby, et al. paper shows that the mathematical connection between the Rasch model and additive conjoint structures is stronger than had previously been proved.

Armed with this stronger proof, a simulation study was defined in the hopes of showing that the Rasch model could return known location parameters from unidimensional, equal interval scales, cutting through any independent multidimensional error, and thus providing the best basis for constructing construct-linked scales. Such error is due to the existence of different degrees of item uniqueness and sample error. Together, these two sources of error yield different slopes in the "a" parameter. The simulation study is reported in Pelton (2002). As is often the case in research, hopes are not always borne out. In Pelton's calibrations of simulation data where the true location parameters were known, the 2PL calibration program returned the true originating parameters and true order more closely than did the Rasch calibration program, especially when the distribution of students was shifted to the high end (as if they had progressed and learned). The Rasch calibration programs were found to be superior only in cases where there was minimal multidimensional error or

guessing. In this simulation study the effects of various sources of error, interacting with the features of existing calibration programs, apparently overwhelmed the mathematical ideal that shows additive conjoint structures and Rasch structures to be inseparable. It is still true that there is no easy road to the ideal of invariant, construct-linked measurement scales that span a domain.

PRINCIPLED DESIGN EXPERIMENTS

The idea behind a design experiment is that learning or instructional innovations may be incorporated into a theory-based design and implemented in a live setting over a series of continuing research cycles, with the teachers and students serving with educational researchers as research and design partners.

The roots of the idea of a design experiment in education can be traced back before the accelerating growth of interest in the previous decade (e.g., Dewey, 1916 cited in Tanner (1997); Glaser, (1976), and Reigeluth, Bunderson, & Merrill (1978)). Despite this longer history, most proponents writing today cite Allan Collins' 1990 paper: *Toward a Design Science of Education*, wherein he introduces the term design experiment, and Ann Brown's 1992 paper *Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings*. Both of these authors, as well as Glaser (1976), who described much of the idea of a design experiment, without using the term, use another term *Design Science*, and cite *Sciences of the Artificial* (Simon, 1969; 1981).

The term *design experiment* has been appropriated by those interested in intuitive and reflective types of action research, where the teacher takes the lead role, rather than a partnership role with researchers. The teacher journals personal reflections on possible designs and outcomes as a way to improve practice. Theory, measurement, and the validity of causal inferences are very important issues in research generalizable beyond single practitioners. For this reason, it is useful here to use the term *principled design experiment.* To be principled, the design must be theory based, with implications/predictions testable through the measures collected during the living experiment. In addition, the design must be well documented and its prescriptive guidelines well formulated for use in future designs ("principled design"; see Berkeley, (1999) and diSessa (1991)). Most importantly, a design experiment needs measurement scales that are fully comparable, even invariant, over the multiple cycles of implementation and research.

A design experiment takes place over multiple cycles (blocks of time, e.g., semesters for school, or quarters for the corporate world). The results of previous cycles serve as controls for the next cycle, reducing the need for control groups. Knowledge of each learner's starting position reduces the need for randomization. Transfer of the principal findings to other situated groups over time strengthens the argument for population representativeness when random sampling from a population is not possible. Thus, design experiments reduce the almost impossible problem of trying to meet the requirements of good experimental or quasi-experimental design where variables must be randomized, controlled and manipulated across groups of classrooms (McGee & Howard (1998)).

Design Experiments and Experimental Design.

It is useful to consider the nature of good experimental design to fully appreciate what can be accomplished with a principled design experiment using invariant measurement scales. Campbell and Stanley (1963) gave guidance much used over the years in how to reduce twelve common threats to the validity of causal inferences by employing good experimental design. While many others over subsequent years have expanded on their useful work, it remains a clear and cogent exposition of the central issues that will be used here to highlight how these issues are dealt with in a design experiment. They discussed three pre-experimental designs, three true experimental designs, and ten quasi-experimental designs. There has been a shift away from significance testing since their work was published in favor of effect sizes, but the need to consider threats to the validity of inference has not abated. The simplest inference is that the introduction of treatment *X* did indeed cause the increase in observation (measure) *O*. As discussed in Bunderson (2002), there is a need for design disciplines both to assure that treatment *X* does indeed involve construct-linked learning and that the features of the instruction are related to the construct to be improved. Instructional-design theories based on a domain theory related to both *X* and *O* can help give this design the six aspects of construct-validity. The domain theory with construct-linked scales of learning and growth, along with a validity argument, can assure that the *O* indeed measures valid levels of progress in the construct. Individual difference theories may also be involved in the design, in which case a pretest design is needed to identify groups of learners who would receive different treatments. In this case we are measuring an O_{id} in addition to the measure of progress on the domain construct, *O*. In order to measure change from before and after the administration of treatment *X*, we need a pretest-posttest design on the construct-linked outcome measure *O*.

For simplicity, we will not in this article discuss designs that require adaptation to individual difference measures O_{id}. Consider a pretest-posttest design, one of Campbell and Stanley's three true experimental designs. It can be diagrammed as follows, where R means random assignment to either group, and R O X O is a temporal sequence for the group:

R O X O experimental group
R O O control group

The gain score can be calculated as O postest – O pretest if the measurement scale *O* supports equal intervals. Then we may graph the gain for the experimental group and compare it to the gain for the control group.

Consider a series of such experiments, each with a well-documented change in the design specifications for treatment *X*, or perhaps only changes in the implementation procedures for the treatment, which are a part of the specification of *X*.

	Cycle 1	**Cycle 2**	**Cycle 3**	**Cycle 4**	**Cycle 5**
Baseline Measure O_0 Control for cycles 1-5.	X_1 O_0 Control for cycles 2-5	X_2 O_1 Control for cycles 3-5	X_3 O_2 Control for cycles 4-5	X_4 O_2 Control for cycle 5	X_5 O_3 Control for future cycles

Table 1. Repeating cycles of a principled design experiment using invariant scale *O*

The design experiment in Table 1 assumes that all repeating cycles of principled design intervention, X_1-X_5 take place in the same class in an educational institution (or group of similar classes using the same treatments and outcome measures, *X* and *O*). The designers of the experiment in Table 1 altered the treatment condition X_i each time, being careful to conform each design change to a prescriptive theory and assuring that each version of *X* was well documented. No *R* for random assignment is shown in this diagram, as it is neither possible nor necessary for random assignment to occur in these classes. Assume that both the registration procedures of the educational institution and ethical considerations bar such a practice. Thus we must depend on the sample-invariance and interpretive invariance properties of outcome measure *O*, and on repeated near-replications with other groups flowing as samples, presumably from the same or similar population to substitute for random assignment. Our construct-linked variable of interest, *O*, has the appropriate properties to reduce the risks to internal validity that randomization aimed to reduce.

Repeated measurements of subsequent groups over the cycles of a design experiment give further basis for ruling out the effects of a peculiar group of students one semester. But the students who flow through the classes of any one institution are unique and slanted in their own way, so population representativeness has not been obtained. For this function, randomization is a poor tool as well. To obtain evidence of generalizability to other groups, with other language, gender, racial, and special conditions requires design experiments to be set up in other locations wherein samples of students with subsets of these other characteristics abound. Nevertheless, it is no small benefit to causal inference to be assured that each semester's group was either equivalent on the highly interpretable pretest-post-test measure *O*, or of known deviation. It is of no small benefit to know that the use of a gain score is appropriate because we have achieved a close approximation to equal interval properties in our measurement scale(s) *O*. Finally, the quest for the validity of causal inference itself is, in the framework of a design experiment, set in a larger context similar to the quest for total quality management. After all, there are many aspects to treatment *X* and its procedures for implementation. To which specifically is the causal inference to be made? An outstanding result is a tribute to the entire group of people and the roles they assumed, the rules they followed, and the tools they used to administer the treatment. An outstanding result is evidence of the possibility that such results can be obtained by managing the implementation of the treatment well, and if in one group, why not in another?

In addition to substituting known starting positions on an invariant scale for random assignment, the design experiment in Table 1 also replaces the control group, using a series of comparisons to previous groups. Starting with the baseline measurement on O_0, each succeeding cycle can compare its outcome not only to this baseline measure, but also to each of the preceding outcomes, using invariant scales O_0 -- O_2. (The subscripts indicate that in this experiment, it was only necessary to modify the outcome measure twice by adding or deleting tasks, while during the same 5 cycles the treatment *X* was modified five times, after each cycle and before the next one). By the specific objectivity property of scale invariance, we can add or subtract questions or tasks from the most recent version of the instrument each semester. Then, after assuring that the versions are equated to the same meaningful scale each semester, we can make inferences that the version of *X* used that semester was likely the main cause of whatever improvement in gain score was observed.

Surely new guidelines to update good experimental design prescriptions must be developed for principled design experiments. Scientists writing in Campbell and Stanley's era could not have foreseen today's opportunities for unprecedented control of a complex interplay of many treatment variables possible in live settings through interactive technology. Neither could they have envisioned sophistication of on-line measurement, nor new measurement methods that can provide close approximations to the invariance properties needed to realize the design experiment scenario. Certain of Campbell and Stanley's quasi-experimental designs are similar to parts of this design experiment scenario, but scientists of their day could not have anticipated the full extent of reliable replication (using technology) with well-documented design changes. Advances in the technology of learning, measurement, and management of complexity changes how well we can control each of the twelve common threats to validity of concern to these earlier scientists. Moreover these advances both introduce new threats and help to find the means to reduce their effects.

It is interesting that the use of the term "validity" by experimental design writers and validity concepts from psychometrics, such as Messick's (1989,1995) unified validity concept, do not coincide. For example, the idea of "external validity" is used in entirely different ways. By bringing these two views together through validity-centered design and the methods of principled design experiments, perhaps we are taking a step toward bridging the gap between the "two disciplines of scientific psychology" decried by Cronbach (1965, 1975).

The "*J*-Curve of Implementation" and Design Experiments

Consider a possible graph of the results of the 5-cycle design experiment specified in Table 1. Figure 1 is a graph representing the sequence of outcomes. The Y-axis is the outcome measure *O*. The *X*-axis plots the subsequent gains due to treatments, starting with the baseline and for each of the 5 cycles. The bars represent gains. The case is depicted where not all groups started at the same average. Another graph could plot the *J*-curve of pure gain.

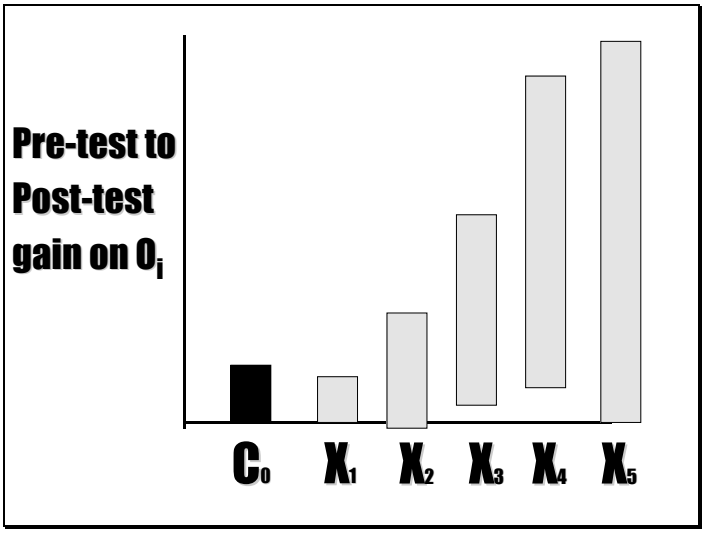

Figure 1. Graphing the results of a 5-cycle design experiment. Illustrating the "*J*-Curve of Implementation."

Sometimes, indeed, not infrequently, when complex interventions are introduced into live settings, there is a dip in the outcome measure below the baseline. There are many reasons for this, but most can be summarized under the need for the team of teachers, assistants, students, and administrators to "learn new tricks" and to make the new process gel. This kind of curve is called a "*J*-curve" in international exchange rate economics. We refer to it here as the "*J*-curve of implementation" because the effects of learning how to implement and manage the new system properly in a live setting speaks to the need for another disciplined design process. It could be called implementation design, change management design, or any of a number of titles. In experimental design, it would be equivalent to separating the specific treatment, *Xt*, from the management methods (*Xm*) for administering the treatment before, during, and after the introduction of the experimental treatment itself. *Xm* would include the method for administering, analyzing, and using outcome measure *O*. Even as there should be a disciplined Instructional Design Theory as a source of prescriptions for designing *Xt*, there should be an implementation management design theory for designing and conducting *Xm*. Figure 2, below, depicts this needed design theory.

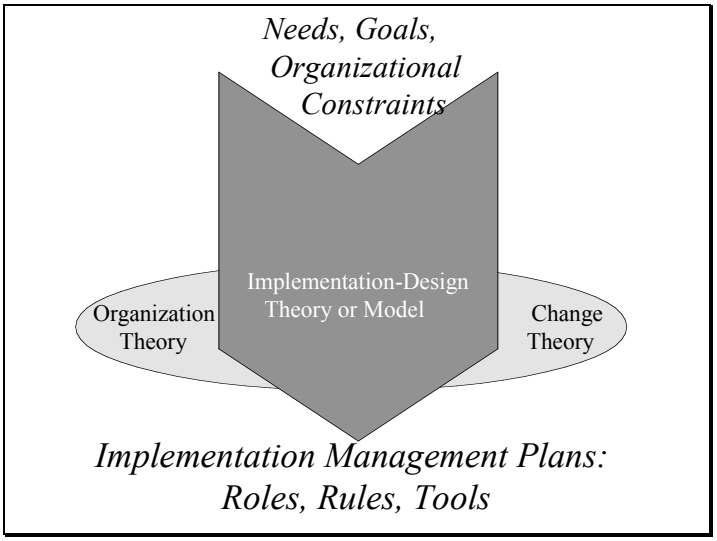

Figure 2. Disciplined design method or theory for implementation/change management

Figure 2 depicts only the implementation-design theory or model, using the symbol of a transforming arrow as with the instructional design theory. Relevant descriptive theory behind the prescriptions of such an implementation model or theory are organizational theory and change-management theory. Not depicted are the two needed descriptive theories shown in the other diagram, domain theory and individual differences theory. In addition, since the implementation model gives prescriptions for the design of implementation management aspects of treatment *X* (*Xm*), while the instructional design model gives prescriptions for the learning and instructional system aspects of treament *X* (*Xt*), both are needed and could be shown in a somewhat complex diagram.

Powerful effects due to implementation management. The effects of improvements in the implementation plan and design on the success of the entire intervention may be greater than the effects of the focal treatment change itself. An example will be given of a design

experiment in the introduction of a computer-delivered instructional program in Freshman English, including measures for both grammar and mechanics and good writing.

This example was seminal to the author both as an early instance of a design experiment, and as a compelling illustration of the importance of designing and executing a powerful implementation plan. This example is taken from implementation of the TICCIT system at BYU some years ago. TICCIT stands for "Time-Shared, Interactive, Computer-Controlled Information Television. It was one of the two largest NSF-funded Computer-Assisted Instruction projects of the 1970's. (PLATO was the other one.) The Mitre Corporation designed and developed the hardware and system software, and instructional designers/researchers at BYU developed and implemented the courseware. When the TICCIT computer system was completed and began to be used at BYU, the English Grammar and Composition course on TICCIT was used in several sections of English to prepare students to take the university-wide general education (GE) writing exam.

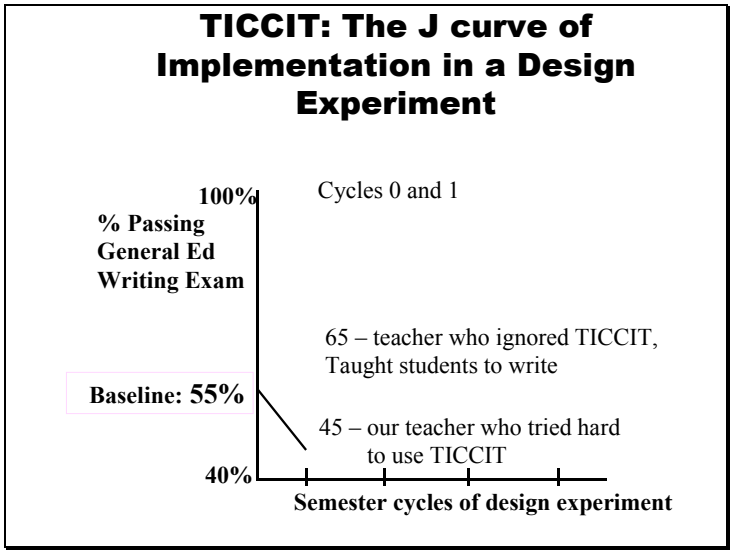

Figure 3. Results after one cycle of TICCIT design experiment.

All sections at BYU took the same test at the end of each semester, which consisted of a multiple choice grammar and mechanics test and a written essay graded by teachers. It is unusual for a university to develop but one standard measurement system for all sections in a particular domain, but this was one of those unusual times. The tests were scored by classical means and lacked desirable invariance properties, but were suitably comparable for the large effects noted. The cycle for this college class was one semester in length. The first semester two graduate student instructors taught two TICCIT sections. The first one tried hard to integrate the computer, which had a set of tutorial lessons in grammar and mechanics. It also had writing lessons dealing with audience, purpose, structure, outlining, etc. The second teacher did not believe in the computer, but was confident that she could teach students to write. The baseline and the results of the first cycle are illustrated in Figure 3.

After the first cycle it was found that the teacher who sincerely tried to use TICCIT had poor success. On average only 45% of her students passed the GE exam. The other teacher said to her students in effect: "Go use that computer we're supposed to use in what ever way you wish, but here in my class I'll teach you how to write". The result of her efforts:

65% of her students passed the GE exam. The baseline average was 55% across the university. After one cycle, the logical conclusion would be to teach the methods of the second teacher to teachers in the other sections, and throw out the computer system.

But the English faculty were not attentive to data, and none of them were interested in what was going on with TICCIT anyway, so we were spared. The team of developers and researchers, working with the first teacher, studied data available on the classes taught by both teachers. This group designed better implementation plans for the second semester, plans designed to help the students understand their roles better, and the teacher to redefine her role. No changes were made in the hardware or courseware, just in the implementation of new roles and new rules. Students were organized into small groups and scheduled to meet at computer terminal tables where they could see and talk to one another. At the end of the second cycle the TICCIT teacher's class did better than the 2nd teacher's previous semester, 72% compared to 65%, and substantially higher than baseline. The TICCIT team still believed their teacher, who was a colleague and participant in the design experiment, could do better. They believed also that the students, backed by professional developers and researchers, could do much better, so more changes in roles and rules were made. At the end of the third cycle 84% of that semester's group of students in the TICCIT section passed the GE exam. The students in these subsequent semesters were accepting their role to learn and practice grammar and mechanics on the computer, and the teacher was spending more time with small groups teaching them how to write. The team made further refinements and tried again. At the end of the 4th semester 93% of this new group of TICCIT students passed the two-part GE exam. See Figure 4.

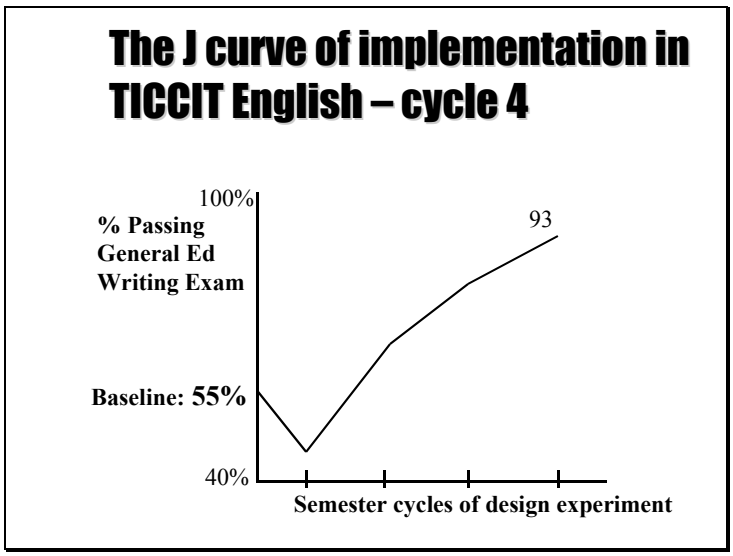

Figure 4. TICCIT design experiment results over four semesters.

The improvement in cycle 4 was attributed to the teacher beginning to use the reports generated by the computer system to identify the number of TICCIT lessons the students were completing. This implementation tactic was added to further refinements to the previous successful implementation methods for working with small groups. Using the reports, those not making good progress could be identified early and encouraged and taught individually,

while most of the class was involved practicing and studying on the computer or working on their writing.

This large effect was attributable to variables entirely from the implementation or change management area. There were no NSF funds left at that time to redesign the hardware, software, or courseware.

The Need and Hope for Invariant Measurement Scales

The case for principled design experiments presented above, and the nature of the implementation process and the *J*-curve of implementation both depend on measurement scales for outcome variables that have strong invariance properties. The TICCIT design experiment could succeed only because of the unusual adoption by the university (for a short time) of a common measurement system for all sections. The classically scored tests did not have these properties, but were comparable as roughly parallel tests. Because of the specific objectivity properties of the Rasch model, users of this approach are well aware that the first two of the four kinds of invariance listed below are needed and are available:

1. **Sample Invariance**: Measures do not depend on the sample of persons initially used to calibrate.
2. **Task Invariance**: Measures do not depend on the particular set of items (tasks or testlets) involved in the initial calibration, nor is the same set of items needed in any particular assessment of an individual.
3. **Unit Invariance**: Measures can be interpreted as having at least approximately equal intervals – the measurement scale supports the computation and interpretation of gain scores.
4. **Interpretive Invariance**: Measures are set in a coherent interpretive framework, and this framework remains constant from occasion to occasion (e.g., from cycle to cycle of a design experiment).

Interpretive Invariance. Surely the last form of invariance is the most important. It requires a strong validity argument and an excellent map or other display representing progress from lower to higher in each domain. This display must be fully understandable to learners, teachers, and others. In short, it requires all nine of the aspects of an excellent validity argument outlined in Bunderson (2002) elsewhere in this volume. User-centered design is necessary to make the visual representations – learning progress maps – accessible and useful to the users. These maps may also provide navigational utility in the on-line computer or web environment. Interpretations at different levels of the learning constructs for each separate scale in the domain-spanning maps rely on linkages to constructs in a domain theory.

Beyond the nine aspects of a validity argument, domain theories and their associated construct-linked scales can free capital, in the sense described by Fisher (2002) in this volume. Consider accomplishments of the brutal but effective Chinese emperor, Qin Shi Huang in 220 BC. Among other notable accomplishments, he established a common set of weights and measures, which galvanized great economic and scientific growth. In our era we can free capital in domains like education and health by establishing common measurement scales with a common framework of interpretation. To do this, a social process must occur in which the domain theory and its scales become adopted widely by many other users –

including many teachers and learners in the same domain. Some flexibility in the selection of tasks, and even scales, is appropriate, but a basic agreement must exist, and data must be shared at a central data center and analysis center that operates above reproach in attending to the interest of all users and maintaining a common standard. At this point, we will be able to achieve Interpretive Invariance, and may approach the other forms of invariance ever more closely as well.

Eva Baker(1998) speaks eloquently for the need to know clearly what should be taught and what should be measured:

> An important step toward a solution requires an understanding of the conceptual and scientific basis of student learning. Rather than continue to patch and accrete more and more incompatible solutions, we need a clear, fully articulated, descriptive system. A major scientific effort is needed to specify and goals, instructional requirements, and potential measured outcomes of learning.

Baker uses the metaphor of the human genome mapping project to argue for the importance and potential benefit of the two projects, "although the learning map is more difficult and requires some arbitrary definitions." The goal of domain theory is similar, but another goal of domain theory development is to find ways to make it easier and less costly through doing it incrementally in a variety of domains and settings, using systems that build in the capability for conducting principled design experiments. We would like to replicate another finding of the human genome project -- as the competing groups got into it, the rate of accomplishment and completion greatly accelerated over earlier expectations.

Further evidence that the Rasch Model can produce additive scales. This paper has provided a case for the need for a stochastic measurement theory that will produce stable equal-interval scales. Issues of concern in this paper provided motivation for the mathematical work reported by Newby, Conner, Grant, and Bunderson (2002). Principled Design Experiments as described herein require all four forms of scale invariance listed above, especially the crucial interpretive invariance. IRT models in general offer invariance across different samples of people and sets of items. Beyond that, it has been argued that the Rasch model provides equal interval scales through its link to Additive Conjoint Measurement (ACM). The mathematical analysis in the Newby, et al. paper shows that the mathematical connection between the Rasch model and additive conjoint structures is stronger than that of the Rasch model being merely a "special case" of additive conjoint measurement. There is a necessary connection between additive conjoint representations and Rasch representations, as defined in the mathematical paper. The analysis in Newby, et al. also provides a link to probabilistic representation. It gives a greater reason to make the claim that the Rasch model is, or could become, a way to provide a stochastic realization of ACM.

The analysis in Newby, et al. relates a mathematical model of ACM to a mathematical model of a Rasch representation. There are a variety of different axiom sets proposed for proving theorems of representation in different domains (e.g., Luce & Tukey (1964), Krantz, et al. (1971), Fishburn (1988); see also Newby, et al.'s explanation of Brogden's (1977) shortened list of axioms). The mathematical analysis in Newby, et al. does not require, or refer to, any axiom set. However, it does provide a link to probability not found in the deterministic and unbending structures of axiomatic representation theory of measurement. However, the

results found in Newby, et al. should not be misinterpreted as proving that if a particular set of data fits the Rasch model reasonably well, then that fit is sufficient to claim equal intervals. Michell (2002) in this volume makes a case that this has not been achieved, and that the best we can do with the Rasch model is still ordinal. (Cliff (1992) and McDonald (1999) have made a similar point in different ways.) Nevertheless, this and other mathematical analyses let us hold out the hope that further progress may be made in the human sciences in developing sufficiently invariant and sufficiently widely accepted scales to free the enormous scientific, educational, social, and economic capital that common measures can release.

Further evidence that the Rasch Model calibration programs cannot overcome independent multidimensionality. Independent multidimensionality is the combined effect of sampling error and item uniqueness. Uniqueness is manifest in IRT by different slope parameters on the logistic ogive functions that model the probability of correct response to an item. Unequal slopes violate the conditions for additive conjoint measurement. An early version of the mathematical analysis in Newby, et al. (2002) was influential in the initiation of the study by Pelton (2002). The proof shows that if you have an additive conjoint structure, you have a Rasch representation. Therefore, the Rasch model and calibration programs ought to lead more directly to the four types of invariance discussed above than 2PL or 3PL calibration programs. Wright (1999) has argued this point many times that crossing item characteristic curves will create inversions of the true order of item location parameters. A domain theory of learning and growth cannot achieve interpretive invariance if the order of tasks shift about as students became more mature. Pelton was seeking the best foundation for invariance of person and item parameters as a foundation for domain theories with invariant, construct-linked scales. The evidence from Pelton's study is that the Rasch calibration programs do indeed return the known originating parameter estimates best in the case where the conditions of ACM are most closely approximated: perfect unidimensionality and equal error distributions associated with each item. Unfortunately, the superiority over the 2PL model was very small and of no practical significance in these cases, and in all the cases where the data had different amounts of independent multidimensionality, the 2PL model calibration programs returned the known *location parameters* more accurately than the Rasch calibration programs. Also, contrary to the expectation that in the presence of crossing trace lines, the 2PL model would show more inversions of the true parameters than the Rasch programs, the reverse was found. The 2PL calibration program showed a lower mean-square error from the true locations -- especially when the distribution of students was shifted to the high end (as if they had progressed and learned). Implications of this type of result to the decision of how best to develop invariant, construct-linked scales of learning and growth are not yet clear.

Leaving these issues, let us turn to building a domain theory that could be accepted by a wide user community in a given domain, thus freeing the scientific, educational, and economic capital that comparable, easily interpretable domain scales could bring. This is the issue of interpretive invariance, and it could presumably be achieved with only an increasingly improving approximation to sample, task, and unit invariance. After all, length was initially measured by cubits, the length of a person's forearm and hand. Standardizing on the king's forearm didn't last, and it was centuries, millennia, before new measurement rods could be compared to a standard meter stick by visiting the standard meter in its controlled vaults in Paris.

Theory-based development requires a cooperating scientific and user community and cannot depend on local-sample, unique scores or scales

Diane Strong-Krause is well into a multi-year project in a design experiment format to develop a domain theory of English as a second language (ESL), starting with the speaking scale. In the first cycle of her design experiment (Strong-Krause, 2001), she examined a practical model of ordered tasks and a standardized oral proficiency interview. The American Council of Teachers of Foreign Language (ACTFL) developed the model and proficiency measurement system over a number of years. Scholars respected the codification of combined experience, but wanted more substantive process theory. Strong-Krause examined the process theories of leaders, including Bachman (1990). Based on these existing candidate pieces of a domain theory, one that emphasized tasks and one that emphasized processes, she used principled design methods consistent with Validity-Centered Design to develop tasks grouped into ten levels with four tasks each. In the paper in this volume (Strong-Krause (2002)), she reports on the success in creating three different sets of theoretical calibrations based on the judgements of experts coming from three perspectives. Each of these three judged the set of 40 tasks by estimating the ratings they would give to imagined students of given proficiency levels on a standard rubric. The three judges were:

- 1. A person expert in the ACTFL proficiency model,
- 2. A designer who had used principled design methods to produce the 40 tasks, and
- 3. An experienced teacher who had taught in the BYU English language center for several years.

The intent of these ratings was to perform an exploratory study to determine the reliability of predictions from these three different sources of human domain expertise. The results of the study are reported in Strong-Krause (2002). In this paper it is our task to clarify the reasons that such an anti-empiricist idea would even be considered.

Consider the problem of developing a construct-linked scale tied to a domain theory that could be seen to be of value to a broad community. Figure 5 is a representation of the population distribution as seen by a group that has had success in developing a set of standard descriptors of expertise levels, and a standardized oral proficiency interview.

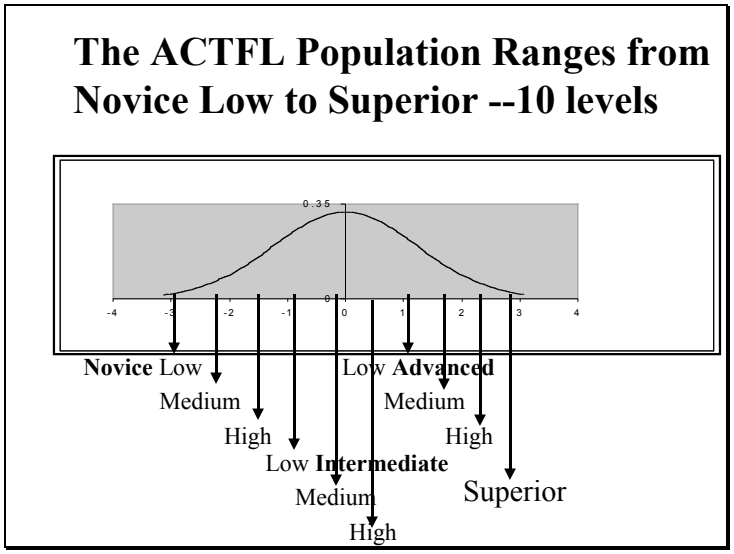

Figure 5. A theoretical population distribution and designated proficiency levels established by one national group.

Across the whole range of proficiency, labels have been established for four major categories: novice, intermediate, advanced, and superior. Each of the first three of these are broken down further into low, medium, and high, giving a total of ten categories. For a design experiment to strive toward a domain theory applicable to an entire population, the sample available at a given location may not be representative. At BYU, a large and successful English language center program, the sample is truncated at both ends, as seen in Figure 6.

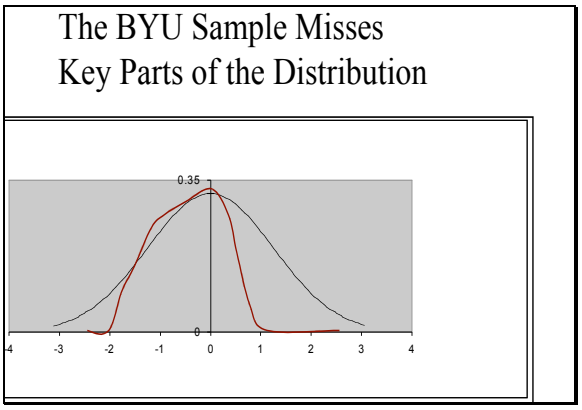

Figure 6. Even larger programs may not encompass the full population of persons, nor tasks in a domain.

The BYU program is fairly large, and has enrolled around 250 persons most years. These persons are ranked under headings that map to the ACTFL levels novice medium to intermediate medium. BYU does not admit those who would be classified as novice low, and graduate students to other college work or employment when they reach the intermediate low level, with some reaching intermediate medium. To develop construct-linked domain scales for the whole range would involve collaboration with other programs that have people and that require tasks at the lower and higher levels. One of the disciplines of construct-linked scale development is to consider the definition of person and task alpha and person and task omega. What is the easiest task that fits on a given scale that could just barely be passed by the most minimally prepared person that could enter the program to work on the domain of expertise? What is the hardest task and the person who could just pass it? As Messick (1995) has stated, "domain theory is a primary basis for specifying the *boundaries* and structure of the construct to be assessed".

Finding the boundaries and having access to the entire range of persons and tasks is not the only problem in developing a domain theory and associated construct-linked scales. Even if we could assemble all the people in one study, develop the entire range of tasks to span the domain scales from easiest to hardest, how could we ever collect the data? Not one institution could be found that has the complete range of people or of tasks. Even if they did, consider the BYU English language center, whose program spans three semesters. Semester 1 people would choke at the hard tasks. Semester 3 people would be bored and insulted by the

easy tasks. It would be impossible to assemble the people, to give all the tasks in a reasonable time, or to assure the quality of the responses by a lot of annoyed people.

Theory-based calibrations can have a high correlation with later calibrations using real data. Judges who have worked with tasks, like those found in the ACTFL proficiency interview, and with students of different levels of learning can make predictions about which level of persons will score at which scores in a rubric on a set of real tasks. These estimates can be averaged over a suitable number of judges, the distribution of simulees can be expanded to match the hypothesized population distribution, and a large data matrix can be generated with suitable models for error. Later, real tasks can be calibrated and linked into the same scale, so that person estimates during learning can be obtained from their responses to these tasks. The tasks share the interpretive framework of the domain theory, and thus have the property of interpretive invariance.

In Strong-Krause's (2002) results, the correlations between theory-generated calibrations and actual calibrations were around .80, the value of a respectable reliability coefficient and a remarkable predictive validity coefficient. With further work within the validity-centered design framework will it be possible to obtain theory-based calibrations from multiple judges that correlate around .90 with actual calibrations? We are working to find out.

SUMMARY

Design Experiments. Compared to experimental design, design experiments fit into naturally occurring environments and are more feasible and authentic than experimental designs in contrived settings. Furthermore, if we can consider our measurement scales suitably invariant, then comparable measures over repeated cycles replace the main control group, because knowledge of actual starting points replaces the assurance that groups *X* and *C* are comparable on the outcome variable due to randomization. It is still necessary to seek evidence of generalization to other groups with other mixes of students and tasks. This evidence is part of a good validity argument, but it must be obtained from evidence gathered from other classes, schools, and businesses who also set up design experiments in their locations. To do this, the complexity of design experiments must be buried in the technology used to transport the learning and measurement systems to other locations.

Using Validity-Centered Design and design experiments, we can test predictions of domain theories. Domain theory predictions about task locations can be made by domain experts using actual tasks, and these predictions can be confirmed or disconfirmed. The same is true about predictions of the locations of hypothesized (cognitive or other) processes presumed to be used by people at different levels of learning and growth. Prescriptions of instructional-design and implementation design theories can be tested and confirmed or disconfirmed. Over cycles, a strong validity argument can be built for both measurement instruments and theories, or changes can be made to strengthen the argument for revised theories and systems.

On theory-based calibrations. Building the construct-linked scales of complete domain theories cannot be accomplished with empirical data alone, but theory-based calibrations are feasible and can be reliable. They can form the basis for freeing the scientific, educational,

social, and economic capital by having "a common system of weights and measures" in each of the key domains of human learning and growth. Like measurement in other human domains where enormous efforts are expended at assuring exactness and standardization, domain theory-based measurement scales can improve over the years. The goal of the line of work summarized in the papers referred to herein is to show that the development of such domain scales can be accomplished a step at a time. It is doable, starting from small things like the experience of a group of teachers over some years with a flow of students and a set of tasks. It can then evolve over the cycles of design experiments, increasingly in multiple sites, into scales having widely accepted interpretive invariance, and increasingly, the other kinds of scale invariance. In likening the development of common standards for domain knowledge in important educational domains to "a human genome project", Baker (1997) used a compelling and apt metaphor for the social importance of such an enterprise – the capital it could release – but we are trying to develop design methods that will make it not as difficult and costly as Baker implies.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Oxford University.

Baker, E. L. (1997). *Understanding educational quality: Where validity meets technology*, Fifth annual William H. Angoff memorial lecture, Educational Testing Service Policy Information Center, Princeton, N.J., 08541-001.

Berkeley (1999). *Field Guide to Design Experiments* http://www.soe.berkeley.edu/sandhtdocs/guide.html

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika, 42(4),* 631-634.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, 2, 2, 141-178.

Bunderson, C. V. (2002). How to build a domain theory: On the validity centered design of construct-linked scales of learning and growth, in review at *Proceedings of the IOMW*.

Campbell, D. T., & Stanley, J. C. (1971). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research.* Skokie, IL: Rand McNally.

Cliff, N. (1989). Ordinal consistency and ordinal true scores. *Psychometrika* 54(1): 75-91.

_____ (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ, Lawrence Erlbaum Associates.

Collins, A. (1990). Toward a design science of education *[Technical Report #1]; Cambridge, MA: Bolt Beranek and Newman*. Also found in Scanlon, E.,& O'Shea, T. (Eds.). (1992). New directions in educational technology. New York: Springer-Verlag.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671-684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116-127.

Dewey, J. (1916). *Democracy and education*; New York: Free Press.

diSessa, A. A. (1991). Local sciences: Viewing the design of human-computer systems as cognitive science. In J. M. Carroll (Ed.), *Designing Interaction: Psychology at the Human-Computer Interface,* (pp. 162-202). Cambridge, England: Cambridge University Press

Embretson, S. E. & Reise, S. P. (2000*). Item response theory for psychologists* (multivariate applications book series), Lawrence Erlbaum Association.

Fishburn, P.C. (1988). *Nonlinear preference and utility theory*, Johns Hopkins University Press, MD.

Fisher, W. (2002). The benefits of the Rasch focus on Fundamental Measurement, , in review at *Proceedings of the IOMW*.

Glaser, R. (1976). Components of a psychology of instruction: Toward a science of design. *Review of Educational Research*, 46(1), 1-24.

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Volume 1: Additive and polynomial representations.* Academic Press, NY.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1,* 1-27.

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology **34**: 100-117.

_____ (1999). Test theory: a unified treatment. Mahwah, NJ, Lawrence Erlbaum Associates.

McGee, S. & Howard, B. (1998). Evaluating educational multimedia in the context of use. *Journal of Universal Computer Science*, 4(3), 273-291.

Messick, S. (1989). Validity, in R. L. Linn, ed., *Educational Measurement* (pages 13-103), New York: Macmillan.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, 50(9), 741-49.

Michell, J. (2002). The Rasch model: necessary (perhaps) but not sufficient for scientific measurement, in review at *Proceedings of the IOMW*.

Michell, J. (1999). Measurement in psychology: a critical history of a methodological concept. Cambridge, Cambridge University Press.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2002). The Rasch model & additive conjoint measurement, in review at *Proceedings of the IOMW*.

Pelton, T. (2002). What are the limits of the Rasch advantage?, in review at *Proceedings of the IOMW*.

Reigeluth, C.M., Bunderson, C. V., & Merrill, M.D., (1978). Is there a design science of instruction? *J. Instructional Development*, 1(2), 11-16.

Simon, H. (1969). *The Sciences of the Artificial* (also (1981) 2nd Edition with additional chapters) MIT Press, Cambridge MA.

Strong-Krause, D. (2002). Toward a domain theory in English as a second language, in review at *Proceedings of the IOMW*.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality, *Psychometrika*, 52, 589-617.

Tanner, L. N. (1997). *Dewey's laboratory school: lessons for today*; New York: Teachers College Press.

Wright, B.D., (1999) Fundamental measurement for psychology, in Embretson, S. E., and Hershberger, S. L., *The new rules of measurement*, Lawrence Erlbaum Associates, Inc., Mahwah, N.J.