Theory Referenced Measurement: Combining Substantive Theory and the Rasch Model

A. Jackson Stenner

Abstract A construct theory is the story we tell about what it means to move up and down the scale for a variable of interest (e.g., temperature, reading ability, short term memory). Why is it, for example, that items are ordered as they are on the item map? The story evolves as knowledge regarding the construct increases. We call both the process and the product of this evolutionary unfolding "construct definition" (Stenner et al., Journal of Educational Measurement 20:305–316, 1983). Advanced stages of construct definition are characterized by calibration equations (or specification equations) that operationalize and formalize a construct theory. These equations, make point predictions about item behavior or item ensemble distributions. The more closely theoretical calibrations coincide with empirical item difficulties, the more useful the construct theory and the more interesting the story. Twenty-five years of experience in developing the Lexile Framework for Reading enable us to distinguish five stages of thinking. Each subsequent stage can be characterized by an increasingly sophisticated use of substantive theory. Evidence that a construct theory and its associated technologies have reached a given stage or level can be found in the artifacts, instruments, and social networks that are realized at each level.

1 Level 1

At this stage there is no explicit theory as to why items are ordered as they are on the item map. Data are used to estimate both person measures and item difficulties. Just as with other actuarial sciences, empirically determined probabilities are of paramount importance. When data are found to fit a Rasch Model, relative differences among persons are independent of which items or occasions of measurement are used to make the measures. Location indeterminacy abounds: Each instrument/scale pairing for a specified construct has a uniquely determined “zero.”. At Level 1, instruments

Paper Presented at the Pacific Coast Research Conference. 2002.

A. J. Stenner (✉)
MetaMetrics, Inc., Durham, NC, USA

W. P. Fisher and P. J. Massengill (eds.), *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement*,
https://doi.org/10.1007/978-981-19-3747-7_9

don't share a common "zero" i.e., location parameter. A familiar artifact of this stage is the scale annotated with empirical item difficulties (Artifact 1). Most educational and psychological instruments in use today are Level 1 technologies.

2 Level 2

A construct theory can be formalized in a specification equation used to explain variation in item difficulties. If what causes variation in item difficulties can be reduced to an equation, then a vital piece of construct validity evidence has been secured. We argue elsewhere that the single most compelling piece of evidence for an instrument's construct validity is a specification equation that can account for a high proportion of observed variance in item difficulties (Stenner et al., 1983). Without such evidence only very weak correlational evidence can be marshaled for claims that "we know what we are measuring" and "we know how to build an indefinitely large number of theoretically parallel instruments that measure the same construct with the same precision of measurement".

Note that the causal status of a specification is tested by experimentally manipulating the variables in the specification equation and checking to see if the expected changes in item difficulty are, in fact, observed. Stone (2002) performed just such an experimental confirmation of the specification equation for the Knox Cube Test—Revised (KCT_R), when he designed new items to fill in holes in the item map and found the theoretical predictions coincided closely with observed difficulties. Can we imagine a more convincing demonstration that we know what we are measuring than when the construct theory and its associated specification equation accord well with experiments (Stenner & Smith, 1982; Stenner & Stone, 2003)?

Similar demonstrations have now been realized for hearing vocabulary (Stenner et al., 1983), reading (Stenner & Wright, 2002), quantitative reasoning (Enright and Sheehan, 2002), and abstract reasoning (Embretson, 2002). Artifacts that signal Level 2 use of theory are specification equations, RMSE's from regressions of observed item difficulties on theory, and evidence for causal status based on experimental manipulation of item design features (Artifact 2).

3 Level 3

The next stage in the evolving use of theory involves application of the specification equation to enrich scale annotations. We move beyond using empirical item difficulties as annotations. One example of this use of the specification equation is in the measurement of text readability in the Lexile Framework for Reading. In this application, a book or magazine article is conceptualized as a test made up of as many "imagined" items as there are paragraphs in the book. The specification equation is

then used to generate theoretical calibrations for each paragraph which then stand in for empirical item difficulties (Stone et al., 1999).

For instance, the text measure for a book is the Lexile reader measure needed to produce a sum of the modeled probabilities of correct answers over paragraphs, qua items, equal to a relative raw score of 75%. We can imagine a thought experiment in which every paragraph (say 900) in a Harry Potter novel is turned into a reading test item. Each of the 900 items is then administered to 1000 targeted readers and empirical item difficulties are computed from a hugely complex connected data collection effort. The text measure for the Harry Potter novel (880L) is the amount of reading ability needed to get a raw score of 675/900 items correct, or a relative raw score of 75% (Artifact 3).

The specification equation is used in place of the tremendously complicated and expensive realization of the thought experiment for every book we want to measure. The machinery described above can also be applied to text collections (book bags or briefcases) to enable scale annotation with real world text demands (college, workplace, etc.).

Artifacts of a Level 3 use of theory include construct maps (Artifact 3) that annotate the reading scale with texts that, thanks to theory, can be imagined to be tests with theoretically derived item calibrations.

4 Level 4

In biochemistry, when a substance is successfully synthesized using amino acids and other building blocks, the structure of the purified entity is then commonly considered to be understood. That is, when the action of a natural substance can be matched by that of a synthetic counterpart, we argue that we understand the structure of the natural substance. Analogously, we argue that when a clone for an instrument can be built and the clone produces measures indistinguishable from those produced by the original instrument, then we can claim we understand the construct under study. What is unambiguously cumulative in the history of science is not data text or theory but is rather the gradual refinement of instrumentation (Ackerman, 1985).

In a Level 4 use of construct theory there is enough confidence in the theory and associated specification equation that a theoretical calibration takes the place of an empirical item difficulty for every item in the instrument or item bank. There are now numerous reading tests (e.g., Scholastic Reading Inventory- Interactive and the Pearson PA Series Reading Test) that use only theoretical calibrations. Evidence abounds that the reader measures produced by these theoretically calibrated instruments are indistinguishable from measures made using the more familiar empirically scaled instruments (Artifact 4). At Level 4, instruments developed by different laboratories and corporations share a common scale. The number of unique metrics for measuring the same construct (e.g., reading ability) diminishes.

5 Level 5

Level 5 use of theory builds on Level 4 to handle the case in which theory provides not individual item calibrations but rather a distribution of "potential" item calibrations. Again, the Lexile Framework has been used to build reading tests incorporating this more advanced use of theory. Imagine a Time magazine article that is 1500 words in length. Imagine a software program that can generate a large number of "cloze" items (see Artifact 5) for this article. A sample from this collection is served up to the reader when she chooses to read this article. As she reads, she chooses words to fill in the blanks (missing words) distributed throughout the article. How can counts correct from such and experience produce Lexile reader measures, when it is impossible to effect a one-to-one correspondence between a reader response and an item and at theoretical calibration, specific to that particular item? The answer is that the theory provides a distribution of possible item calibrations (specifically, a mean and standard deviation), and a particular count correct is converted into a Lexile reader measure by integrating over the theoretical distribution (Artifact 6).

6 In Conclusion

"There is nothing so practical as a good theory." Kurt Lewin.

References

Ackerman, R. J. (1985). *Data, Instruments, and Theory*. Princeton.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380–396.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: implications for item generation. In Irvine S. H., & Kyllonen P. C. (Eds) *Item Generation for Test Development*. Hillsdale, NJ. Lawrence Erlbaum Associates.

Stenner, A. J., Burdick, H., Sanford, E., & Burdick, D. How accurate are Lexile text measures? Manuscript accepted. *Journal of Applied Measurement*.

Stenner, A. J., & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills, 55*, 415–426.

Stenner, A. J., Smith, M., & Burdick, D. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20*(4), 305–316.

Stenner, A. J., & Stone, M. H. (2003). Item specifications vs. item banking. transactions of the rasch SIG; *17*(3) 929–930.

Stenner, A. J.,& Wright, B. D. (2002) Readability, reading ability, and comprehension. Paper presented at the association of test publishers hall of fame induction for Benjamin Wright D., San Diego. In Wright, B. D., & Stone, M. H. (2004). Making Measures. (Chicago: Phaneron Press.

Stone, M. H. (2002) Knox Cube Test - revised. Itasca, IL. Stoelting.

Stone, M. H., Wright, B. D., & Stenner, A. J. (1999). Mapping variables. *Journal of Outcome Measurement, 3*(4), 308–322.