Realizing a Rasch measurement through instructionally-sequenced domains of test items.

E Matthew Schulz[1]
561 Junipero Ave, Pacific Grove, CA, 93950, USA

E-mail: mattschulz54@gmail.com

Abstract. This paper presents results from a project in which instructionally-sequenced domains were defined for purposes of constructing measures that that conform to an ideal in Guttman scaling and Rasch measurement. A fundamental idea in these measurement systems is that every person higher on the measurement scale can do everything that lower-level persons can do, plus at least one more thing. This idea has had limited application in educational measurement due to the stochastic nature of item response data and the sheer number of items needed to obtain reliable measures. However, it has been shown by Schulz, Lee, and Mullen [1} that this ideal can be can be realized at a higher level of abstraction -- when items within a content strand are aggregated into a small number of domains that are ordered in instructional timing and difficulty. The present paper shows how this was done, and the results, in an achievement level setting project for the 2007 Grade 12 NAEP Economics Assessment.

1. Conceptual Background

Cliff remarked that a Guttman scale is one of the best examples of a good idea in all of psychometric measurement [1]. A Guttman scale typically consists of a relatively small number of ordered tasks or observational trials where each task represents a level of proficiency [2]. Persons in the population to which the scale applies can generally be expected to have mastery of tasks up to and including their assigned level, and non-mastery of tasks representing higher levels. The universal ordering of task difficulty for all persons allows one to predict a person's mastery of each of the levels defining the scale, solely from the person's assigned level. Moreover, from observing a person's performance on just one level, one can predict performance on any lower level (if the performance was successful) or on any higher level (if the performance was not successful).

Using an example of a four-item test, Andrich [3] showed that the Rasch model [4] is a probabilistic version of a Guttman scale. In the Rasch model, Guttman levels correspond to points on the latent proficiency scale, and there are as many levels as there are binary items or rating scale categories in the assessment (as in traditional Guttman scaling). The Rasch model provides a probability for success as a function of the difference between the measure of examinee ability and the measure of task, or level, difficulty. Mastery of a level can be viewed as a matter of degree, quantified by the probability of success, rather than as an all-or-none phenomenon. Importantly, however, Andrich showed that in order for items to be Guttman-scalable their item characteristic curves on the latent proficiency scale must not cross. That is, the ordering of levels must be the same at all levels of ability and for all probabilities of success.

[1] To whom any correspondence should be addressed.

Schulz, Kolen & Nicewander [5] noted two characteristic features of educational tests, besides the stochastic nature of test items pointed out by Andrich [3], that has made it difficult to put Guttman scaling ideals into practice in education. One is the sheer number of items on educational tests. Guttman scales typically consist of 4 to 7 levels, each represented by a single task that can be observed and scored with near-perfect reliability. The distance between Guttman levels on a Rasch scale would be very large and there would be little chance that their characteristic curves would overlap. Large numbers of items are required for educational testing because their reliability is so low. Items are so closely spaced on the measurement scale that their characteristic curves do tend to cross. The sheer number of crossings, however, discourages educators from attaching a great deal of meaning to this particular violation of a measurement ideal.

The other way that educational test items differ from Guttman tasks is in that they are universally regarded as exchangeable, random sampling units of larger domains. A general area of skill, such as mathematics, or even a specific skill such as working with fractions, may be represented by hundreds, if not thousands, of test items. Virtually any skill that is a target of assessment and general inference is assessable with multiple, exchangeable test items, at least in theory. The tasks or observations comprising a Guttman tasks are typically treated as essential components of the scale. It is not common practice to exchange one task for another or for there to be two separate versions, or forms, of a Guttman assessment, each consisting of a different set of tasks measuring the same thing.

Recognizing test items as exchangeable sampling units of a broader domain of skill, Schulz, et al., [5] argued that item parameters such as the difficulty and discrimination parameters in item-response theory (IRT) models should be treated as random variables underlying similar parameters of domains. In subsequent studies, Schulz, et al., [6, 7, 8, 9] developed and applied a technique for defining a relatively small number of instructionally-relevant, difficulty-ordered domains within the broader domains of educational achievement tests including the National Assessments of Educational Progress (NAEP) in Grades 8 and 12 mathematics. Expected percent correct scores on the domains supported the notion that the domains had the same order of difficulty for all persons and that higher-level persons could do what lower-level persons could do, plus at least one more thing were realized at the level of percent correct scores on domains. Item parameters were considered only to the extent that they were expected to be normally distributed around the domain parameters with which they were associated and outlier status in this regard could cause items to lose their association with a domain.

In this paper, the domain definition process developed by Schulz et al., [7] is illustrated in a subject area, economics, that is not generally regarded as containing naturally-ordered progressions of skill. In mathematics, skills such as addition, subtraction, multiplication, and division seem to have a natural ordered in difficulty. In a subject such as economics, the prospect of defining difficulty-ordered domains based on instructional sequence is less clear. The domain development work in economics conducted by ACT for the NAEP economics standard setting project [10] has not been reported in detail previously except in technical reports and presentations delivered by the contractor to its technical advisory committee and to the committee on standards, design and methodology of the National Assessment Governing Board (NAGB).

2. Application

The achievement level setting project for the 2007 NAEP in grade 12 economics included a process of defining difficulty-ordered domains within the grade 12 economics assessment. The domains were intended to support the same Guttman-scale relationships as domains previously developed using items in the Grade 8 and Grade 12 NAEP mathematics assessments [6,7,8,9]. A key step in the domain-development process used in those studies was is to quantify the instructional timing of the test items. In a mathematics assessment covering primary grades, or even high school mathematics, instructional timing can be associated with grade levels or courses in a standard high school sequence. In economics, the rating scale shown in Figure 1 was used. Five curriculum and content experts used the rating scale to rate the NAEP Grade 12 economics items. An average rating was computed for each item.

The content experts then used the average ratings of items to organize the items into a series of "teacher" domains that they expected to be coherent, instructionally useful, and ordered ininstructional timing. Item-response-theory (IRT) estimates of item difficulty were available, but were not used at this stage of domain definition. The researchers leading the project felt it was felt important to define domains primarily on the basis of instructional timing and content and to consider less substantive characteristics of the items, such as difficulty only later.

Item Rating Scale	
Rating	Meaning
5	The knowledge, skills, and abilities (KSAs) required to get full credit on this item are usually mastered after the vast majority of other KSAs in an Economics curriculum have been mastered, the goal of which is mastery of all the benchmarks in the NAEP framework.
4	Mastery of the KSAs required to get full credit on this item typically follow mastery of the majority of other KSAs in Economics.
3	The KSAs required to get full credit on this item are mastered about midway through mastery of all the KSAs in Economics.
2	The KSAs required to get full credit on this item typically require mastery of some earlier KSAs in Economics.
1	Most of the KSAs needed to get full credit on this item are mastered early in a learning sequence in Economics.
DOES NOT APPLY	The KSAs required by this item are mastered in an Economics instruction in no particular sequence in relation to other KSAs. They may occur early in some Economics curricula and late in others.

Figure 1. Rating scale for quantifying the instructional timing of the knowledge, skill, and/or ability (KSAs) needed to correctly answer NAEP grade 12 economics assessment items.

Domains were expected to exhibit a positive, but not necessarily perfect, correlation between instructional timing and difficulty. Item difficulty was expected to be less variable within than across domains, but substantial overlap in item difficulty across domains was also expected. Figure 2 shows the dispersion of IRT difficulty parameters within and across domains that were ultimately defined within one of the three main strands of the NAEP Grade 12 economics assessment, National Economy. The other two strands are Market Economy and International Economy. The domains are ordered from left to right by the average instructional timing of their items. The correlation between mean instructional timing and mean item difficulty among the domains met expectations. The variability of item difficulty within and across domains also met expectations.

Figure 3 shows percent correct curves of the ten teacher domains defined within the National Economy strand. It can be concluded from this plot that the number of domains is too large to exhibit the desired Guttman and Rasch patterns of widely-spaced, non-crossing percent correct curves. The emphasis at this stage, however, was on defining domains that were meaningful and useful to teachers (teacher domains) as evidence by the ability of teachers to classify the items reliably into the domains using only a brief narrative description and approximately three exemplar items for each domain. The study did in fact find high agreement among teachers independently classifying items into the domains. To some extent, the degree of crossing and variability in the slopes of the percent correct curves may be due to the small number of items and differences in item types comprising the domains at this stage. A teacher domain might consist of as few as three items due to the limited total number of items available and the breadth of the assessment.

The four circles superimposed on the plot in Figure 3 shows how the teacher domains were combined to form a smaller set of domains for supplying relatively stable, Guttman-style descriptions of growth and differences in achievement. Teacher domains were combined through primarily through consideration of similarities among the domains in instructional timing and content.

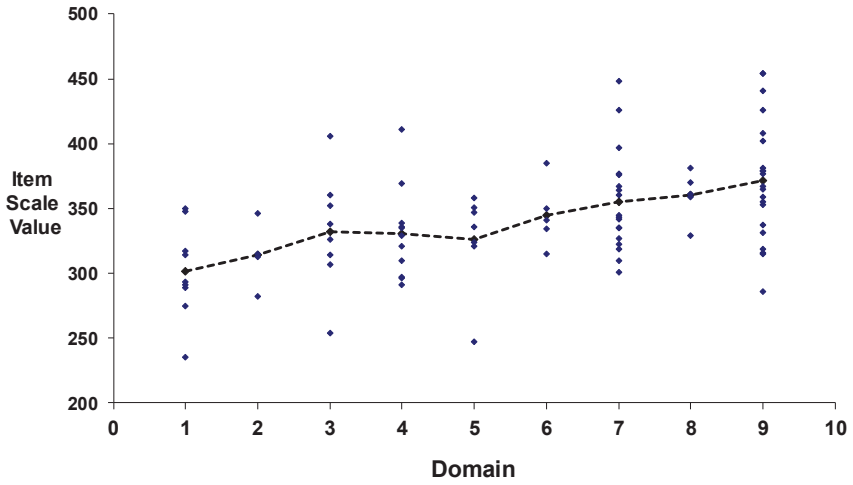

Figure 2. Item scale values within teacher domains defined for the National Economy strand of the 2007 Grade 12 NAEP Economics Assessment.

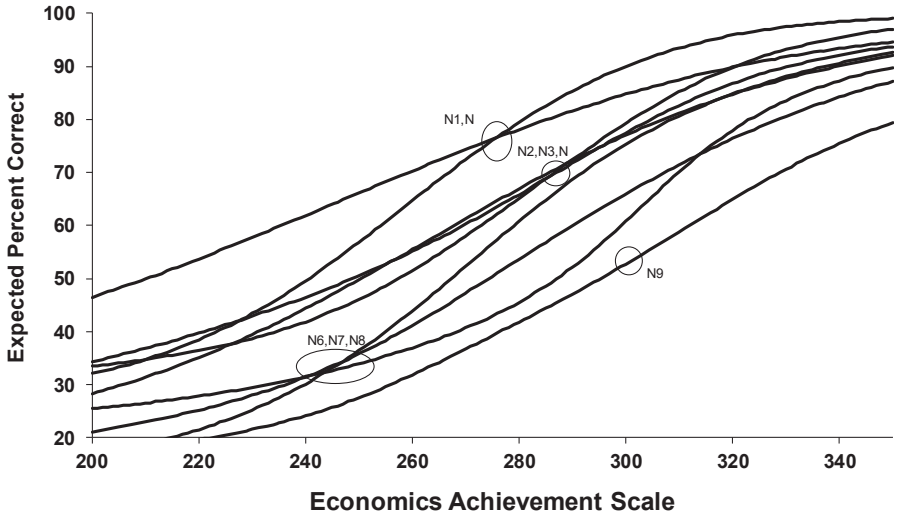

Figure 3. Percent correct on teacher domains within the National Economy substrand of the 2007 NAEP Grade 12 Economics assessment as a function of achievement.

Figure 4 shows that the resulting 'reporting' domains were relatively widely-spaced on the achievement scale and that their percent correct curves do not cross. Importantly, the domains will be mastered in the same order not matter what percent correct criterion for mastery is used. This pattern was felt to be stable enough for reporting and descriptive purposes. The expected stability of the mastery pattern is due to the spacing between the curves and to the larger (relative to the teacher domain) number of items comprising each domain.

Finally, Figure 5 shows how the teacher domains were used in an achievement level setting pilot study. Percent correct scores conditional on the lower boundaries of the cut scores being considered

for the achievement levels helped the panellists understand what students at the cut scores could and could not do and what students at a higher achievement level could do that students at a lower achievement level could not do. The brief titles representing the teacher domains in Figure 5 may convey to the reader a sense of growing complexity and skill with as students move up the scale. In addition to these titles, the panellists read brief narrative descriptions of the domains and studied the content of items within the domains. Process evaluations during the course and at the conclusion of the workshop showed that the panellists felt the domain-information was very helpful in the standard setting process.

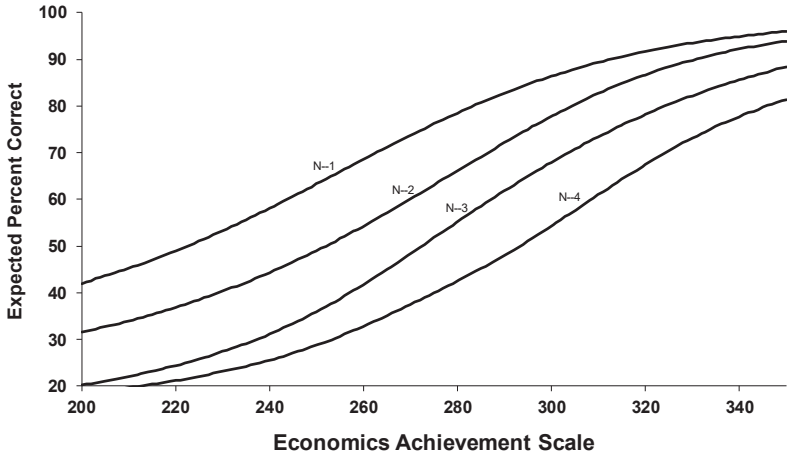

Figure 4. Percent correct on reporting domains in the National Economy substrand of the 2007 NAEP Grade 12 Economics assessment as a function of achievement.

Content Area	Domain	Expected Percent Correct at Lower Borderline of…		
		Basic	Proficient	Advanced
National	N1. Money, Loans, and Interest Rates	56%	74%	91%
	N2. Spending, Income, and Related National Measures	41%	71%	97%
	N3. Resource Allocation	42%	59%	89%
	N4. Economics Growth and Productivity	38%	60%	87%
	N5. Government Programs and Taxes	38%	57%	92%
	N6. Real Interest Rates	24%	51%	87%
	N7. Inflation and Unemployment	27%	46%	79%
	N8. Money Supply	29%	40%	82%
	N9. Fiscal and Monetary Policy	21%	36%	68%

Figure 5. Domain percent correct scores at lower boundaries of achievement levels in a pilot study for 2007 NAEP Economics achievement-level setting.

3. Educational Significance

The results of this, and previous work by Schulz, et al. adds something to the currently hot topic of learning progressions, where growth in achievement is seen as a sequential mastery of skills. Mastery in educational achievement testing is typically based on an arbitrary criterion percentage correct score, sometimes by policy and sometimes established through standard setting workshops. In Figure 5, the percent correct scores highlighted in yellow are above the 67% criterion for mastery used in the NAEP grade 12 economics standard setting workshop. If the order in which skills are mastered (and by implication the order in which skills should be taught) is to be independent of an arbitrarily-set mastery criterion, percent correct curves representing performance on the skills in the sequence must not cross. If descriptions of growth in student achievement can become independent of arbitrary mastery criteria and at the same time capture the simplicity and inferential power of Guttman scales, teachers, parents and policy makers will place more confidence in the results and implications of educational achievement tests for curriculum and instruction.

4. References

[1] Cliff, N. (1983). Evaluating Guttman Scales: Some old and new thoughts. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederick M. Lord*. Hillside, NJ.

[2] Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. A. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction* (pp 60-90). Princeton: Princeton University Press.

[3] Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Tuma (Ed.), Sociological Methodology (pp. 33-80). Jossey-Bass.

[4] Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research. Chicago: The University of Chicago Press. (Original work published 1960).

[5] Schulz, E. M., Kolen, M., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23* (4), 347-362.

[6] ACT, Inc. (September, 2003). *Defining Multiple Content Domains for Describing Achievement Levels on the NAEP Grade 8 Mathematics Assessment. A preliminary report to the National Center for Education Statistics*. Iowa City, iA: Author.

[7] Schulz, E. M., Lee, W., & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement*. *42*, 1-26.

[8] ACT, Inc. (April, 2005). Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Domain Development Report. Iowa City, IA: Author.

[9] ACT, Inc. (April, 2005*). Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Process Report.* Iowa City, IA: Author.

[10] ACT, Inc. (April, 2007). *Developing achievement levels on the 2007 national assessment of educational progress in grade twelve economics: Special Studies report.* Iowa City, IA: Author.