The things taught in schools and colleges are not an education but the means of an education. Ralph Waldo Emerson.

There is no such thing as measurement absolute; there is only measurement relative. Jeanette Winterson.

We dream about measuring cognitive status so effectively that we can monitor progress over the student's career as confidently as we monitor changes in height, weight, and time for the 100-meters. We're not there yet but we aren't where we were. Partly because of Rasch. Celsius and Fahrenheit probably did not decide in their youth that their mission in life was to build thermometers; when they wanted to understand something about heat, they needed measures to do that. Educators don't do assessment because they want to build tests; they build tests because they need measures to do assessment.

Historically, we have tried to build longitudinal scales by linking together a series of grade-level tests. I've tried to do it myself; sometimes I claimed to have succeeded. The big publishers often go us one better by building "bridge" forms that cover three or four grades in one fell swoop. The process requires finding, e.g., third grade items that can be given effectively to fourth graders and fourth grade items that can be given to third graders, and onward and upward. We immediately run into problems with *opportunity to learn* for topics that haven't been presented and *opportunity to forget* with topics that haven't been re-enforced. We often aren't sure if we are even measuring the same aspect in adjacent grades.

Given the challenges of building longitudinal scales, perhaps we should ponder our original motivation for them. For purposes of this treatise, the following assertions will be taken as axiomatic.

1. *Educational growth* implies additional capability to do increasingly complex tasks.
2. *Content standards* that are tightly bound to grade-level instruction can be important building blocks and diagnostically useful, but they are not the goal of education.
3. **Status Model** questions that Standards-based assessment was conceived to answer are about school accountability and improving lesson plans, e.g., *Did the school do what it needed to do so students finishing third grade have what they need to succeed in fourth grade; if not, what* tools *are they lacking?*
4. **Improvement Model** questions were added as annual grade-level data began to pile up in the superintendent's office, e.g., *Did the school improve so that third graders this year have more and better tools than third graders last year?*
5. **Growth Model** questions are personal, *Is this individual (enough) better at solving complex tasks now than last year, or last month, or last week?*

Meaningful open-ended tasks[1] are more direct instances of the goals implied by axiom 1 than multiple-choice items used across grades. This suggests an alternative strategy to replace the historical model for building the longitudinal scale. A more authentic scale could be built through a graduated series of these tasks; the scale would align naturally with the stated goal and should be more apparent to stakeholders outside the assessment community. Let the students in

[1] I have gotten ahead of myself here. This section really needs to use extended samples of student work, which typically are scored *0* to some small number *m*. Everything I have done so far has assumed *m* is *1*. I'll get to *m>1* eventually.

grade *y* demonstrate that they can actually provide more sophisticated essays, solutions, and analyses than students in grade *y-1* when placed in the appropriate situations.

This approach may not provide the diagnostic information about which tools students have or have not but it is appealing on a number of fronts, although most of the following should be taken as a list of dissertation proposals, not established truths:

1. *If you want to know if a student can deal effectively with increasingly complex tasks, the best evidence would be to ask the student to perform a complex task*.

   This is vaguely reminiscent of Rasch's original remedial reading project, which did not involve asking questions about main idea, topic sentences, literary devices, or meaning from context; just direct observation of the student reading.

2. *Success on these tasks is less tightly bound to grade-level curricula than standards-based MCQ*.

   Effective instruction of any sort should improve the students' performances on a relevant task, regardless of the specific topics presented or re-enforced, which mitigates the problem of identifying items acceptable for use across grades. It cannot matter what textbook series, teaching philosophy, or scope and sequence the system has adopted.

3. *Non-identical tasks can be scaled together, which is central to the issue of building scales that span grades and eliminates the problem of out-of-grade testing*.

   Raters are not evaluating the work sample against a common rubric and determining which steps were performed correctly. Rather the judge is comparing two work samples and deciding which is the more (or less) impressive. They may be responding to the same problem or one may be doing geometry and the other algebra; they may be in the same grade or a grade or two apart. None of this matters to the judge; it is just a matter of which work sample is the better response to the question that this student was asked[2].

4. *Scaling non-identical tasks is the solution to the problem in content areas for which schools may not follow the same sequence,* e.g., secondary math or science.

   If school *A* teaches algebra in ninth grade and geometry in tenth and school *B* reverses them, then a ninth grader in *A* should do better on algebra than a ninth grader in *B*. But the *B* student should do better on geometry than the *A* student, all other things being equal. And if the schools are comparable, the algebra samples from *A* should be comparable to geometry samples from *B*, for any grade.

5. *The scaling could be accomplished using data routinely collected in the regular course of testing and would require neither additional testing nor the administration of off-grade items*.

   Most large scale assessments now include some non-dichotomous items, perhaps open-ended, constructed response, and extended response. Because the scaling can be done across and within grades using responses to different tasks, any work samples

[2] This is not to say that it is necessarily easy to compare an outstanding response to a simple problem with a marginal response to a complex problem.

collected as part of the normal testing can be used, requiring no additional input from the students or effort by front-line educators.

6. *Validity of the MCQ-heavy within-grade standards-based assessments would be supported by a demonstration of a relationship with the measure based on the more authentic extended-response-based longitudinal scale*.

   Because the work samples were also scored traditionally and calibrated with the all other items in the assessment, the scale defined by the common calibration can be related to the new longitudinal scale, which will provide the link to equate all forms and all items. The degree of agreement between the two metrics will provide additional support for the validity of both, while facilitating the analysis of growth.

The order and spacing of student responses to authentic tasks should be consistent with educational and cognitive theories about development in the content area. Educators and stakeholders should be able to look at work samples and agree that progress along the scale demonstrates increased proficiency.

It will also show less favoritism to programs using conventional instructional models and textbooks. If an innovative approach is effective, it should be able to demonstrate its effectiveness on tasks less bound to the standard approaches.

Logistics of Thurstone Paired Comparisons

The process for scaling student work samples involves presenting pairs of samples to a judge. The two instances may represent the efforts of two students or one student on two occasions and may involve the same or different problems. At this point all we care about is we have two samples of student work. The question for the judge is, *Which sample is the more sophisticated solution?* The process is repeated for multiple judges and multiple pairings until sufficient judgments are available to scale all samples together with sufficient precision. It is not necessary for all work samples to be paired with all others; nor is it necessary that all judges see all pairings.

Alastair Pollitt (2004) has been investigating this method for several years as an alternative method for scoring the high-stakes English exams. While he talks some about using non-identical tasks, he doesn't mention building longitudinal scales. Pollitt's research suggests that making the paired judgment requires less time than the traditional scoring of two papers, but because 20 or 25 ratings seem to be needed for a pair, the overall effort is greater with paired comparisons; however, we are talking about a relatively small sample, one time (or every few years); perhaps 25 judges, two to three days; the judges require minimal training and could be in-state teachers (with some controls on what student work they see.)

The process for developing a longitudinal scale expects the work samples to have previously been scored before starting. This helps organize the work efficiently but also will be used later to anchor the scale. The following describes a manual process, which would require the judges to physically handle the work and to record the winner and loser of each pair. This is written as though it were being sent to the programmer so don't be offended by the level of detail. With digitally imaged samples, the entire process would be better managed by computer but don't run until you can walk.

1. Select work samples to cover the range of proficiency more or less uniformly.
2. Order the work samples by the traditional scores, lowest to highest.

3. Divide the samples into two stacks to be nearly parallel on the traditional scores by randomly placing the first sample into either the right stack or the left stack; place the second sample in the other stack and continue alternating.
4. Assign the stacks to a judge who compares the top sample in the *right* stack to the top sample in the *left* stack.
5. Place the more sophisticated sample in the *winner* stack face down; the other in the *loser* stack face down. Record the *winner* and *loser*.
6. After exhausting the first two stacks (*left* and *right*), turn the *winner* and *loser* stacks face up.
7. Remove the top sample from the *loser* stack and the bottom sample from the *winner* stack. This will pair the winner of the lowest pair with the loser of the second lowest pair, etc.
8. Send the two stacks to a different randomly chosen judge for another round; there will always be equal numbers in the stacks but one less than the previous,
9. Assign this judge the *winner* and *loser* stacks from a different, randomly chosen judge.
10. Continue with additional rounds until enough judgments have been made to link all tasks with sufficient precision and confidence.

In this process, the judges are the common, overlapping element; there need not be overlap in the students or in the tasks. It is not necessary to pair every sample with every other sample nor is it necessary that every judge see every pairing.

Scaling

The scaling of work samples is accomplished from a pair-wise matrix *N* in which element $n_{i,j}$ is a count of the number of times work sample *i* was rated higher than sample *j*. From this, a skew-symmetric matrix *D* is generated by computing $d_{ij} = \ln \frac{n_{ji}}{n_{ij}}$. If matrix *D* is complete and we're happy with the convention that the center of the scale is zero, then the scale value for sample *i* is the row mean of row *i*; if *D* is not complete, the solution requires solving a system of simultaneous equations; I've dealt with that a couple times. In either case, the result is a set of scaled work samples that cover the range of proficiencies expressed in a common unit of measurement.

Equating to the Regular Assessment

That last sentence perhaps needs a little expansion: "*proficiencies expressed in a common unit*" really means a "*unit common among themselves, not a unit common with the regular assessment*." That requires another step and that step is called equating. We have a logit scale value for each person used in the Thurstone scaling of the samples and we have an estimated logit ability for these same people based on the calibration of the regular MCQ-dominated assessment. In a perfect Rasch world, we can plot the two estimates and adjust one (or the other) scale by the adding (or subtracting) the *x* (or *y*) intercept.

Once again, equating should be used as another control on the model. In the process, it could answer some inconvenient questions or give inconvenient answers to some questions.

- If the plot doesn't define a line with slope one, then the two scales aren't using the same units. That's hardly fatal, like changing Fahrenheit to Celsius, but we should also weight the work samples differently in calibration with MCQs. (It's still a Rasch model but the simple raw score isn't the sufficient statistic anymore. This is not the same as including discrimination in the model and estimating it with the calibration data.)

- If the plot doesn’t define a straight line, then the relationship isn’t linear. We would be in the market for a linearizing transformation but I would want a defensible explanation, better than it seemed to work here.
- If the plot doesn’t define a line at all, then the two scales aren’t measuring the same construct. That’s good to know but we should stop scaling the work samples in with the multiple choice and report separate scale scores.

If the plot defines a line, then we can assert with more confidence than we had before, but somewhat short of total confidence, that we have a legitimate measure and what that measure means can best be seen by studying the progression of student work samples. The work samples, which can be expensive and time-consuming to collect, provide authenticity; the multiple choice bring efficiency and reliability. When they give us the same message from different sources, then we should have a good idea of what the aspect is we have a hold on, or hold of.