

Dear Prof. Hake,

I am writing in response to your post on the AERA-D listserv at <http://lists.asu.edu/cgi-bin/wa?A2=ind0603&L=aera-d&T=0&F=&S=&P=339>, concerning normalized gain score <g>. You posited that psychometricians seemed unaware of <g>. Based in part on his conversations with me, Physicist Joe Redish at the University of Maryland suggested that the case was less one of unawareness than one of uninterest, due to its limitations. I offer below some amplifications of this line of thought. The topics I discuss are, in turn, item response theory (IRT) as a way of dealing with floor and ceiling effects, IRT as a way of characterizing change in terms of conceptual development, further extensions, and lines of contemporary work. As an introduction to the second and third of these points, I recommend chapters 2 - 4 of the National Research Council's recent volume "Knowing What Students Know," an examination of the interplay between advances in psychometrics and cognitive psychology. The publisher, National Academies Press, provides without charge an online version of this volume for browsing: <http://www.nap.edu/books/0309072727/html/>

Best regards,
Bob Mislevy

I. IRT as a way of dealing with floor and ceiling effects

It is definitely the case that gain scores based on total scores or percent scores on a particular test is subject to floor and ceiling effects. This is among the vicissitudes of analyzing change in these terms, as explored in the Cronbach & Furby (1970) paper mentioned in your post. <g> is one way to attack the problem of making gains at different parts of the test score scale comparable. The alternative that represents the path toward tackling this problem in the psychometric literature (starting in the 1950's, taking hold in the 1970's) is item response theory (IRT). Here is the Wikipedia intro to IRT: http://en.wikipedia.org/wiki/Item_response_theory

IRT bears a key resemblance to <g> but three important differences that make it, rather than <g>, of more interest to psychometricians and educational measurement theorists. The similarity is the use of some sort of logistic transformation to account for the decreasing capacity of total scores to reflect change in proficiency as one approaches the top of the score range. Here is a succinct form of the argument:

Fischer, G. Some Probabilistic Models for Measuring Change. In De Gruijter, D. N. M. & L. J. Th. van der Kamp (eds), *Advances in Psychological and Educational Measurement*, London: John Wiley, 97-110, 1976.

The first difference is that IRT framework is applicable to any collection of tasks that have been calibrated together and show satisfactory fit to the model. This allows for assembling harder or easier tests, shorter or longer tests, even tests assembled on the fly

adaptively to each examinee, and still obtain the same estimates of status and change (in expectation). This advantage offers many practical advantages in test use and construction. In contrast, measures of $\langle g \rangle$ for easier and harder tests assembled from the same pool of items differ with choice of items. This is not a problem if the collection of items you have is the only one you care about, but it is like having thermometers that are each calibrated in their own degrees in different laboratories; maybe high quality and useful in their limited context, but not having generality or extensibility. Since IRT and $\langle g \rangle$ perform similarly in a given special context, but IRT enjoys many other advantages, the psychometrician prefers IRT to measure change.

The second is that IRT is grounded in the framework of probability-based reasoning. All of the practical uses mentioned above for IRT are accompanied by rigorous statistical methodologies for model criticism and parameter estimation. This allows for characterizing the accuracy of estimates for individuals, group differences, etc.

The third is that while IRT originated to solve practical problems that could be expressed but not solved under classical test theory, the way of thinking opens the door to new families of models that use the same kinds of reasoning and modeling approach but accommodate a wide range of kinds of responses, tasks, student-knowledge dimensions, etc. Among these are more interesting ways of characterizing change, for example,

Fischer, G. H. (1995). Linear logistic models for change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 157-180). New York: Springer.

II. IRT as a way of characterizing change in terms of conceptual development

The Cronbach & Furby paper pointed to by Prof. Glass is indeed an important one. It has a well-thought analysis of the problem of measuring change under classical test theory, including the implications of measurement error. Yet to me one of the most interesting parts is almost an aside, toward the end of the paper, which presages more recent developments. Through the use of standard test theory, evidence can be characterized and brought to bear on inferences about students' overall proficiency in behavioral domains, for determining students' levels of proficiency, comparing them with others or with a standard, or gauging changes from one point in time to another. But they cautioned that characterizations about the nature of this proficiency or how it develops fall largely outside the paradigm's universe of discourse:

Even when [test scores] X and Y are determined by the same operation [e.g., a true-score or IRT model for a specified domain of tasks], they often do not represent the same psychological processes. At different stages of practice or development different processes contribute to the performance of a task. Nor is this merely a matter of increased complexity; some processes drop out, some remain but contribute nothing

to individual differences within an age group, some are replaced by qualitatively different processes (Cronbach & Furby, 1970, p. 76).

Because IRT models performance at the level of individual tasks rather than at the level of total scores, it is possible to examine more closely the nature of the proficiency (at least in terms for its implications for success on tasks with different features) at different points along the scale. The “progress maps” employed by Mark Wilson and his colleagues at the Bear Educational Assessment Research (BEAR) center (see examples in chapter 4 of “Knowing What Students Know”) give examples of this type.

Such models are a kind of hybrid of a way of measuring progress with a common measurement scale (through distributions of IRT student scores) while acknowledging and beginning to characterize the qualitative nature of change. Admitting the latter raises the question of the comparability of a change of unit in performance across the full range of a scale, no matter what the metric, be it total scores, IRT, or the proportion toward 100% (<g>).

III. Extensions that model the nature of change

Sometimes the aim of a study is to compare programs or instructional methods with respect to gain on a common dependent variable, using comparable groups of students. The methods discussed in the preceding sections are used for this purpose. But sometimes the aim of a study is to analyze the nature of proficiency and its change, in which case the preceding methods may fall short. More recent work in psychometrics takes up this challenge. There is a self-imposed (by the field) constraint of doing so within the framework of probability-based reasoning. The issues addressed in this work include modeling performance in terms of production rule acquisition and differential strategy choice. Again some of this work is described in Chapter 4 of “Knowing What Students Know” at an introductory level. A more advanced review of one particular strand of this work (called “cognitive diagnosis”) is Brian Junker and Klaas Sijtsma’s (2000) paper “Cognitive assessment models with few assumptions, and connections with nonparametric IRT.” You can download a copy from Prof. Junker’s page of technical reports <http://www.stat.cmu.edu/%7Ebrian/bjtrs.html>. Another approach—for modeling students’ misconceptions using FCI-like tasks—is discussed in <http://www.cresst.org/reports/r646.pdf>. Modeling physics problem-solving in terms of a production system in an intelligent tutor is illustrated in the work of Prof. Kurt VanLehn, for example, at http://www.pitt.edu/~vanlehn/Stringent/PDF/95IJHCS_JM_KVL.pdf.

In summary, <g> has uses for comparing change between groups of similar students for a particular test form in relieving ceiling effects, but is limited to that form, is questionable for comparing groups starting at very different points, and is not grounded in a probability framework. Alternative approaches (especially IRT) have been developed in

psychometrics that are not subject to these limitations, and have provided a springboard for analyzing change from a wider range of perspectives, including ones motivated by developments in both cognitive psychology and physics education, that lie outside the reach of standard test score analyses.