The Continued Search for Nonarbitrary Metrics in Psychology

Susan E. Embretson
Georgia Institute of Technology

H. Blanton and J. Jaccard (2006) examined the arbitrariness of metrics in the context of 2 current issues: (a) the measurement of racial prejudice and (b) the establishment of clinically significant change. According to Blanton and Jaccard, although research findings are not undermined by arbitrary metrics, individual scores and score changes may not be meaningfully interpreted. The author believes that their points are mostly valid and that their examples were appropriate. However, Blanton and Jaccard's article does not lead directly to solutions, nor did it adequately describe the scope of the metric problem. This article has 2 major goals. First, some prerequisites for nonarbitrary metrics are presented and related to Blanton and Jaccard's issues. Second, the impact of arbitrary metrics on psychological research findings are described. In contrast to Blanton and Jaccard (2006), research findings suggest that metrics have direct impact on statistics for group comparisons and trend analysis.

Keywords: scaling, item response theory, change measurement

Measuring psychological constructs is fundamental to both psychological research and practice. As a science, psychology has been characterized by the development of an extensive collection of measures and tests. And yet, the means of establishing the scale intervals and corresponding interpretations is a recurring issue. The debate has taken various forms. Very early in psychology, Thurstone (1925, 1928) attempted to establish interval scaling of mental tests and even to set a zero point for intelligence. Although Thurstone's methods foreshadowed contemporary psychometric methods, it was Otis's (1917) notion of standard scores computed from normative groups that provided the intervals and meaning to test scores for most of the 20th century. Other dissatisfaction with metrics has involved debate on the importance of achieving scaling scores to meet the properties of fundamental measurement (Townsend & Ashby, 1984) or not (Lord, 1953).

Blanton and Jaccard (2006) examined the arbitrariness of the metrics in psychology in the context of two examples: (a) the measurement of racial prejudice by comparing decision times under varied conditions and (b) the establishment of clinically significant change. Blanton and Jaccard were concerned primarily with the linkage of the observed scores to the metric of an underlying psychological construct. They defined a measure to have an arbitrary metric if the location of the observed scores on the continuum (i.e., whether they are high or low) or the meaning of a one-unit change is unknown. For example, the observed scores on an attitude rating scale may be monotonically related to the continuum, but the relative score intervals and their meaning for the construct are not specified. According to Blanton and Jaccard, although research findings are not undermined by arbitrary metrics, individual scores and score changes may not be meaningfully interpreted. They showed how two popular strategies for score interpretation, "meter reading" and norming, fail to give meaningful interpretations.

Blanton and Jaccard (2006) illustrated their points with easily understandable examples and referents. I believe that their points are mostly valid and that their examples were appropriate. However, their article does not lead directly to solutions, nor did it adequately describe the scope of the metric problem in psychology. This article has two major goals. First, some prerequisites for nonarbitrary metrics are presented and related to the issues noted by Blanton and Jaccard. Second, the impact of arbitrary metrics on psychological research findings is described. In contrast to Blanton and Jaccard (2006), research in this area has indicated that metrics can have direct impact on statistics for both theoretical and applied research. Examples of relevant research in this area are given.

Prerequisites for Nonarbitrary Metrics

The Nature of Psychological Measurement

Psychological constructs are usually conceptualized as latent variables that underlie behavior. As noted by Cronbach and Meehl (1955), psychological constructs are theoretical constructions to explain behavioral consistency over varying contexts. Thus, a person's standing on a psychological construct has broad implications for his or her behavior, and consequently, the construct cannot be represented by a single index.

This conceptualization of a psychological construct has several implications for measurement. First, a person's

Correspondence concerning this article should be addressed to Susan E. Embretson, School of Psychology, Georgia Institute of Technology, 654 Cherry Street, Atlanta, GA 30332-0170. E-mail: susan.embretson@psych.gatech.edu

Vol. 61, No. 1, 50–55 DOI: 10.1037/0003-066X.61.1.50

Susan E. Embretson

standing on the construct must be inferred from his or her behavior. For practical reasons, measurement of persons in a natural context is the exception rather than the rule in psychology. So, psychological measures and tests are developed to observe responses. Responses to test items can be indicators of the construct if they are related to the postulated behavioral consistency in the broader context. Second, adequate measurement involves repetition over situations or items. Thus, psychological measures typically consist of multiple items or tasks that vary in content. Some items are more likely than others to be endorsed or solved. But if a construct is measured by the items, consistency of behavior across items should be observed. Third, because constructs are theoretical constructions, there is no natural metric. Instead, if a person's position on a construct is to be represented numerically, it is important to justify estimates of the latent construct on the basis of a measurement theory.

Currently, two different models are commonly used to obtain construct measurements from responses to measuring tasks: classical test theory (CTT) and item response theory (IRT). Although both models require response consistency and repeated observations to minimize measurement error, they differ in several ways that impact the score metric. First, differences between items in their psychometric properties (i.e., item difficulty, item intercorrelations) are treated differently to achieve meaningful metrics (see Embretson & Reise, 2000, for further elaborations). In CTT, item psychometric properties are regarded as fixed and must be balanced to obtain equivalent measures. In IRT, item psychometric properties are included directly in the model, and hence, comparable trait estimates may be obtained from tests with nonequivalent items. The most simple IRT model, the Rasch model for binary data, gives the probability that person *s* passes item *i*, $P(\theta)$, as follows:

$$P(\theta) = \exp(\theta_s - \beta_i)/(1 + \exp[\theta_s - \beta_i]), \quad (1)$$

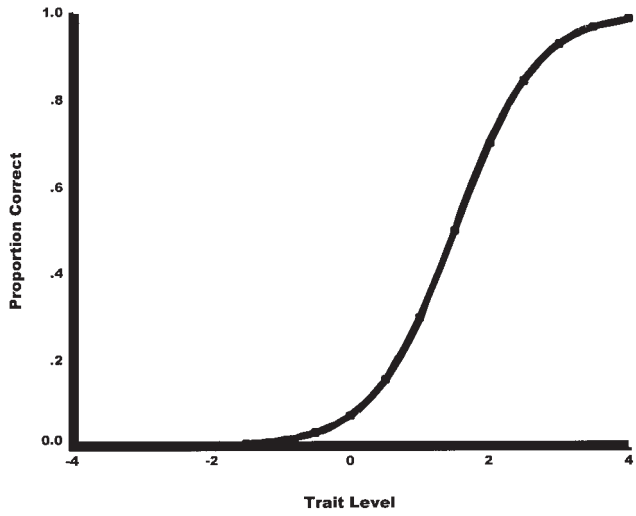
where the person's trait score is θ_s, and the item's location (or difficulty) is β_i.

Because Equation 1 is a logistic model, the probability of a response has a nonlinear relationship to trait level, as shown in Figure 1. Further, because item psychometric properties are included in the IRT model, trait levels with comparable metrics can be estimated from responses to any subset of items.

Second, raw scores are converted to a "meaningful" metric in qualitatively different manners under CTT and IRT. According to Michell (1990, p. 63), the meaning of a measurement requires specification of a comparison. Comparisons involve specifying the standard with which the score is to be compared as well as the numerical basis of the comparison. For example, the standard to measure a person's height is a ruler, and the numerical basis is a ratio. If the height is 6.3 ft, then the ratio of the person to the ruler is 6.3.

In CTT, the meaning of the score results from a comparison standard of other people, with order as the numerical basis. That is, the score has meaning by its relative position of the person in the norm group. It must be assumed that the norm group has obtained scores on tests with the same or equivalent items. If the raw scores are directly converted to a *z* score or a standard score, then the relative distances between scores are proportional to the raw score distances. That is, the raw score metric is linearly related to the standard score metric. However, if a norm group does not have a normal distribution, scores are often normalized, and the metric consequently changes. That is, some score intervals are stretched while others are reduced. In this case, the relationship of raw scores to standardized

Figure 1
Relationship of Observed Scores (Proportion Correct) to Item Response Theory Trait Level Estimates

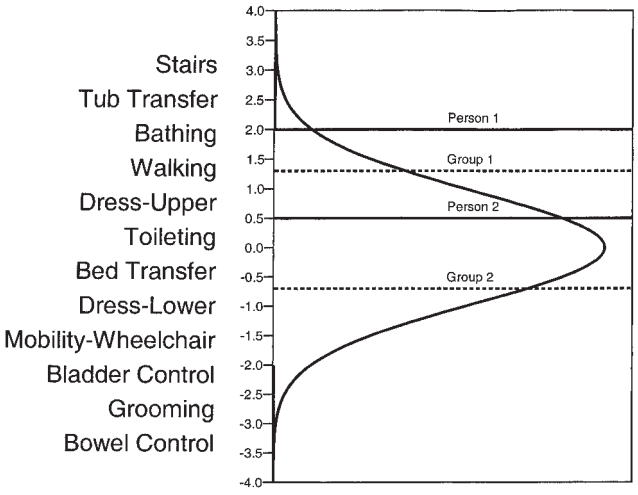


scores depends on the score distribution in the specific normative population. Different normative populations will lead to different score intervals.

In IRT, however, both the comparison standard and the numerical basis are different than in CTT. The comparison standard is items, and the numerical basis is difference. That is, scores are compared with items for differences. In IRT, items and persons are placed on a common scale by modeling item responses jointly from person and item parameters. Persons are compared most directly with items in the Rasch model for binary item responses in Equation 1. The probability of an item response is determined by the simple difference between the person's standing on the trait and the item's location. If the person and the item have the same location on the trait, then the probability that the person passes or endorses the item is .50. The probability of item endorsement or success increases with increasing distance of the person's location above the item.

Common scale measurement can be illustrated by the data presented on the functional independence measure (FIM) in Figure 2. FIM is a behavioral survey for elderly adults. The Rasch model has been shown to fit FIM data reasonably well, which indicates that the scale locations describe adequately the relative order in which these functions are lost in the aging population. Figure 2 locates the Rasch scale values (β_is) of each item. The items on the top describe difficult activities, such as climbing stairs, whereas items on the bottom describe easier activities that are maintained relatively well. Because persons and items are measured on a common scale, trait level scores (θ_ss) for particular persons may be located on the same scale as the items. For example, Person 1 has a probability of .50 for independent bathing and has an increasing probability of performing the lower level activities on the scale. Person 2 is located between two items, which means that the probability of successfully completing upper dress and then tasks above is less than .50, whereas the probability of successfully completing toileting and all lower items is greater than .50. IRT scores may also be compared with norms, if desired. Figure 2 shows the position of the individuals in a norm group of persons in their 80s. Notice that both persons are above average.

Figure 2
Item and Person Locations on the Functional Independence Measure



Estimation in IRT has several important features as compared with CTT (see Embretson & Reise, 2000, for further explication). First, the item estimates are not totally dependent on a particular population, as they are in CTT. Because there is a complete model of item responses, estimates of item difficulty are controlled for trait level. Second, trait levels are estimated in the context of an IRT model that includes item psychometric properties. The estimated trait level for a person is the one that yields the highest likelihood given the item responses and the item psychometric properties. Third, the intervals between persons have uniform meaning for the (log) likelihood that items are solved. This principle is most direct in the Rasch model, because the only difference between items is their location or difficulty. The relative distance between persons reflects in their difference in performance (i.e., log likelihood) for any item. It is sometimes argued that optimal measurement properties, such as specific objectivity and fundamental measurement, are achieved in the Rasch model (e.g., Fischer, 1995). Other models, such as the two-parameter logistic model, involve the further complexity of varying item discriminations. Fourth, consistent with traits as theoretical constructs, no natural metric is associated with the IRT model. Both the zero point and the score-interval width depend on two numerical constraints that are specified by the researcher. But, significantly, the relative differences between IRT trait levels retain meaning with linear transformations. Fifth, IRT models may be applied to a variety of item formats. Special IRT models have been developed for rating scale responses (e.g., Muraki, 1993) and continuous formats, such as response times (e.g., Mellenbergh, 1994).

Implications for Nonarbitrary Metrics

Blanton and Jaccard (2006) discussed at length the importance of a nonarbitrary metrics for a *meter-reading* approach to score interpretation. In meter reading, scores have meaning due to theoretically or clinically important benchmarks. This approach, in general, seems outside accepted guidelines for test interpretations, such as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Theories, at best, are incomplete, and they are often insufficiently operationalized. Clinical experience is subject to a long list of limitations, including memory biases and the representativeness of experience.

Setting aside the general limitations of meter reading, Blanton and Jaccard (2006) noted that arbitrary metrics

lead to further problems with meter reading. First, the meaning of a specific point to benchmark a trait is impacted by many features that can make it arbitrary. Blanton and Jaccard's example is the zero point that is obtained by subtracting response times obtained under varied conditions. According to them, this method is used to separate positive and negative racial attitudes on the Implicit Association Test. Of the several problems noted by Blanton and Jaccard, the most important impact on the zero point is probably item differences (e.g., the Black vs. the White faces) in psychometric properties and their possible interactions with conditions. Item differences and interactions can shift the observed zero point away from the construct-neutral point. Second, and related to the factors that impact the zero point, the score intervals may not be equal. Blanton and Jaccard noted that even if scores are based on indicators with justifiably equal intervals, such as response times, equal intervals on the construct are not guaranteed. That is, the response time indicator may not be linearly related to the construct.

These concerns emphasize the importance of having a model to map changes in the construct to changes in the items. The construct, as noted above, reflects consistency in item responses. Applying an IRT model to scale the data can be an effective means to map the construct to the items. For binary data, Figure 1 shows an item-to-construct mapping given by the Rasch model. For rating scale and other polytomous data, Blanton and Jaccard's (2006) Figure 1 could represent the mapping of specific item thresholds to the construct, as would be given by an IRT model, such as the generalized partial credit model (Muraki, 1993).

Blanton and Jaccard (2006) argued that a second method of score interpretation, norming, also fails to provide sufficient information about the construct. Consider the norms shown in Figure 2. Without the item information, the scores shown would be related only to other persons and would indicate little about what the person could do. However, if the comparison with items is available, as with the common scale measurement in IRT, additional meaning sometimes may be obtained. With the FIM example, the items have direct meanings for behavior, and hence, the likelihood of various activities is part of the interpretation of the score. For other constructs, particularly abilities, cognitive models of item difficulty may provide further meaning about the kinds of items at various points in the scale.

Blanton and Jaccard (2006) recommended using multiple groups to guide interpretation of the construct. Figure 2 also shows the hypothetical placement of the means of two groups of elderly people: Group 1, elderly persons in assisted-living residences, and Group 2, elderly persons in nursing homes. Again, the common scale measurement allows meaningful linkages of scores, group means, and group variability to items.

Model-based measurement, which includes IRT, does not provide a universal metric with true zero points and interval widths. How such metrics could be obtained is difficult to envision for most psychological constructs. Instead, model-based measurement can provide justifiable relative distances between scores and the possibility of linking scores to items and external data for meaning.

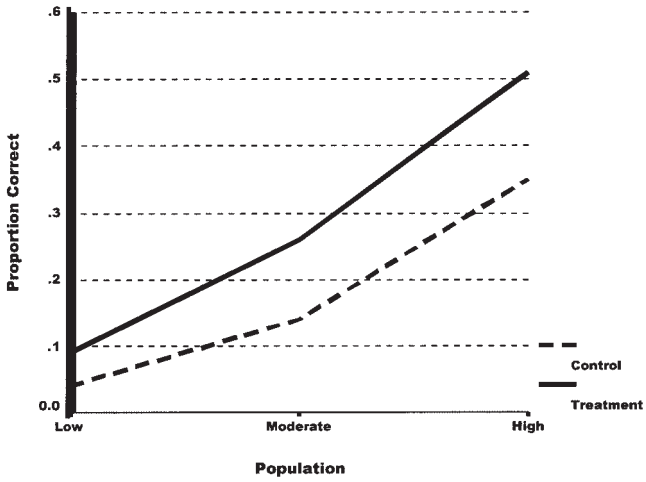
Impact of Measurement Scale on Statistics

Although Blanton and Jaccard (2006) do not believe that arbitrary metrics have great impact on research findings, application of parametric statistics to ordinal data (i.e., scores with uncertain intervals) is, in fact, a recurring topic in psychometrics and psychological statistics. Under certain restricted conditions, Davison and Sharma (1990) showed that relationships found for observed scores will also hold for the latent variables, but the same error levels and power may not apply. Of course, this can impact inferences that are drawn about the latent construct from the observed scores, because significance levels will be impacted.

Under other conditions, results from the observed scores can be quite misleading. For example, Maxwell and Delaney (1985) showed that two groups with equal true means can differ significantly on observed means if the observed scores are not linearly related to true scores. The effect occurs when differently shaped score distributions are coupled with inappropriate test-difficulty levels. Similarly, significant interactions can be observed from raw scores in factorial analysis of variance designs (Embretson, 1996) when no interaction exists in the true scores. When the raw scores are not linearly related to the latent construct, spurious interactions can occur.

Consider the proportion-correct means for six groups that are plotted in Figure 3. The groups differ in initial levels on the trait (low, moderate, and high) and conditions (control vs. treatment). Figure 3 shows an interaction of treatment with group level; the difference between treat-

Figure 3
Observed Score Means for Three Populations in Treatment Versus Control Conditions



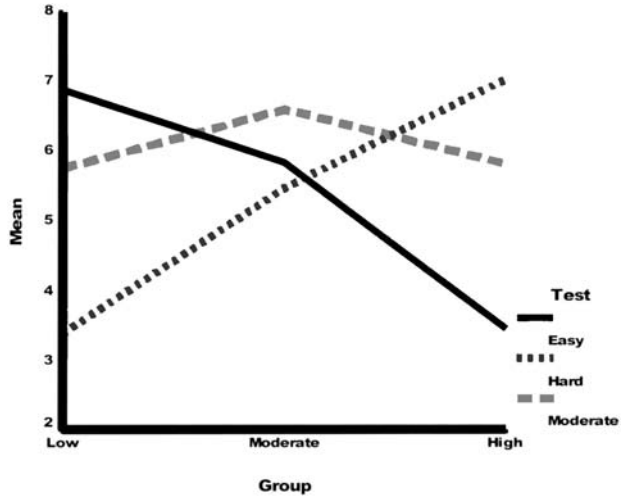
ment and control is greatest for the high-level population. Because these are simulated data with main effects but no interaction, the interaction observed for proportion-correct scores is spurious. In general, the greatest differences between conditions will be found for the population for which test-difficulty level is most appropriate (see Embretson, 1996). The test was hard, so the high-level population had the greatest difference.

Sometimes the impact of measurement scale on mean comparisons can be quite dramatic. Consider the mean gain scores for three groups presented in Figure 4. These data were based on simulations (Embretson, in press) with 1,000 cases per group. True means were −1.5, −0.5 and 0.5 for the low, moderate, and high populations, respectively. Mean true gain was equal across populations: 1.0. Item response data were simulated for easy, moderate, and hard tests, which were most appropriate for the low, moderate, and high populations, respectively. Scores were calculated from the generated data at each time point, and then gain scores were computed.

Figure 4 shows that the populations generally differ in mean raw gain. Further, which population gains the most depends on the test. For the moderate test, for example, the moderate population gains the most. In general, the most gain is observed for the population for which the test is most appropriate. For each test, significant differences were observed, and nontrivial effect sizes were observed. Yet, in fact, true gains were equal.

Group comparisons of change and trend are also impacted by observed scores that are not linearly related to the construct. Several simulations have shown that the group that changes the most is the one for which the test-difficulty level is most appropriate (Embretson, 1994, in press). Similarly, longitudinal studies on trend will also be impacted by uncertain score metrics (Embretson, in press).

Figure 4
Mean Raw Gain Scores From Three Groups on Tests of Varying Difficulty



Thus, research findings are not immune to scaling artifacts. Impact on statistical tests, effect size estimates, and parameter estimates can be expected. To the extent that psychological research is based on arbitrary scaling, the implications for research findings could be quite broad.

Conclusion

The adequacy of the scaling of scores impacts not only test interpretations but also psychological research findings. What, one might ask, constitutes a nonarbitrary scaling? That, of course, has been the subject of considerable debate. According to some scholars, appropriate applications of Rasch models can achieve the necessary desirable qualities. But of course, for some psychological measures, the Rasch model simply will not fit. For other measures, achieving fit to the Rasch model leads to significant narrowing of the construct.

Achieving nonarbitrary scales requires mapping the observed outcomes to fallible items onto the latent construct. Applying contemporary methods in model-based measurement is one possible solution. Applications of model-based measurement are rapidly increasing in the testing industry, but applications to psychological research are lagging. Although a wide variety of models with explanatory potential are now available and accessible through popular software (see DeBoeck & Wilson, 2004), they will not be applied effectively unless psychologists are better prepared in measurement and statistics. Meeting this challenge will require a refocusing of efforts on several levels in the training of psychological researchers.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Davison, M. L., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regressions. *Psychological Bulletin, 107,* 394–400.

DeBoeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

Embretson, S. E. (1994). Comparing changes between groups: Some perplexities arising from psychometrics. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 211–248). Ottawa, Ontario, Canada: Edumetric Research Group, University of Ottawa.

Embretson, S. E. (1996). Item response theory models and inferential bias in multiple group comparisons. *Applied Psychological Measurement, 20,* 201–212.

Embretson, S. E. (in press). Impact of measurement scale in modeling development processes and ecological factors. In T. Little & J. Bovaird (Eds.), *Issues in development in an ecological context.* New York: Praeger.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fischer, G. (1995). Derivations of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York: Springer-Verlag.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.
Maxwell, S., & Delaney, H. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97,* 85–93.
Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29,* 223–236.
Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Erlbaum.
Muraki, E. (1993). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.
Otis, A. S. (1917). A criticism of the Yerkes-Bridges Point Scale, with alternative suggestions. *Journal of Educational Psychology, 8,* 129–150.
Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433–451.
Thurstone, L. L. (1928). The absolute zero in the measurement of intelligence. *Psychological Review, 35,* 175–197.
Townsend, J. T., & Ashby, G. (1984). Measurement scale and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.