

The Problem of Units and the Circumstance for POMP

Patricia Cohen

New York State Psychiatric Institute and Columbia University

Jacob Cohen

New York University

Leona S. Aiken and Stephen G. West

Arizona State University

Many areas of the behavioral sciences have few measures that are accepted as the standard for the operationalization of a construct. One consequence is that there is hardly ever an articulated and understood framework for the units of the measures that are employed. Without meaningful measurement units, theoretical formulations are limited to statements of the direction of an effect or association, or to effects expressed in standardized units. Thus the long term scientific goal of generation of laws expressing the relationships among variables in scale units is greatly hindered. This article reviews alternative methods of scoring a scale. Two recent journal volumes are surveyed with regard to current scoring practices. Alternative methods of scoring are evaluated against seven articulated criteria representing the information conveyed by each in an illustrative example. Converting scores to the percent of maximum possible score (POMP) is shown to provide useful additional information in many cases.

One of the most fundamental tasks in building a science is the establishment of standard operationalizations of the major constructs used in its theory (Mandel, 1964; Meehl, 1967, 1990; Tukey, 1969). The presence of agreed-upon measures of scientific constructs characterized in agreed-upon units of measure permits the progression from speculations about “the way things are” to a science in which facts and theory accumulate in mutually refining ways. Ideally, theory and measurement are intertwined in a series of reciprocal successive approximations. However, with the possible exception of the area of psychophysics, even the area of experimental psychology with one of the longest research traditions in behavioral science and many highly sophisticated measurement efforts, has not yet established a consensus on units of measurement. Other areas of psychology

Data for this illustration were collected as part of NIMH Grants R01- 36971 and R01-54161, Patricia Cohen, P. I. Stephen G. West was supported by NIMH Grant P50-MH39246. We thank Peter Killeen, editor Roger Millsap, and three anonymous reviewers for their helpful comments on an earlier version of this manuscript. Correspondence should be directed to Patricia Cohen, Epidemiology of Mental Disorders Research Unit, 100 Haven Avenue, New York, NY 10032; e-mail:prc2@columbia.edu.

face an even more difficult task given the literally thousands of different measures of abilities, personality traits, attitudes, motives, moods, and other dispositions that have been employed in research, and the plethora of units in which the scales are measured.

In this article we consider a number of issues associated with the development of meaningful units for scales in the behavioral sciences, units that are useful both in communicating findings to other behavioral scientists and in formulating theory and prediction from theory. We propose a rarely used but highly communicative method of scoring, the *POMP*. In *POMP*, the score assigned to each individual is a percentage, reflecting the individual's position on the scale as a *Percent of the Maximum Possible* score achievable on the scale. We argue that the development of meaningful, easily communicated units such as *POMP* is required in order that we undertake a critical task facing behavioral scientists — characterization of *material effects*. By material effects we mean effects of sufficient magnitude to be taken seriously by substantive researchers and, perhaps, by practitioners, policy makers, and even the public (Menand, 1996) as well.

Impact of Consensus on Measures and Units

Lack of consensus on measures of constructs and on the units in which scales are measured hinders communication among scientists who are working in the same or adjacent areas of research. Each of a number of measures of the same or closely related constructs may be assessed in units that are idiosyncratic to the particular response options and scoring algorithm used by a given investigative team. Because familiarity with the units has not been established among different investigators in the area, little attention is paid to findings expressed in the units of these scales. Familiarity is such an important prerequisite for the understanding of units that without such a framework it can hardly be said that a unit has meaning. One of the most important tasks of science is to provide such a framework (Cliff, 1993).

Meaningful units allow investigators to express both the magnitude of observed relationships between variables and the impact of experimental manipulations in terms of these units. Further, meaningful units permit examination of whether we have material effects without reliance on statistical conventions. As Tukey has pointed out, without meaningful units the theory of elasticity would come down to “when you pull on it, it gets longer” (1969, p.86). Perhaps the over-reliance on null hypothesis significance testing that is characteristic of the behavioral sciences (American Psychological Association, 1996; J. Cohen, 1994; Meehl, 1967; 1990; Tukey, 1969) may be attributed, in part, to the difficulty of formulating more than directional scientific hypotheses in the absence of agreed-upon standard measures with known units. When there is

no other way to convey the meaning of, for example, a four point difference on a scale, we often revert to the convenient fiction of describing anything that is statistically significant in our study as “important” and anything that is not as “not important” (J. Cohen, 1994; Schmidt, 1996). Or worse yet, the researcher or reader may consider the smallness of the p -value to reflect the importance of the phenomenon. But it is clear that the scientific significance of a phenomenon does not reside in the level of statistical significance achieved or having confidence intervals that do not overlap zero, both of which strongly reflect the size of the sample used in the study (Thompson, 1992).

How a Framework is Gradually Constructed Around Units, and Meaning is Developed in Context

Within the physical sciences, measurement generally begins with some arbitrary reference unit around which information builds and familiarity is created (Mandel, 1964). Perhaps some of the examples that come most easily to mind are degrees of temperature, measures of length, time, and counts of discrete objects.

Degrees of temperature were a scientifically and practically useful measurement unit long before absolute zero was defined (Middleton, 1966). Equality of units of temperature itself depends on the method of measurement. When water was the frame of reference and the zero and one hundred degree points Celsius (C.) were defined in terms of boiling and freezing, there was the problem of additional cooling required to change the state of water at zero degrees C. to ice at zero degrees. Although scientific procedures used to define units of temperature have changed over time, decisions about which units to use have often been based more on practical concerns than on scientific necessity. Nevertheless, agreement among scientists on the meaning of a degree of temperature was sufficient to permit refinement of experiment and theory in the physical sciences.

Even when the units are agreed upon, the meaning is defined in context. Thus, for example, a difference of four degrees Fahrenheit may be immaterial when one is cooking, of modest importance when one is deciding whether to go to the beach, and of considerable impact when one is measuring human body temperature. The difficulty in accomplishing a conversion from Fahrenheit to centigrade (Celsius) measures of temperature in the United States, and more generally to the metric system, is testimony to the importance of familiarity in context for ordinary interpretation of units.

There are a few cases in the behavioral sciences in which long familiarity with a scale (and usually substantial standardization data) has made relatively familiar units out of measures of abstract constructs. In psychology we have become familiar enough with the IQ unit to convey a sense of what are meaningful

differences, even though the methods used to assess and calculate IQ vary. A long clinical tradition with the Beck Depression Inventory has led to a shared consensus about the meaning of a score of 15, for example, which is widely used as a cut-off indicating clinically significant depression (Beck & Beamesdorfer, 1974). However, evidence suggests that in the case of many other scales, researchers have ignored the meaning of units. Different sets of items are used to measure the same construct, scales are lengthened or shortened arbitrarily, and different schemes may be used to assign a total score even when the same set of items is used.

In this article, we begin by reviewing the use of units as reported in two leading American Psychological Association journals, one basic and one applied, to provide an illustrative overview of some of the different ways units are currently used in practice in psychology. We then consider what information is available in each of these representations of units. We then highlight our proposed POMP scoring and argue that this method may have certain advantages over traditional scoring methods. We propose several criteria that may be useful in determining whether results can be usefully communicated to other researchers. The degree to which different scoring methods meet these criteria is illustrated by showing the differential interpretations that result when a common data set is analyzed. Finally, we consider some of the strengths, weaknesses, extensions, and future directions for the use of POMP scoring as a potential solution to the problem of having interpretable units.

A Review of Recently Published Studies

To provide a taxonomy of current normative practice with respect to scale units, we reviewed volumes for 1996 from two leading American Psychological Association journals, the *Journal of Consulting and Clinical Psychology (JCCP)* and the *Journal of Personality and Social Psychology (JPSP)*. Note that JCCP primarily addresses applied research issues whereas JPSP primarily addresses issues of basic research. All regular empirical articles from each volume were reviewed and classified in terms of the predominant methods used to identify scale units in the one or more studies reported. Based on a preliminary examination of the literature, we coded the scoring of scales in each article into one of five categories. (The properties of these categories will receive more extended consideration in the next section).

1. *Scores without intrinsic or supplied meaning (Item Sums)*: Studies reporting unnormed item sums, or scales for which no information was reported on the means (*Ms*) and standard deviations (*SDs*).

2. *Single or averaged Items*: Studies reporting single items or other scores for which the potential range was provided, including average item scores.

3. *Standardized (z) Scores*: Studies reporting z -scores based on sample data.

4. *Percentages*: Studies using percentages rather than scaled variables. Percentages were almost always based on dichotomous or dichotomized variables such as the percent of subjects with a particular disorder.

5. *Meaningful Units*: Studies that used units with established meaning, typically based on physical measurements (e.g., blood pressure, weight) and studies reporting units for which a referent meaning was provided (e.g. the Beck Depression Scale for which the authors provide “normal”, “marginal”, and “clinical” ranges based on previous work).

Journal of Consulting and Clinical Psychology

Volume 64 (1996) of the *Journal of Consulting and Clinical Psychology* included 29 regular (not special section) research articles. Of these, four reported findings only or predominantly in terms of percentages, and four used measures with meaningful units. No studies relied predominantly on single or averaged items. One study used a mixture of techniques and one study was not classifiable. In 19 studies most or all scales used item sums or did not report means and standard deviations. Thus, for about two thirds of the reports no interpretable value was provided by the M s or SD s of variables, or by raw-unit statistical coefficients.

Journal of Personality and Social Psychology

Volume 70 (1996) of the *Journal of Personality and Social Psychology* showed a strikingly different pattern. In the fields of study published in this journal, single item scales are quite common and, perhaps because of this tradition, it is also quite common to create scale scores in the item metric by averaging items. Of the 45 of 46 articles that could be characterized by a predominant way of scoring, three used percentages, five used measures with meaningful units, and one used z scores. The potential range of the scales was reported in 14 studies; 11 of these used average item responses and provided the response options. About half of the studies (22) used item sums or reported no scale characteristics other than reliabilities and correlates.

These two journals illustrate an interesting division that now exists among different specialty areas in psychology and other behavioral sciences with regard to the scales commonly used in research studies. In certain substantive areas, such as some of those reflected in the volume of the *Journal of Personality and Social Psychology* as well as in *Health Psychology*, brief *ad hoc* scales are commonly devised to reflect the constructs of interest, with reliabilities assessed by coefficient alpha, and validities inferred on the basis of face validity of items

and theoretical coherence of findings. Frequently the scores used for these scales reflect the mean of items responded to on a common Likert scale. Although some scales are employed across research settings and investigators, it is not typical that there is an explicit interest in accumulating information about the scale, nor an expectation that scales will be developed to stand as reference standards for certain constructs. Other research areas, such as the studies of clinical populations reflected in the *Journal of Consulting and Clinical Psychology*, achievement and intelligence assessment, and some areas of developmental psychology, are much more likely to look askance on investigations in which novel measures are employed to represent previously studied constructs. Face validity and internal consistency are generally considered to be inadequate evidence of construct representation, and population referencing is far more common. Nevertheless, even in these substantive areas, scale units are frequently ignored unless it is possible to scale them with reference to some population. Even frequent scale users may find it hard to guess whether most items were responded to in one direction or the other, that is, whether a score of 22 means that the respondent tended to endorse or not endorse most items.

Scoring Alternatives for Collections of Items and Their Implications

Having identified the units that are most commonly used in leading journals in personality, social, and clinical psychology, we now consider the usefulness of each of these units in communicating information to other scientists. Our focus here is on the information communicated about collections of items (scales) when each method of scoring is used. We examine the information available from single studies rather than that generated by studies involving measure development or representative populations. We also highlight our proposed Percent of Maximum Possible (POMP) score, which may often have certain advantages in communicating scale information.

Item Sums

As shown by our survey, one of the most common methods of scoring collections of questions for a scale is to simply sum items (either subtracting or reversing and adding questions for which high values indicate lower amounts of the quality being measured). Given the variable numbers of items and the potentially variable range of item responses, the consequent scores will convey virtually nothing to most investigators. However, a research team that repeatedly employs the same measure and same scoring system may gain familiarity with what should be considered high or low scores for their particular version of the measure.

Single or Averaged Items

As noted, scales may be represented by an average item score, especially when all items use the same Likert or dichotomous response format. These scores are potentially more informative than item sums because the respondent's location on the Likert scale can add substantive knowledge. For example, if a Likert score ranges from 1 = Not at all true to 5 = Completely true, a score of 4 indicates that the respondent thought that the items, on the average, were mostly true. When items have different response scales, average item scores are rarely used as they have no clear interpretation. Units of single item measures have many of the same advantages and disadvantages of averaged item scores with respect to the information conveyed to other researchers.

Standardized Scores

Another approach that is frequently used is to express the scales in standard deviation units as z scores based on the observed data in the sample. Like scores based on population norms, z scores permit normative interpretations. However, the quality of the normative interpretation is entirely dependent on the quality of the observed sample. That psychology often bases its research on "accidental samples of convenience" (Cook & Campbell, 1979, p 71) means that the sample means and standard deviations on which standardized measures are based may be arbitrary, and therefore not generalizable to any useful population. In addition, z scores based on different samples from any given population will vary as a function of sampling error. Moreover, z scores can be misleading when distributions are highly skewed, a situation that is quite common in certain areas of research such as psychopathology, health, and substance abuse (see Micceri, 1989). Finally, employing z score units can encourage unwary investigators to think of the units as comparable, even when based on very different populations, for example, clinic patients and college students. Nevertheless, a meaning of individual scores with regard to a particular sample on which the standardization is based is conveyed by z scores and not by most other methods of scoring.

Scores Based on Population Norms

When scale developers or other researchers have collected data from a representative sample of a well defined population, scale scores can be converted to T scores ($\mu = 50$, $\sigma = 10$), or to z scores ($\mu = 0$, $\sigma = 1$) or deciles based on the representative sample rather than the sample at hand, or to other scores that provide an established framework for interpretation. Such scores convey much additional useful information, and research findings formulated in terms of

normed scores may be a theoretical step ahead of scores for which no such framework can be provided.

Scores Based on Established Units

In some theoretical and empirical work it is possible to use established units such as seconds or days, temperature, blood pressure readings, distances, or counts. However, as employed in behavioral research, even counts should often be best considered to be arbitrary units (see Mosteller & Tukey, 1977). When counts are used we “understand” the unit, but it may or may not translate into something more meaningful than any other arbitrary score. Absent a familiar context that supplies meaning to counts (e.g., as the number of key pecks a pigeon will complete as a function of the rate of reinforcement), the fact that they are counts is not enough to supply a useful meaning. Once again, experimenters who work in the operant reinforcement area are likely to develop both familiarity with their particular counts and consensus as to what constitutes a material effect, just as those working with biological assays do. When that happens and the operationalization is widely accepted we have one of the foundational building blocks for theory development. Very often, however, those outside the area are likely to have to take their word for it, having no framework for understanding what is a lot and what is a little. For example, in the current psychiatric nosology (DSM IV, American Psychiatric Association, 1994) the number of a specified set of symptoms or criteria is used to determine diagnoses. Without knowledge of the nosological framework, the number of symptoms in an area is not interpretable. Thus, even for such transparent scales as counts the generalization holds: Virtually all scales require a context to supply meaning.

Moreover, it cannot always be assumed that the framework will hold constant over time. For example, even in the case of dollars we must constantly adjust for inflation, putting the dollars into the presumably familiar units of a specific constant year. To talk about unadjusted differences (e.g., in welfare supplements, national debt, price of bread) can be highly misleading (see Sechrest, 1985).

Scores Represented as the Percent of Maximum Possible (POMP)

As an alternative to conventional scoring schemes, we would like to highlight an alternative which we term percent of maximum possible or POMP scoring. This alternative may often be useful in communicating information to other researchers. Given a set of items that are combined to produce a scale, it is possible to determine what the maximum and minimum possible scores would be and to express those as 100 and 0, respectively. Specifically,

- (1) $POMP = [(observed - minimum)/(maximum - minimum)] \times 100$,
 where *observed* = the observed score for a single case,
minimum = the minimum possible score on the scale, and
maximum = the maximum possible score on the scale.

Similar scales are, of course, very widely used in grading school tests, and thus have a generally understood context. Conventions have developed that tend to lend differences in these units an intrinsic meaning. The school grades example is instructive because the meaning is generally understood despite the fact that it cannot be assumed that a zero score means the student had learned nothing, and a score of 100 does not necessarily mean that all course material was completely mastered. How close these scores come to this meaning depends (among other things) on appropriate sampling from the content domain. There is also usually an informal consensus on the approximate appropriate mean and range for such scores as well. Different instructors may devise tests of the same material that are seen as inappropriately easy or difficult (as shown by means that are too high, e.g., 98, or too low, e.g., 20), or as insufficiently discriminating among students (too small a range or *SD*). Such descriptions make it clear that there is a generally understood framework for grade scores, more or less common to a range of academic settings.

This method of scoring is not common in psychological measures; indeed, our literature review found no instances of POMP scoring. When the items are dichotomous and scored (0,1), it is equivalent to average item scores (times 100) or “percentage correct,” the percent of responses in the keyed positive direction.

Scoring Methods are Linear Transformations: Implications

It is important to recognize that each of the scoring methods considered in this paper is a linear transform of the other methods. Other scoring methods, such as decile scores, that shrink or expand ranges of the scale differently will present somewhat different issues.

Scale Adequacy

Questions about scale adequacy, including reliability and validity, are quite independent of questions about the information content of these scores. Classical test theory (Lord & Novick, 1968) and many of the modern approaches to measurement (e.g. item response theory, Hambleton & Swaminathan, 1985; generalizability theory, Shavelson & Webb, 1991) have ignored the question of exactly what the unit is for the final measure, since it is clear that whether items are summed or averaged, and whether items scored in the opposite direction are

subtracted or reversed and added, the resulting internal consistency, reliability and correlations with other measures will be the same.

Statistical Outcomes and Statistical Power

That the scaling methods, including POMP, are linear transforms of one another also means that test statistics (F , t , etc.) resulting from statistical analysis will remain constant across transformations. While some aspects of an analysis will change with scaling, for example, raw differences between means and raw unit regression coefficients, the test statistics from the analysis will remain constant. Furthermore, any computations of statistical power will remain the same across scaling approaches, because power analysis is based on *standardized* effect sizes, which do not change under linear transformation of the original scales.

Statistical Findings and Scale Units

An important goal of scientific investigation is the expression of relationships among variables in terms of accepted units of the variables, functional forms of the relationship, and “constants” expressing the empirical values of elements in the equation (see e.g., Anderson, 1970). For example, $e = mc^2$ is a testable theoretical statement only when appropriate measurements for its components can be specified (although this does not mean that the relationship applies only to a *particular* set of scales, Falmagne, 1992). In the behavioral sciences we usually assume that such equations can be represented by linear, simple curvilinear, or (usually linear) interactive relationships that can be usefully estimated by regression analysis. The “constants” in these equations are represented by the intercept and raw unit regression coefficients. Equivalently, in some analyses (often investigated by t tests or analysis of variance) the scientific estimates are provided by differences in group means. Statistical measures such as raw regression coefficients and intercepts, and raw differences between means depend on an understanding of the units for their information value. Arbitrary scoring impairs communication with colleagues as well as the lay public, to the extent that one journal actually bans the reporting of regression coefficients (*American Journal of Epidemiology*, 1997). Even M s and SD s of the variables under study communicate little useful information when they are expressed in completely unfamiliar or arbitrary units. When measures are shared by researchers and are scored using the same scoring scheme, there is a possibility of comparing scale M s and SD s across research samples; however, such comparisons are relatively rarely reported in the current literature. Readers may use M s and SD s to compute *standardized* measures of effect size (such as d) even when the study’s authors do not; thus reporting M s and SD s may enhance the

possibility of including the study's results in meta-analyses. Nevertheless, all too frequently for most measures, these fundamental statistics are of sharply limited utility in conveying information about what was found in the study.

The Role of Units in Estimates of Effect Size

Theorists may try to avoid the problem of non-standard measurement by formulating theories in terms of effect size metrics that avoid raw-unit implications or by dichotomizing scales to produce meaningful units (e.g., case versus non-case).

Standardized Effect Size Measures

Some conventions have been proposed (J. Cohen, 1962, 1988) and widely employed, for the consideration of small, medium, or large effects formulated in standardized ("unit-free") coefficients, based on a rough normative estimate of effect sizes actually reported by psychological researchers. According to this convention, small, medium, and large effects are represented by Pearson r s of .1, .3, and .5, respectively, or standardized mean differences (d s) of .2, .5, and .8, respectively. However, even for such unit-free indices, what is considered a small effect in one research area may be considered large in another research context. For example, Rosenthal (1990) has shown that the effect sizes calculated from successful randomized trials of several recent medical innovations, such as using aspirin to prevent heart attacks ($r = .034$) or using AZT to prevent deaths in AIDS patients ($r = .23$), have been very small to moderate in terms of these norms.

Dichotomization of Scaled Variables

Another means of avoiding unit problems that has become increasingly popular in some research areas such as public health and psychiatry is the dichotomization of scaled variables. Of course, there are times when such a treatment of scales is justified by practical or theoretical considerations (Meehl, 1992). However, all too frequently the major motivation is the provision of an effect size (e.g., in epidemiology, odds or odds ratio, proportion, rate, or risk ratio) that is readily understood even by those who are unfamiliar with the scales employed. The dependence of these statistics on the selected cut-point is often not well recognized. And, of course, the loss of information resulting from dichotomization has long been a concern (J. Cohen, 1983).

Defining a Material Difference in Meaningful Units

Defining a material difference or effect that is large enough to be taken seriously in a research area is an important substantive scientific task, one which has not often been given the attention it deserves. Although these theoretically justified effects may be expressed in terms of standardized units, expression in the form of well understood units that are independent of the characteristics of the sample in a particular investigation is the norm for a well-developed science. Thus, a goal for the behavioral sciences is accrual of information about study outcomes measured in agreed upon and widely used meaningful units that are *independent* of the research design or sampled population. Then, over time, material effect sizes within areas of research can be defined in those units, rather than in units that change as a function of these study characteristics. As an illustration, imagine that we measured body temperature by expert ratings, and that each expert used an idiosyncratic scale but was very accurate. One research reports that two aspirin lowered the temperature ratings of a set of patients with scarlet fever $.2 SD$, and another study reports that two aspirin lowered the temperature of a set of children without known illness $.4 SD$. Do we not have much more information when we know that for both groups the temperature decline was 1 degree, or that for the patients the decline was 2 degrees and for the well group it was $.3$ degrees, both of which value sets expressed in degrees (and an infinite number of others) are potentially consistent with these standardized findings? Thus we see that using *SD* units may not always allow appropriate aggregation over studies, as critics of meta-analysis have noted.

Towards Criteria for Judging the Information Content of a Score

Thus far, we have alluded to the information communicated to researchers by the units of measurement but have not offered any formal criteria for judging the quality of our units in this regard. As a first step towards this end, it is useful to consider what substantive knowledge may depend on how items are scored. We provide four over-arching standards for measures and then propose 7 specific criteria that operationalize those standards.

Standard 1

Commonly used summary statistics on single variables should convey immediate meaning. Means and mean differences should be informative, as should measures of variability such as ranges and standard deviations.

Standard 2

The coefficients by which relationships between variables are expressed should have immediate meaning. Another statistic of general substantive interest is the change in one variable per unit change in another (whether in a linear regression equation, in comparisons of groups, or in some other non-linear equation implied by the theory). In applications based on a well-developed theory the intercept in a regression equation may provide useful scientific information.

Standard 3

Scoring should be free of sample characteristics so that comparisons of scores across populations can be readily accomplished. Given that measures may be applied to alternative populations in different studies (e.g., males and females; clinical and non-clinical), their units should contribute useful information to the comparison of these populations, and to the relationships among variables in these populations.

Standard 4

Comparisons of scores across constructs or across alternative measures of the same construct should be facilitated by the scoring. Given alternative operationalizations of our constructs, a measure's units will be most useful when they contribute useful information, such as difficulty level or general endorsement tendency or potential floor or ceiling effects, to the empirical comparison of conceptually similar measures. Based on these four standards, we propose the following seven criteria for examination of different methods of representing the units of scores on a scale.

1. Can the means and mean differences be substantively interpreted?
2. Can the standard deviation and differences in *SD* be substantively interpreted?
3. Can the range and differences in ranges be substantively interpreted? Note that the range may provide useful information about features of the data such as ceiling/floor effects or scale usage that are not immediately apparent from the *SD*.
4. Can functional relationships between the variables (here expressed as unstandardized coefficients in linear regression equations) be substantively interpreted?
5. Does the regression intercept have a useful meaning? With some units the intercept may represent the predicted value of *Y* at impossible values of *X*, whereas with other units, the intercept may represent a meaningful value.

6. Are the scores independent of sampling error or population differences? If the mean responses or response variability differs from sample to sample or from study to study, do the scores reflect these differences? Or are the scores dependent upon the sample at hand?

7. Do the scores provide potentially useful comparisons of alternative operationalizations of a construct or of related constructs in terms of overall response tendencies?

An Illustration of Scaling Options in Real Data

Earlier in this article, we reviewed several different types of units that can be used in psychological research. Four of these methods can be applied to develop scales for collections of items in a single sample when we do not have meaningful units or population norms: (a) *item sums*; (b) *averaged item scores*; (c) *standardized scores* (i.e., *z scores*); and (d) *POMP scores*. Below we apply each of these scoring methods to a real data set and evaluate the results in terms of the seven criteria outlined above.

The data set we use was collected from a sample of 18 to 27 year olds that has been studied since early childhood (Kogan, Smith, & Jenkins, 1977). The measures we report were gathered in the fourth wave of data collection, at which time the sample was still representative of the population in the upstate New York counties from which it was originally sampled (Cohen & Cohen, 1996).

We obtained responses to one item on overall life satisfaction (LifeSat) on a 3-point response scale from 1 = life is far from what you want it to be, to 3 = life is about what you want it to be. These young adults also responded to 5 items asking about material satisfaction (MatSat; satisfaction with variety of activities in life, with opportunity to travel, with the place you live, with provision for basic needs, and financial adequacy). These items were on a scale from 1 to 4, except for financial adequacy which was on a 1 to 5 scale. Three other items were taken to represent affiliational satisfaction (AffSat; satisfaction with relationship with closest friends, with sex life, both on a scale from 1 to 4, and with involvement in clubs and organizations on a scale 1 to 3). The 1 to 4 scales all range from not satisfied to completely satisfied.

The substantive goals of the study included determination of the contribution of material and affiliative satisfaction to overall life satisfaction, and investigation of the age and gender differences in these variables and in the relationships among them. The regression equation predicting life satisfaction from age, gender, material satisfaction and affiliative satisfaction yielded a multiple $R = .537$; and $F(4, 724) = 73.31$. Because the different methods of scoring the scales discussed below are linear transforms of one another, these statistics are the same for all

analyses, as are *t* tests on individual IVs. Equations including interaction terms are discussed in the next section.

In the present example, each of the four methods was used to develop scale scores for each participant. In order to make possible comparisons in which all scores were *exact* linear transforms of one another, the response alternatives on the item on each IV scale (AffSat, MatSat) with a different response scale was rescaled to a 1-4 range.¹ The response scale for the single DV LifeSat item retained a 1-3 range. (a) *Item sum scores* were created by adding the sum of the positively worded items and subtracting the sum of the negatively worded items; (b) *average item scores* were created by reversing negative items and calculating the mean score across the items comprising the scale; (c) *Standardized scores* (*z* scores) were created based on the mean and standard deviation of the total sample, using the item sum scores as the basis of computing the *z* scores; and (d) *POMP scores* were created from the item sum scores by taking the difference between the score for each subject and the lowest possible score and dividing by the difference between the highest possible score and the lowest possible score *as per Equation 1 on page 323*.

For simplicity in interpreting the results of those regression equations containing interactions (see Aiken & West, 1991), age was measured in years but was centered (i.e., the mean age of the sample was subtracted). The resulting mean of centered *age variable* in the sample is zero. Sex was coded with unweighted effects codes (Cohen & Cohen, 1983, p. 198; West, Aiken, & Krull, 1996), with female = -.5 and male = .5, and labeled "male" to indicate which sex was given the higher code.² The value zero on this variable is the approximate mean of the entire sample. Analyses were based on 729 subjects for whom data are complete.

¹In the case of the affiliative satisfaction (AffSat) scale one item measured on a 3-point scale was transformed by $[(\text{original score} - 1) \times 3/2] + 1$. Similarly, on the material satisfaction (MatSat) scale, one item was measured on a 5-point scale and was transformed by $[(\text{original score} - 1) \times 3/4] + 1$. Our decision to carry out these transformations was a consequence of the demands of average item scoring and our desire to keep all scores strict linear transforms of one another. Most uses of average item scoring would be confined to circumstances in which all items had been responded to on the same scale. We ignore psychometric issues of establishing equivalency here.

²The use of *unweighted* effect coding for sex would have led the predictor sex to have a mean of precisely zero (0.00) if there had been an equal number of males and females. With $n = 365$ females and $n = 364$ males, the mean of the effect coded sex variable is $-.001$.

Descriptive Statistics and Simple Linear Regression Coefficients

Item Sums

Table 1 presents these data as item sums. The variable names in this section are given the subscript *s* to identify them as item sums. *Ms*, *SDs*, and scale ranges convey very little information about how satisfied either group is in any of the three areas. Nor can we easily determine whether the mean difference between male and female respondents, for which the leading digit is less than .1 on each of the satisfaction measures, should be considered material or not. Examination of the *M*, range and *SD* for the total group as well as for males and females separately shows a mean score for material satisfaction (MatSat_s) that is about 25% larger than the mean of affiliative satisfaction (AffSat_s), which, in turn, is more than three

Table 1
Item Sum Scores of Study Scales and Regression Equation

Group	<i>n</i>	Statistic	Scale		
			MatSat _s	AffSat _s	LifeSat
Total Population	735	Mean	11.030	8.978	2.536
		Range	-1 to 15	4 to 12	1 to 3
		<i>SD</i>	2.400	1.585	.622
Males	366	Mean	11.025	8.922	2.489
		Range	-1 to 15	4 to 12	1 to 3
		<i>SD</i>	2.367	1.579	.662
Females	369	Mean	11.036	9.034	2.584
		Range	2 to 15	4 to 12	1 to 3
		<i>SD</i>	2.437	1.594	.576

Regression Equation

$$\text{LifeSat}_s = .120 \text{MatSat}_s + .038 \text{AffSat}_s - .029 \text{age} - .092 \text{male} + .865$$

se^a (.009) (.013) (.007) (.039) (.125)

^a se = standard error of regression coefficient.

times the magnitude of life satisfaction (LifeSat_s). With this method of scoring these differences cannot be interpreted substantively. The minimum observed score for MatSat_s is less than zero, indicating that for at least one person the response on the keyed negative item was higher than the sum of the keyed positive items. The range on LifeSat_s covered the entire range of the (single) item, but the range on AffSat_s is essentially uninformative for this scoring. SDs differed between variables in the same directions as did their Ms , but less extremely. Of course, since these measures had different numbers of items and different item scales, this information is essentially useless.

When we examine the regression model for LifeSat_s , we see that both MatSat_s and AffSat_s contributed independently to the prediction of LifeSat_s , as did age (with a decline in LifeSat_s with age), and male gender (where females were higher on LifeSat_s than were males). The magnitude of the changes in LifeSat_s per unit of the independent variables is, however, not interpretable. The intercept provides the estimate of LifeSat_s when MatSat_s , AffSat_s , age, and gender are all at zero (a score that happens to be impossible for AffSat_s given that the minimum possible score is 4). When we examine our criteria for these scores, we see that we “flunk” criteria 1-5 on meaningful Ms , ranges, and SDs , regression coefficients (B_s), and intercepts (B_0s). Although, sum scores based on the same set of items in different samples can be subjected to a statistical test, the size of the mean difference *per se* does not convey anything about the importance of the difference. Nevertheless, it would have the same meaning regardless of the sample characteristics; that is, a higher M in one study than in another would indicate more of the quality. Thus criterion 6 is minimally passed. Sum scores provide little information that is useful in understanding results across different scales measuring related constructs, such as whether the subjects were more satisfied with their material situation than with their affiliative situation (criterion 7 is failed), or whether one measure of material satisfaction is more “difficult” than another measure.. Thus, as noted in Table 2 (next page), Column 1, most of the seven criteria are flunked by this popular method of scoring.

Average Item Scores

Table 3 provides the same data for scores consisting of average item scores. In this section variables are given the subscript a to indicate they are average item scores. Starting with the ranges of observed scores, we see that for two of the three satisfaction measures individual respondents covered the full range from the minimum (1) to the maximum average score (3 for LifeSat_a ; 4 for MatSat_a and AffSat_a). In all cases the average response was above the midpoint of the scale, and, on the average, respondents were somewhat more satisfied with their material situation (MatSat_a) than their social situation (AffSat_a), a comparison

Table 2

Evaluation of Four Scoring Methods Against Seven Criteria for Scale Interpretability and Utility

Criterion	Scoring Method			
	Item Sums	Average Item Scores	Standardized Scores	POMP Scores
1. Means and mean differences	--	++	+	++
2. Standard deviations (<i>SD</i>) and <i>SD</i> differences	--	++	+	++
3. Ranges and range differences	--	++	+	++
4. Regression coefficients	--	+	++	++
5. Regression intercepts	--	+	--	++
6. Sample and population differences	+	++	--	++
7. Group and scale comparability	-	+	--	++

Note. ++ clearly meets criterion; + marginally meets criterion; -- fails criterion.

that we were able to make because these two scales used the same response scale. We also found that the *SD* of LifeSat_a was larger despite the smaller range of the scale than that of the other measures. The gender difference in overall LifeSat_a was nearly one tenth of a scale point.

The regression coefficients show the effect on LifeSat of a change of 1 average item unit was more than 5 times larger for MatSat_a (.602) than for AffSat_a (.115). Note that these can potentially be compared directly because the same four point scale was used for these two measures; in other circumstances such a comparison would be meaningless.³ There was a .033 scale point decline in LifeSat_a for each year increase in age. The intercept is the estimated score on LifeSat_a for an individual who scored zero on every other predictor. Such a score

³As noted by Cooper and Richardson (1986), Kenny (1979), and others, comparison of standardized or unstandardized regression coefficients involves consideration of several psychometric criteria regarding the equivalence of the independent variables. Among these are whether each independent variable has a comparable reliability, comparable scale units, provides comparable sampling of each underlying content domain, and represents the full range of the construct.

Table 3
Average Item Scores of Study Scales and Regression Equation

Group	<i>n</i>	Statistic	Scale		
			MatSat _a	AffSat _a	LifeSat
Total Population	729	Mean	3.206	2.993	2.536
		Range	1 to 4	1.33 to 4	1 to 3
		<i>SD</i>	.480	.529	.622
Males	364	Mean	3.205	2.974	2.489
		Range	1 to 4	1.33 to 4	1 to 3
		<i>SD</i>	.473	.526	.662
Females	365	Mean	3.207	3.011	2.584
		Range	1.55 to 4	1.33 to 4	1 to 3
		<i>SD</i>	.487	.531	.576

Regression Equation

$$\text{LifeSat}_a = .602 \text{MatSat}_a + .115 \text{AffSat}_a - .029 \text{age} - .092 \text{male} + .263$$

se^a (.043) (.039) (.007) (.039) (.149)

^a *se* = standard error of regression coefficient.

is outside the range of possible values for MatSat_a and AffSat_a. Note that, if the bottom scale points had been coded at zero, the intercept would have been directly interpretable as the predicted score for a respondent of average age who had scored zero on both MatSat_a and AffSat_a.

In terms of our 7 criteria, average item scores fare better than item sums, given that the same scale range is used for all items for averaging (see footnote 1). The mean and standard deviation of average item scores are interpretable in the original item units, as is the range (criteria 1, 2, and 3 are clearly passed). The regression coefficients (e.g., .602 for MatSat_a) can be substantively interpreted only if the reader can keep in mind the meaning of the units of both independent and dependent variable, which may be different, as they are here. Thus, we now know that a one unit increase on the four point MatSat_a scale was associated with a .602 unit increase on the three point LifeSat_a scale. The inherent difficulties in

interpreting this lead us to judge that average item scores only minimally pass criterion 4. If the value of zero is a meaningful value on each predictor, then the regression intercept is potentially interpretable as well, but if it is not (as here) it is not, so that we also give average item scores a minimally passing grade on criterion 5. Average item scores have the same interpretation in different samples or populations, which may be usefully compared in terms of differences on the original item scale (criterion 6 is clearly passed). Different scales can be usefully compared and regression coefficients can be compared among scales, but only under the proviso that they are scored with the same number of scale points, so that we give average item scores a minimally passing grade on criterion 7. Thus, as summarized in Table 2, Column II, average item scores are a clear improvement over sum scores, but still have interpretational deficiencies as suggested by the minimally passing grades on criteria 4, 5, and 7.

Standardized Scores (z Scores)

Table 4 presents the same data for which each of the three satisfaction scales has been standardized using the M and SD of the total sample. In this section variables are given the subscript z to indicate that they are z scores. The M s and SD s of the total sample on the satisfaction scales provide no information, being fixed at 0 and 1 respectively by the method. However, we can make a comparison of gender differences on the different scales, and see that while the scores on $MatSat_z$ hardly differed, men were $.07 SD$ units lower on $AffSat_z$ and $.15 SD$ units lower on $LifeSat_z$. Furthermore, it appears that men were a little more variable in $LifeSat_z$ scores than were women. The asymmetry in the range suggests that the scores may be somewhat negatively skewed, although this interpretation must be made cautiously because the range depends only on the two most extreme scores. Given the absence of utility for the total sample M s and SD s, and the presence of useful comparisons between subgroups on these same statistics, z scores only minimally pass criteria 1-3.

Turning to the regression coefficients, we see that $MatSat_z$ was a more influential predictor of $LifeSat_z$ than was $AffSat_z$ (.465 versus .097). Note that although this comparison is meaningful for both item average scores and z scores, it does not have the same numerical value or meaning across the two scoring schemes. The z scores are expressed in SD units, which are not equal in raw units for $MatSat_z$ and $AffSat_z$. In contrast the average score units for $MatSat_a$ and $AffSat_a$ are on comparable 4-point scales.⁴ $LifeSat_z$ declined about $.05 SD$ for each year of age; $LifeSat_z$ was also about $.15 SD$ units lower for men than women. These equations also show that for each SD unit increase in $MatSat_z$ there was

⁴POMP scores and z scores each create an equivalency of the meaning of scale units across scales, of meaningful, but different kinds. (see footnote 3 for other psychometric criteria).

Table 4
Standardized Scores (z scores) on Study Scales and Regression Equation

Group	n	Statistic	Scale		
			MatSat _z	AffSat _z	LifeSat _z
Total Population	735	Mean	0	0	0
		Range	-4.60 to 1.65	-3.14 to 1.90	-2.47 to .75
		SD	1.000	1.000	1.000
Males	366	Mean	-0.002	-.036	-.076
		Range	-4.60 to 1.65	-3.14 to 1.90	-2.47 to .75
		SD	.986	.995	1.065
Females	369	Mean	0.002	.035	.076
		Range	-3.45 to 1.65	-3.14 to 1.90	-2.47 to .75
		SD	1.015	1.005	.926

Regression Equation

$$\text{LifeSat}_z = .465 \text{MatSat}_z + .097 \text{AffSat}_z - .047 \text{age} - .148 \text{male} + .000$$

se^a (.033) (.033) (.012) (.063) (.031)

^a se = standard error of regression coefficient.

nearly a one half *SD* unit increase in LifeSat_z. The intercept is the *M* of LifeSat_z when all predictors are zero, that is for the full sample *M* on age, MatSat_z and AffSat_z; thus for these centered and standardized variables it is necessarily zero (see footnote 2), and thus not informative.

Applying our seven criteria, there was substantially more information in the *z* transformed scores than in the item sums. Although *z* scores fix the *M*s and *SD*s of the total sample at 0 and 1, respectively, differences in *M*s, *SD*s, and ranges between subsamples could be meaningfully evaluated so that criteria 1-3 are only minimally passed. The magnitude of regression coefficients was meaningful, both individually and in comparisons with one another (criterion 4 is clearly passed). However, the intercept did not add information (criterion 5 is failed), and the estimates are all dependent on the adequacy or comparability of the standardization in comparisons with other samples (criterion 6 is failed). *z* scored

scales cannot be usefully compared with one another with regard to their (fixed) M s and SD s (criterion 7 is failed), although different measures of a presumed constant construct can be compared with regard to regression coefficients, providing the samples are the same or comparable.

On the whole, we have some loss and some gain in information when we compare standardized scores to average item scores (see Table 2, columns II versus III). With average item scores M s and SD s are potentially interpretable in themselves, and not just in subgroup comparisons as they were in standardized scores. Provided that they are scored on the same response scale, different scales can be usefully compared for average item but not standardized scores. Regression coefficients are likely to convey much more information for z scores. They can also be compared among scales with the same number of scale points scored as item averages. However, because different scales will have different SD s, the meaning of equivalence is not the same for average item and standardized scores. Providing that a zero scale point is within the range, the intercept is potentially interpretable for average item scores, but provides no new information for standardized scores. Average item scores can be compared for different measures if they use the same response scales, unlike z scores.

POMP Scores

Table 5 provides the same information using POMP scores for the same data. Variables in this section are given the subscript p to identify them as POMP scores. Now the fact that the observed scores cover the full potential range (0 to 100) for two of the three measures is even more obvious. We also note that all three satisfaction measures had an average response somewhere in the vicinity of 75% of the maximum possible score, but also that $AffSat_p$ in particular was lower than $MatSat_p$. In fact, the information provided in these scores led us to examine the question of whether those mean differences were statistically significant. The comparisons of $MatSat_p$ with $AffSat_p$, $MatSat_p$ with $LifeSat_p$, and $AffSat_p$ with $LifeSat_p$ yielded t s of 7.15, 3.72, and 7.50, respectively, all statistically significant given our large sample size. Note that examination of the difference between $MatSat_m$ and $AffSat_m$ is also potentially meaningful for Average item scores, but that the comparisons with $LifeSat_m$ are not, because of the different scale. Whether these differences are interpreted for Average or POMP scores, we have at least two available interpretations, namely that young Americans are less satisfied with affiliative aspects of their lives than with material aspects or, alternatively, that we have written our items assessing affiliative satisfaction in such a way as to be intrinsically more difficult than the material satisfaction items. The reasonableness of one or another interpretation would be logically buttressed by psychometric and

Table 5
Percent of Maximum (POMP) Scores and Regression Equation

Group	<i>n</i>	Statistic	Scale		
			MatSat _{<i>p</i>}	AffSat _{<i>p</i>}	LifeSat _{<i>p</i>}
Total Population	735	Mean	73.534	66.422	76.818
		Range	0 to 100	11 to 100	0 to 100
		<i>SD</i>	16.002	17.627	31.080
Males	366	Mean	73.498	65.796	74.451
		Range	0 to 100	11 to 100	0 to 100
		<i>SD</i>	15.777	15.777	33.087
Females	369	Mean	73.571	67.047	79.178
		Range	0 to 100	11 to 100	0 to 100
		<i>SD</i>	16.245	17.710	28.792

Regression Equation

$$\text{LifeSat}_p = .903 \text{MatSat}_p + .172 \text{AffSat}_p - 1.468 \text{age} - 4.611 \text{male} - 1.021$$

se^a (.065) (.059) (.361) (1.950) (5.129)

^a *se* = standard error of regression coefficient.

normative data that goes beyond the discussion here.⁵ We found that the *SD* of LifeSat_{*p*} was larger relative to the range than that of the other measures, probably because it consisted of a single item (Epstein, 1983). The male-female difference in LifeSat_{*p*} was about 5 percent of the potential range of the scale.

The regression coefficients tell us that for each percentage point increase in MatSat_{*p*} there was an average .9 percentage point increase in LifeSat_{*p*}. The effect of AffSat_{*p*} on LifeSat_{*p*} was about one-fifth as large. (Note that despite the fact that both of these scales are linear transforms of the *z* scores, they are *different* linear transforms, so this ratio is not the same.) The intercept tells us that the estimated LifeSat_{*p*} for an individual who gave the lowest possible scores on MatSat_{*p*} and

⁵ Equality of means and *SD*s similarly cannot be assumed to show that two scales can be considered "equal", an assertion that needs to be established by other psychometric methods. Nevertheless, comparisons of means and *SD*s provide useful, if limited, information about the meaning and implications of each scale as compared to certain others.

AffSat_p was -1.021 , not significantly different from zero. Although the score of zero on AffSat_p is outside the range of scores actually observed in this sample, this prediction is potentially meaningful. It suggests that total dissatisfaction with material and affiliative aspects of one's life implies total dissatisfaction with one's life as a whole. In addition, POMP scores can be compared in different samples from the same population or different populations. POMP score statistics are intrinsically meaningful, and thus related scales can be usefully compared. Of course, these comparisons necessarily have a particular and necessarily limited meaning; that is, we are comparing the POMP units, and not any other function of the item responses. With this understanding we can say that with POMP scoring we have clearly met all 7 criteria for interpretability.

Regression Models With Interactions

In this study a major interest was in whether the influence of material and affiliative satisfaction on LifeSat was equivalent for males and females, that is, whether the satisfaction predictors interacted with sex. Upon testing we found that there was no evidence of a differential slope of LifeSat on AffSat by sex; however, there was a statistically significant difference for the interaction of Male with MatSat. These findings are displayed for each scoring method in Table 6. Each coefficient in these equations is a comfortable multiple of its standard error and is therefore statistically significant. Because each of the scores is a linear transformation of the others, the t tests of these coefficients are identical. Again we ask which method provides the most information, and whether we can conclude that this difference is "material".

In these equations, we have followed procedures proposed by Aiken and West (1991) and West, Aiken, and Krull (1996) in specifying regression models containing interactions between continuous and categorical variables. We have centered all IVs on the sample M . In this scaling, the intercept equals the LifeSat M in the total sample. Since this number adds no new information, we have omitted it in the table.

With item sum scores we see that the adjusted slope for MatSat_s was $.033$ higher for males than for females, making it $.137$ for males [$.121 + .033(.5)$] and $.104$ for females [$.121 + .033(-.5)$]. These units are not interpretable since we do not know the meaning of a 1 unit change on the scale; however, it is clear that the difference between the slopes for males and females ($.033$) is about $1/4$ ($.27 = .033/.121$) of the size of their average slope ($.121$). This ratio will also be the same for average item scores ($.167/.605 = .27$), for z scores ($.129/.468 = .27$) and for POMP scores ($.250/.908 = .27$), within rounding error.

In addition, with POMP scoring, we may note that the difference in slopes ($.250$) is one quarter of a percentage point, and may feel justified in saying that

Table 6
Regression Prediction with Interactions for each Method of Scoring

Method of Variable Scaling	Regression Equation
1. Item sum scores	$\text{LifeSat} = .121 \text{ MatSat}_s + .038 \text{ AffSat}_s - .029 \text{ age} - .092 \text{ sex} + .033 \text{ MatSat}_s \times \text{sex}$
2. Average item scores	$\text{LifeSat} = .605 \text{ MatSat}_a + .113 \text{ AffSat}_a - .029 \text{ age} - .092 \text{ sex} + .167 \text{ MatSat}_a \times \text{sex}$
3. Standardized (z) scores	$\text{LifeSat}_z = .468 \text{ MatSat}_z + .096 \text{ AffSat}_z - .047 \text{ age} - .148 \text{ sex} + .129 \text{ MatSat}_z \times \text{sex}$
4. POMP scores	$\text{LifeSat}_p = .908 \text{ MatSat}_p + .169 \text{ AffSat}_p - 1.464 \text{ age} - 4.614 \text{ sex} + .250 \text{ MatSat}_p \times \text{sex}$
<i>t</i> values (all analyses)	14.008 2.874 4.068 2.372 2.057

Note: In all analyses MatSat and AffSat are centered; sex is coded male = .5, female = -.5.

MatSat_p contributed powerfully to LifeSat_p in males, LifeSat_p increasing about one percentage point with each percentage point increase in MatSat_p [$1.033 = .908 + .250(.5)$], age and AffSat_p held constant. For females, the increase was about three quarters of a percentage point for each percentage point increase in material satisfaction [$.783 = .908 + .250(-.5)$]. This difference may be then placed in the perspective of findings from other studies, such as sex differences in the effects of other variables (similarly POMP scored) and other effects on life satisfaction (again, within the POMP score framework).

Summary

This example illustrates how linear transformations of conventional summed scores convey additional information to scale users and other scientists. The z scores facilitated subgroup comparisons within a study, but lost information about the entire sample's positions on the scales, and are dependent on the adequacy of sample statistics. Average item scales retain a useful scale metric and are population independent, but can present obstacles to interpretation of regression coefficients and intercepts when the IV and DV have different units or when the low end of the scale is not zero. They also limit useful comparisons if all items do not have the same response scale or when different IVs are on different response scales. POMP scales appeared to provide more information than the other linear transformations we considered, satisfying each of our seven criteria for judging the information content of a scale.

Nonlinear Transformations

Good contemporary data analytic practice often requires statistical transformation of a variable in order to improve its distributional properties and the validity of our statistical inferences (Cook & Weisberg, 1994; Mosteller & Tukey, 1977). For example, Cleveland (1993) illustrates how the relationship between the body weight and brain weight of animals is much clearer following a logarithmic transformation of both variables than when both variables are expressed in their original units (grams). However, as this example also illustrates, even when the original scale is expressed in meaningful units, the interpretation of the results becomes problematic: The means, SDs, ranges, and regression coefficients now reflect the transformed (log) units rather than the original units. Under these circumstances average item scores are also not readily interpretable. However, if scores are transformed and then POMP or z scores are created based on the values of the transformed scores, researchers are provided with a familiar basis with which to interpret the results *in the new metric*.

Some Pros, Cons, and Variations on the Theme of POMP Scores

As we illustrated in the previous section, POMP scores appear to have several advantages relative to other commonly used scoring methods in communicating the meaning of results to other investigators. They are not sample or population dependent and also potentially allow comparisons across different measures of the same construct. As a simple linear transformation of other commonly used scales, they can be easily calculated so long as the maximum and minimum possible scores on the scale are known.

POMP scores may help investigators to develop an understanding of what a meaningful effect is in their research area. Research findings can be put into a common, easily understandable metric and cumulated, leading to a basis for establishing a consensus on the meaning of a material effect in that research area. For example, in laboratory research on cognitive influences of mood, a ten unit POMP score change in reaction to an experimental manipulation may be deemed to be a large effect. However, if we were to evaluate the effects of a major life experience, a ten unit POMP score change on the same mood measure might be considered small. Given the same or similar measures, we can also potentially compare across research areas, for example, concluding that the major life experiences are more important determinants of mood than a particular experimental manipulation. Note that z scores and other standardized measures (e.g., d , r) take the research context for granted by accomplishing different transforms of the original units in the different contexts.

POMP scales foster consideration of certain scale characteristics because of the importance of the lower (0) and upper (100) criteria in their calculation. In the context of Likert-type scales, cognitive psychologists (e.g., Tversky & Kahneman, 1974) have shown the important influences of the labeling of the scale endpoints on the participants' responses. The presence of the end point effect indicates that it behooves investigators to attend to data points regardless of the scoring method, preferably providing reference points that will help guide theoretical and empirical investigations of the scale constructs. One of the long-range issues to be addressed is the definition of a true zero on a construct. As noted, this is likely to require a feedback of theory and empirical investigation. POMP scoring should aid such investigations by improving the information conveyed by scale statistics.

There are cases in which it may be useful to define 100% as some score that is less than the potential maximum on the scale. This situation will most frequently occur when an official category is defined on the basis of numerical values or counts. To cite two examples, family income is often converted to percent of poverty level, which exceeds 100 for most American families. The American Psychiatric Association's Diagnostic and Statistical Manual (DSM IV, 1994)

defines criteria (symptoms) for each diagnosis and the minimum number required before a diagnosis can be made. One may translate the observed number of symptoms for each case into the percent of the number required to meet diagnostic criteria, thus, the potential score may exceed 100%.

One of the advantages of POMP scores is that they provide a quick method of comparing an original reference scale with revisions of that scale. If, for example, an investigator uses a shortened or otherwise revised version of a scale, one might wish to detect whether the “difficulty” of the measure (the POMP score M) was thus altered. Such an investigation can also proceed with Average item scores, providing that the response scales are the same in the original and revised versions. In addition, the internal consistency of a scale (actually the average inter-item correlation) will have a predictable effect on the variance of the scores, and changes in numbers of items will require adjustments for this difference before the resulting POMP scales can be considered strictly comparable. Another useful task for the future is the establishment of implications of POMP scoring for the communication of the results of studies using modern scaling methods such as item response theory.

Scales typically used in research today will often discriminate best in the direction for which there is the greatest distance between the mean and an endpoint. At the extreme, we will see “floor” and “ceiling” effects. In the satisfaction scales examined here we presume that we probably discriminated better among the less satisfied, as suggested both by the distance between the means (of around 75%) and zero, and also by the negative skew in the scores. In clinical scales such as the Beck Depression Inventory, clinical cutoffs occur at a point that is only about 20 on a POMP scale, and the general population mean is about 12 on a POMP scale (calculated from data provided by Beck & Beamesderfer, 1974). The designing and selection of scales in research should take into account these purposes when devising or choosing a scale with a POMP score mean around 25, 50, or 75, for example.

Information from POMP scoring can sometimes be used to suggest possible questions worthy for further psychometric study. For example, suppose we measure stress, depression measure A , and depression measure B on clinical and non-clinical samples of participants, using POMP scales. Using a raw-unit regression coefficient as a measure of effect size, suppose stress relates more to depression measure A than it does to depression measure B in a sample of clinical participants but more to depression measure B than to depression measure A in a sample of non-clinical participants. This finding leads us to examine the characteristics of the POMP scores for A and B (which, of course, we should have done earlier). We find that measure A has M s of 40 and 10 for the clinical and normal samples, respectively, whereas measure B has POMP M s of 55 and 50 for clinical and normal samples, respectively. These observations would suggest

both measurement and substantive hypotheses worthy of further study. In terms of measurement, measure *A* may show better discrimination at the high end of the scale, whereas measure *B* shows better discrimination at the low end of the scale. In terms of substance, depression may have different relationships with stress in the normal and depressed populations. Depression may increase modestly as stress increases in the normal population, whereas depression may increase rapidly once a threshold is reached and the “balance is tipped” in the depressed population. Further psychometric and substantive research that investigated these hypotheses, made more apparent by the POMP scores, could have considerable utility for substantive researchers.

When tests are designed for maximum discrimination in the general population, items distributed symmetrically about their mid-points are optimal. However, this discrimination fails to consider whether such a symmetrical, centered distribution is consistent with the theory on which the construct is based, which may posit a mean closer to one end or the other. Current scale scoring practices have led us to ignore this information. Thus, a potential scientific advantage of POMP scoring is that when two scales of the same construct have very different means in the same population, there are new aspects of the measures that invite study. These differences in “item difficulty” have usually been attended to in the ability area, but often ignored in other substantive fields. Thus, POMP scoring may assist in the integration of findings using conventional scales and those based on IRT methods of construction.

POMP scaling may ultimately facilitate operations such as the determination of theoretically useful confidence intervals on effect size estimates. For example, we may decide that our theory suggests a relationship in the range of .2 to .3 percentage points increase in *Y* per percentage point increase in *X*. Deciding that confidence limits of .15 and .35 would be needed to produce a sufficient degree of precision, we may proceed to determine the necessary sample size. Such a focus on the precision of estimates, rather than on the simple ability to determine the direction of the effect, which is the focus of most current statistical power analyses, will contribute to, and be required by, the more precise theoretical statements that signal a more developed science.

Alternative Methods of Enhancing the Meaning of Scores

The advocacy of POMP scoring here should not suggest that it is an alternative to be preferred to more substantively grounded scale units. Nor is it our belief that measures of effect size in standardized units will not continue to provide essential scientific information, particularly in those areas of basic psychology where very similar samples are consistently used across studies. Other methods of adding useful meaning to scores may also be used in

combination with POMP scores or the original scales. For example, Sechrest, McKnight, and McKnight (1996) have urged that measures of more abstract constructs be calibrated against behavioral and other measures that are in better understood units, such as time or frequency of relevant behaviors. Ozer (1993) has considered the use of classic just noticeable difference measures in personality assessment. In many cases, scores on scales on which relevant population norms are available are more usefully expressed in terms of those norms. Thus centile or *T* scoring may be more useful than POMP, *z*, item sum, or average item scores. And, of course, in those circumstances in which there is already a consensual meaning for units, such as IQ scores, it is unnecessary to start over by employing a new linear transform.

In sum, in many areas of psychology and other behavioral sciences we are not yet ready to settle on one best standard operationalization of each construct we employ in our theories and research. One negative effect of this is our current inability to cumulate evidence about the meaning of scale units for our theoretical constructs, and the potential variations of meaning in different contexts. This lack of standard measures has led to a general inattention to the scale units of the measures that we employ, to the detriment of our science. The use of POMP scores can improve our ability to communicate useful and easily understandable information using a simple linear transformation of the original scale units. POMP scores can be easily calculated from existing measures and thus do not require that we abandon existing measures. At the same time, POMP scores are not a panacea: They suffer from any psychometric problems such as unreliability or poor sampling of the content domain that affect the original measure. But, because POMP scores put the results into such an easily understood metric, similarities and differences in the findings both within and between studies are highlighted. Differences in findings lead researchers to focus on differences between characteristics of the studies that may otherwise be overlooked. Given a mean POMP score of 75 on depression scale 1 and a mean POMP score of 50 on depression scale 2 measured on the same group of patients, we would be led to take a strong look at the item difficulties and the content of the two scales. A mean POMP score of 50 in patient sample 1 and a mean POMP score of 75 in patient sample 2 on the same depression scale would lead us to further explorations of the populations (e.g., inpatient vs. outpatient) from which each sample was drawn. And, finding a mean POMP score of 50 on depression scale 1 in sample 1 and a mean POMP score of 75 on depression scale 2 in sample 2 would alert us to the need to explore whether measurement issues, population differences, or both account for these obtained differences. Thus, the additional information conveyed by POMP scores may help to move us toward a more cumulative science and the development of more standard scales by facilitating the comparison of alternative measures of our constructs and sharpening our focus on the units of measurement.

References

- Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park CA: Sage.
- American Journal of Epidemiology (1997). *Instructions to authors*. Vol. 45, 727.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (DSM4)*. Washington, DC: American Psychiatric Association.
- American Psychological Association (1996). Initial report of the Task Force on Statistical Inference, December 15, 1996. (Available from the Science Directorate, American Psychological Association, 750 First Street, NE. Washington, DC 20002-4242).
- Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review*, 77, 153-170.
- Beck, A. T. & Beamesderfer, A. (1974). Assessment of depression: The depression inventory. In P. Pichot (Ed.) *Psychological measurement in psychopharmacology: Modern problems in pharmacopsychiatry* (Vol. 7, pp. 151-169). Basel, Switzerland: Karger.
- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart press.
- Cliff, N. (1993). What is and isn't measurement. In G. Keren & C. Lewis (eds) *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 59-63). Mahwah, NJ: Erlbaum.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal & Social Psychology*, 65, 145-153.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7, 249-253.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* 2nd ed., Mahwah, NJ, Erlbaum.
- Cohen, J. (1994). The earth is round, $p < .05$. *American Psychologist*, 49, 997-1003.
- Cohen, J. & Cohen, P. (1983). *Multiple regression/correlation for the behavioral sciences (third edition)*. Mahwah, NJ, Erlbaum.
- Cohen, P. & Cohen, J. (1996). *Adolescent life values and mental health*. Mahwah, NJ: Erlbaum.
- Cook, R. D., & Weisberg, S. (1994). *An introduction to regression graphics*. New York: Wiley.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis issues for field settings*. Chicago: Rand McNally.
- Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology*, 71, 179-184.
- Epstein, S. (1983). Aggregation and beyond: Some basic issues in the prediction of behavior. *Journal of Personality*, 51, 360-392.
- Falmagne, J. (1992). Measurement theory and the research psychologist. *Psychological Science*, 3, 88-93.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston: Kluwer-Nijhoff.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.
- Kogan L. S., Smith, J. & Jenkins, S. (1977). Ecological validity of indicator data predictors of survey findings. *Journal of Social Service Research*, 1, 117-134.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mandel, J. (1964). *The statistical analysis of experimental data*. New York, Dover.
- Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66 (Monograph Suppl. 1-V66), 195-244.

P. Cohen, J. Cohen, L. Aiken and S. West

- Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60*, 117-174.
- Menand, L. (1996). How to make a Ph.D. matter. *The New York Times Magazine*, Sept. 23, p 78-81.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.
- Middleton, W.E.K. (1966). *A history of the thermometer and its use in meteorology*. Baltimore: Johns Hopkins Press.
- Mosteller, F., & Tukey, J.W. (1977). *Data analysis and regression, : A second course in statistics*. Reading, MA: Addison-Wesley.
- Ozer, D. J. (1993). Classical psychophysics and the assessment of agreement and accuracy in judgments of personality. *Journal of Personality, 61*, 739-767.
- Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775-777.
- Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.
- Sechrest, L. (1985). Social science and social policy: Will our numbers ever be good enough? In R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 63-95). Beverly Hills, CA: Sage.
- Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist, 51*, 1065-1071.
- Shavelson, R.J. & Webb. N.,M. (1991). *Generalizability theory: A primer*, Newbury Park, CA: Sage.
- Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.
- Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 33*, 1-67.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.
- West, S. G., Aiken, L. S., & Krull, J.L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality, 64*, 1-48.

Accepted September, 1998.