Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement
2017, Vol. 49, No. 2, 97–99

0008-400X/17/$12.00 http://dx.doi.org/10.1037/cbs0000070

False Precision: The Ring of Truth

William L. Roberts
Kamloops, British Columbia, Canada

Articles on best practices in research usually focus on collecting and analysing data. However, an important ethical and practical issue is often ignored: false precision. Researchers, reviewers, and editors often ignore the precision of instruments and the concept of significant digits, familiar from introductory courses in many sciences. The result is that findings are presented so that they appear to be more precise or accurate than they actually are. Imprecision is also ignored (and precision implied) when results are presented without margins of error (confidence intervals). Other practices also increase or mask imprecision. It is not widely appreciated that imprecision is inflated when scale scores are calculated by summing items, a common practice for clinical instruments. The use of global scores can mask the complex, multidimensional nature of constructs such as stress, resilience, and depression. Although these practices are not intended to mislead or deceive, that is their effect when presented to policymakers, clients—and ourselves. Improvements are obvious: Reported results should reflect the precision of measurements; margins of error should be reported; scale scores should be calculated by averaging, not summing, items; and unidimensional scales should be used in research articles instead of global scores.

Keywords: false precision, significant digits

> "Falsehoods all,
> but he gave his falsehoods all the ring of truth.
> —Homer, *Odyssey*, XIX, 235, trans. Fagles (Homer, 1996)

Articles on best practices in psychological research (e.g., Bakeman et al., 2006; Osborne, 2008a, 2013; Wilkinson & the Task Force on Statistical Inference, 1999) usually focus on collecting and analysing data. However, there is an important ethical and practical issue that is often ignored: the issue of false precision. Presenting results so that they appear to be more precise or accurate than they actually are is a common practice in psychology. It even occurs in examples in the *APA Publication Manual* (American Psychological Association, 2010, e.g., Tables 5.10, 5.11, and 5.12). Imprecision is also ignored (and precision implied) when results are presented without margins of error (confidence intervals). Although these practices are not intended to mislead or deceive, that is their effect when results are presented to policymakers, clients—and ourselves.

False precision when reporting results is associated with several other issues. There is a level of imprecision inherent in common instruments and in their use; it is not widely appreciated that imprecision is increased by summing items when calculating scale scores; and finally there is the conceptual imprecision of discussing global, multidimensional constructs as if they had clear, unitary meanings. Each of these will be considered below.

False Precision When Reporting Results

The concept of significant digits is familiar from introductory courses in chemistry, biology, or physics (e.g., http://chemwiki.ucdavis.edu/Core/Analytical_Chemistry/Quantifying_Nature/Significant_Digits). It applies in psychology as well. As noted on p. 137 of the *APA Publication Manual*, reported results should reflect the precision of the measurements made. Thus, when constructs are measured on integer scales (as they are on questionnaires), we are only entitled to report means and standard deviations to one decimal place (as group statistics, they are more stable than individual scores). Correlations, standardized regression coefficients, factor loadings, and Cronbach alphas should be reported to only two decimal places, that is, to two significant digits, matching the number of significant digits for the mean and standard deviation.

Yet the issue of "digits reported" often seems to be treated as a formatting convention, not as an issue of precision. As noted earlier, the *APA Publication Manual* itself contains examples of questionnaire data presented to three significant digits (i.e., to two decimal places, e.g., pp. 136, 157). Occasionally, published questionnaire data have been presented to six significant digits (five decimal places, e.g., Connor & Davidson, 2003). But when constructs are measured on integer scales, it is simply incorrect to suggest that they have been measured accurately to several decimal places.[1]

Margins of error

The problem of precision is also ignored when results are presented without margins of error (confidence intervals). This not only has consequences for theory, it can have important practical (and therefore ethical) consequences, too, when test scores are used as cut scores for clinical or other practical purposes. There is

Correspondence concerning this article should be addressed to William L. Roberts, Independent Scholar, 1869 Robson Lane, Kamloops, BC, V2E 1X5, Canada. E-mail: wroberts@tru.ca

[1] The fact that statistical software reports results to many digits is not a justification, because the software knows nothing of the precision of the original measurements and so cannot round results appropriately. That is the responsibility of the researcher.

a danger that cut scores will be treated rigidly and individual scores misinterpreted as precise and stable, problems long recognised with IQ scores (e.g., Canivez, 2014) and personnel selection tests (e.g., Campion et al., 2001). Cut scores, as sample statistics, have margins of error; margins of error for individual scores are even more substantial (e.g., Kellow & Willson, 2008). Better practice is obvious: Cut scores should be presented as a range or band of scores, reflecting their margins of error (as argued by Campion et al., 2001), and clinical instruments should make clear the margin of error associated with individual scores.

Problems in the Precision of Measurements

Respondents are Imprecise

Approximate as they are, it is widely recognised that the precision of questionnaire data is normally compromised in many ways (see Osborne, 2013, for a discussion). Individuals respond in terms of their own understanding of vague self-report scale anchors (such as "often" or "somewhat characteristic") and inaccurately to more precise anchors (such as "three or four times a week"). As Osborne points out, respondents may also be biased, careless, or intent on manipulating their image. Weathers, Sharma, and Niedrich (2005) present evidence that questionnaire responses are influenced by individual cognitive characteristics as well as the number of response categories. These are difficult problems to address, but this is all the more reason for recognising the imprecision of obtained data.

Instruments are Imprecise

It is a truism in methodology courses that every instrument tells us less than we want and more than we like. That is, every instrument provides only a partial reflection of the construct it assesses and at the same time incorporates error and bias. These are basic conceptual reasons for caution in interpreting findings and for the explicit recognition (using confidence intervals) that scores are approximate. They are also the reasons that the need for data that converge across methods and sources has long been recommended (e.g., Cook & Campbell, 1979).

Although instruments are necessarily imprecise, certain practices make matters better or worse. Measurement requires that units be defined and consistent across individuals and situations. Thus on a rating scale, categories need to be well defined as a necessary prelude to accuracy and comparability across respondents. The need for clear definition limits the number of response categories that can be used, and in fact fewer than 10 response categories are usually recommended for a variety of reasons (e.g., Preston & Colman, 2000; Weathers et al., 2005).

Rating scales from 0 to 100 do not meet these fundamental criteria. What is the evidence that people can make 101 meaningful distinctions in a given construct? What is the evidence that there is a consistent frame of reference across participants, so that responses can be compared? Most fundamentally, what does a unit change on such a scale mean? In research in which judges or observers rate participants (e.g., Block, 1983; Sroufe, Egeland, Carlson, & Collins, 2005), researchers are not only concerned with rank-order stability (correlations across observers) but also with the stability of the scores themselves (assessed by repeated measures). Both dimensions are involved in assessing reliability and change (Block, 1983). Discrepancies in absolute scores are often resolved by discussion, a process important for training observers, clarifying the meaning of rating scales, and insuring their accurate use. Such a process is impossible for disagreements on a 0 to 100 scale, since the units themselves are undefined. Thus, although 101-point rating scales are intuitively appealing to some, and their strong test–retest correlations (e.g., Preston & Colman, 2000) give them a "ring of truth," they give a false impression of precision. They should not be used.

False Precision in the Calculation of Scale Scores

It is not widely appreciated that imprecision increases when scale scores are calculated by summing questionnaire items. This is a common practice, especially for clinical instruments (such as the Beck Depression Inventory; Beck, Steer, & Brown, 1996), but it is misleading. Conceptually, one cannot go from a few crude distinctions (such as 0 = *never*, 1 = *sometimes*, 2 = *always*) to precise, meaningful distinctions simply by adding together crude items. This is the issue of significant digits again, in a different context.

Some argue that summing items yields useful ordinal data: A higher total score means "more." This argument acknowledges that such a score, detached from the framework of the original item ratings, has no meaning beyond rank; but as a rank, it is more unstable (less replicable) than the items from which it was derived. This occurs because the variance of a summed score is equal to the sum of the variances of the items plus an additional amount that increases as the correlations between the items increase (e.g., Howell, 2002)—and items in a scale should be substantially correlated with one another. Summed scores are *far* more unstable than the individual scores from which they are derived.

For both reasons, it is misleading and inaccurate to write as if we can make 20 or 60 (or more) meaningful, stable distinctions in, say, reported depression or resilience. In fact we can make only a few distinctions, based on the original item ratings. Averaging items—which has its own problems—at least makes it clear that we are in the framework of the original ratings, not at some (false) higher level of precision. For example, when averaging 21 items rated 0 to 3, there are 31 possible scores (0, 0.1, 0.2, . . ., 2.9, 3.0), in contrast to 64 possible summed scores (0, 1, . . ., 63). Rumpelstiltskin could spin straw into gold, but psychologists cannot. In lieu of more sophisticated methods, scale scores should be calculated by averaging.

False Precision in Concepts

False precision occurs when global scores for multidimensional constructs (such as aggression, depression, resilience, stress, or intelligence) are presented and discussed as if they had precise meaning. Although we use such constructs in ordinary conversation, global scores are uninterpretable, precisely because they are multidimensional. For this reason, global scores are usually inappropriate in a research article, where we seek to increase our understanding of causes, consequences, and relations between constructs. Understanding requires that multidimensional instruments and concepts be "unpacked" into unidimensional scales and constructs.

Although global scores may be helpful on a practical level, when one has to make a multidimensional judgment (Does this person fit a diagnostic category? Should this child receive a treatment or a service?), there are ethical dangers in ignoring both imprecision of measurement and imprecision of meaning. As noted earlier, one needs to be aware of margins of error, both in global cut scores and in the individual scores compared to them (e.g., Campion et al., 2001).

Recommendations

In conclusion, the need for measures converging across methods and sources has long been recognised (e.g., Cook & Campbell, 1979). Converging measures allow greater precision, clarity of meaning, and reliability. When reporting results, converging or single, we should respect significant digits and also report margins of error. Our measurements should be as good as we can make them. Questionnaires should have a moderate number of well-defined categories that have consistent meaning across respondents (see Block, 2008, for a technique that has these characteristics). In lieu of more sophisticated methods, scale scores should be calculated by averaging. Scale scores need to have clear meanings, and to be interpretable, a scale must be unidimensional. "Evidence-based decisions are only as good as the evidence they are based on" (Osborne, 2008b, p. x). If we are not thinking seriously about the precision and meaning of our measurements, we cannot think seriously about the complex phenomena we seek to understand.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Bakeman, R., Gottman, J., Brewer, D., Bub, K., Burchinal, M., Graham, F., & McCartney, K. (Eds.). (2006). *Best practices in quantitative methods for developmentalists: Vol. 71. Number 3*. London, England: Wiley-Blackwell.

Beck, A., Steer, R., & Brown, G. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Block, J. (1983). *Lives through time*. Mahwah, NJ: Erlbaum.

Block, J. (2008). *The Q-Sort in character appraisal*. Washington, DC: American Psychological Association.

Campion, M., Outtz, J., Zedeck, S., Schmidt, F., Kehoe, J., Murphy, K., & Guion, R. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology, 54,* 149–185. http://dx.doi.org/10.1111/j.1744-6570.2001.tb00090.x

Canivez, G. L. (2014). *Test review of Wechsler Preschool and Primary Scale of Intelligence*. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The nineteenth mental measurements yearbook* (4th ed., pp. 732–737). Lincoln, NE: Buros Institute of Mental Measurements.

Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: The Connor–Davidson Resilience Scale (CD-RISC). *Depression and Anxiety, 18,* 76–82. http://dx.doi.org/10.1002/da.10113

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago, IL: Rand McNally.

Homer. (1996). *The odyssey* (R. Fagles, Trans.). London, England: Penguin Books.

Howell, D. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Kellow, J., & Willson, V. (2008). Setting standards and establishing cut scores on criterion-referenced assessments: Some technical and practical considerations. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 15–28). Los Angeles, CA: Sage.

Osborne, J. (Ed.). (2008a). *Best practices in quantitative methods*. Los Angeles, CA: Sage. http://dx.doi.org/10.4135/9781412995627

Osborne, J. (2008b). Using best practices is a moral and ethical obligation. In J. Osborne, (Ed.), *Best practices in quantitative methods* (pp. ix–xi). Los Angeles, CA: Sage.

Osborne, J. (2013). *Best Practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Los Angeles, CA: Sage. http://dx.doi.org/10.4135/9781452269948

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104,* 1–15. http://dx.doi.org/10.1016/S0001-6918(99)00050-5

Sroufe, L. A., Egeland, B., Carlson, E., & Collins, W. (2005). *The development of the person: The Minnesota Study of Risk and Adaptation From Birth to Adulthood*. New York, NY: Guilford Press.

Weathers, D., Sharma, S., & Niedrich, R. (2005). The impact of the number of scale points, dispositional factors, and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research, 58,* 1516–1524. http://dx.doi.org/10.1016/j.jbusres.2004.08.002

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. http://dx.doi.org/10.1037/0003-066X.54.8.594

Received January 19, 2017
Revision received February 13, 2017
Accepted February 13, 2017 ■