Change

Thomas R. Knapp

Introduction

Mary spelled correctly 3 words out of 6 on Monday and 5 words out of 6 on Wednesday. How should we measure the change in her performance?

Several years ago Cronbach and Furby (1970) argued that we shouldn't; i.e., we don't even need the concept of change. An extreme position? Of course, but read their article sometime and see what you think about it.

Why not just subtract the 3 from the 5 and get a change of two words? That's what most people would do. Or how about subtracting the percentage equivalents, 50% from 83.3%, and get a change of 33.3%? But...might it not be better to divide the 5 by the 3 and get 1.67, i.e., a change of 67%? [Something that starts out simple can get complicated very fast.]

Does the context matter? What went on between Monday and Wednesday? Was she part of a study in which some experimental treatment designed to improve spelling ability was administered? Or did she just get two days older?

Would it matter if the 3 were her attitude toward spelling on Monday and the 5 were her attitude toward spelling on Wednesday, both on a five-point Likert-type scale, where 1=hate, 2=dislike, 3=no opinion, 4=like, and 5=love?

Would it matter if it were only one word, e.g., antidisestablishmentarianism, and she spelled it incorrectly on Monday but spelled it correctly on Wednesday?

These problems regarding change are illustrative of what now follows.

A little history

Interest in the concept of change and its measurement dates back at least as long ago as Davies (1900). But it wasn't until much later, with the publication of the book edited by Harris (1963), that researchers in the social sciences started to debate the advantages and the disadvantages of various ways of measuring change. Thereafter hundreds of articles were written on the topic, including many of the sources cited in this paper.

"Gain scores"

The above example of Mary's difference of two words is what educators and psychologists call a "gain score", with the Time 1 score subtracted from the Time

2 score. [If the difference is negative it's a loss, rather than a gain, but I've never heard the term "loss scores".] Such scores have been at the heart of one of the most heated controversies in the measurement literature. Why?

1. The two scores might not be on exactly the same scale. It is possible that her score of 3 out of 6 was on Form A of the spelling test and her score of 5 out of 6 was on Form B of the spelling test, with Form B consisting of different words, and the two forms were not perfectly comparable (equivalent, "parallel"). It might even have been desirable to use different forms on the two occasions, in order to reduce practice effect or mere "parroting back" at Time 2 of the spellings (correct or incorrect) at Time 1.

2. Mary herself and/or some other characteristics of the spelling test might have changed between Monday and Wednesday, especially if there were some sort of intervention between the two days. In order to get a "pure" measure of the change in her performance we need to assume that both of the testing conditions were the same. In a randomized experiment all bets regarding the direct relevance of classical test theory should be off if there is a pretest and a posttest to serve as indicators of a treatment effect, because the experimental treatment could affect the posttest mean AND the posttest variance AND the posttest reliability AND the correlation between pretest and posttest.

3. Gain scores are said by some measurement experts (e.g., O'Connor, 1972; Linn & Slinde, 1977; Humphreys, 1996) to be very unreliable, and by other measurement experts (e.g., Zimmerman & Williams, 1982; Williams & Zimmerman, 1996; Collins, 1996) to not be. Like the debate concerning the use of traditional interval-level statistics for ordinal scales, this controversy is unlikely ever to be resolved. I got myself embroiled in it many years ago (see Knapp, 1980; Williams & Zimmerman, 1984; Knapp, 1984). [I also got myself involved in the ordinal vs. interval controversy (Knapp, 1990, 1993).]

The problem is that if the instrument used to measure spelling ability (Were the words dictated? Was it a multiple-choice test of the discrimination between the correct spelling and one or more incorrect spellings?) is unreliable, Mary's "true score" on both Monday and Wednesday might have been 4 (she "deserved" a 4 both times), and the 3 and the 5 were both measurement errors attributable to "chance", and the difference of two words was not a true gain at all.

<u>Some other attempts at measuring change</u>

Given that gain scores might not be the best way to measure change, there have been numerous suggestions for improving things. In the Introduction (see above) I already mentioned the possibility of dividing the second score by the first score rather than subtracting the first score from the second score. This has never caught on, for some good reasons and some not-so-good reasons. The strongest arguments against dividing instead of subtracting are: (1) it only makes

sense for ratio scales (a 5 for "love" divided by a 3 for "no opinion" is bizarre, for instance); and (2) if the score in the denominator is zero, the quotient is undefined. [If you are unfamiliar with the distinctions among nominal, ordinal, interval, and ratio scales, read the classic article by Stevens (1946).] The strongest argument in favor of the use of quotients rather than differences is that the measurement error could be smaller. See, for example, the manual by Bell (1999) regarding measurement uncertainty and how the uncertainty "propagates" via subtraction and division. It is available free of charge on the internet.

Other methodologists have advocated the use of "modified" change scores (raw change divided by possible change) or "residualized" change (the actual score at Time 2 minus the Time 2 score that is predicted from the Time 1 score in the regression of Time 2 score on Time 1 score). Both of these, and other variations on simple change, are beyond the scope of the present paper, but I have summarized some of their features in my reliability book (Knapp, 2013).

The measurement of change in the physical sciences vs. the social sciences

Some physical scientists wonder what the fuss is all about. If you're interested in John's weight of 250 pounds in January of one year and his weight of 200 pounds in January of the following year, for example, nothing other than subtracting the 250 from the 200 to get a loss of 50 pounds makes any sense, does it? Well, yes and no. You could still have the problem of scale difference (the scale in the doctor's office at Time 1 and the scale in John's home at Time 2?) and the problem of whether the raw change (the 50 pounds) is the best way to operationalize the change. Losing 50 pounds from 250 to 200 in a year is one thing, and might actually be beneficial. Losing 50 pounds from 150 to 100 in a year is something else, and might be disastrous. [I recently lost ten pounds from 150 to 140 and I was very concerned. (I am 5'11" tall.) I have since gained back five of those pounds, but am still not at my desired "fighting weight", so to speak.]

Measuring change using ordinal scales

I pointed out above that it wouldn't make sense to get the ratio of a second ordinal measure to a first ordinal measure in order to measure change from Time 1 to Time 2. It's equally wrong to take the difference, but people do it all the time. Wakita, Ueshima, & Noguchi (2012) even wrote a long article devoted to the matter of the influence of the number of scale categories on the psychological distances between the categories of a Likert-type scale. In their article concerned with the comparison of the arithmetic means of two groups using an ordinal scale, Marcus-Roberts and Roberts (1987) showed that Group I's mean could be higher than Group II's mean on the original version of an ordinal scale, but Group II's mean could be higher than Group I's mean on a perfectly defensible transformation of the scale points from the original version to another version. (They used as an example a grading scale of 1, 2, 3, 4, and 5 vs. a

grading scale of 30, 40, 65, 75, and 100.) The matter of subtraction is meaningless for ordinal measurement.

Measuring change using dichotomies

Dichotomies such as male & female, yes & no, and right & wrong play a special role in science in general and statistics in particular. The numbers 1 and 0 are most often used to denote the two categories of a dichotomy. Variables treated that way are called "dummy" variables. For example, we might "code" male=1 and female =0 (not male); yes=1 and no=0 (not yes); and right=1 and wrong=0 (not right). As far as change is considered, the only permutations of 1 and 0 on two measuring occasions are (1,1), e.g., right both times; (1,0), e g., right at Time 1 and wrong at Time 2; (0,1), e.g., wrong at Time 1 and right at Time 2; and (0,0), e.g., wrong both times. The same permutations are also the only possibilities for a yes,no dichotomy. There are even fewer possibilities for the male, female variable, but sex change is well beyond the scope of this paper!

Covariance F vs. gain score t

For a pretest & posttest randomized experiment, Cronbach and Furby (1970) suggested the use of the analysis of covariance rather than a t test of the mean gain in the experimental group vs. the mean gain in the control group as one way of avoiding the concept of change. The research question becomes "What is the effect of the treatment on the posttest over and above what is predictable from the pretest?" as opposed to "What is the effect of the treatment on the change from pretest to posttest?" In our recent paper, Bill Schafer and I (Knapp & Schafer, 2009) actually provided a way to convert from one analysis to the other.

Measurement error

In the foregoing sections I have made occasional references to measurement error that might produce an obtained score that is different from the true score. Are measurement errors inevitable? If so, how are they best handled? In an interesting article (his presidential address to the National Council on Measurement in Education), Kane (2011) pointed out that in everyday situations such as sports results (e.g., a golfer shooting a 72 on one day and a 69 on the next day; a baseball team losing one day and winning the next day), we don't worry about measurement error. (Did the golfer deserve a 70 on both occasions? Did the baseball team possibly deserve to win the first time and lose the second time?). Perhaps we ought to.

What we should do

That brings me to share with you what I think we should do about measuring change:

1. Start by setting up two columns. Column A is headed Time 1 and Column B is headed Time 2. [Sounds like a Chinese menu.]

2. Enter the data of concern in the appropriate columns, with the maximum possible score (not the maximum obtained score) on both occasions at the top and the rest of the scores listed in lockstep order beneath. For Mary's spelling test scores, the 3 would go in Column A and the 5 would go in Column B. For n people who attempted to spell antidisestablishmentarianism on two occasions, all of the 1's would be entered first, followed by all of the 0's, in the respective columns.

3. Draw lines connecting score in Column A with the corresponding score in Column B for each person. There would be only one (diagonal) line for Mary's 3 and her 5. For the n people trying to spell antidisestablishmentarianism, there would be n lines, some (perhaps all; perhaps none) horizontal, some (perhaps all; perhaps none) diagonal. If all of the lines are horizontal, there is no change for anyone. If all of the lines are diagonal and crossed, there is a lot of change going on. See Figure 1 for a hypothetical example of change from pretest to posttest for 18 people, almost all of whom changed from Time1 to Time 2 (only one of the lines is horizontal). I am grateful to Dave Kenny for permission to reprint that diagram, which is Figure 1.7 in the book co-authored by Campbell and Kenny (1999). [A similar figure, Figure 3-11 in Stanley (1964), antedated the figure in Campbell & Kenny. He (Stanley) was interested in the relative relationship between two variables, and not in change per se. He referred to parallel lines, whether horizontal or not, as indicative of perfect correlation.]

Ties are always a problem (there are several ties in Figure 1, some at Time 1 and some at Time 2), especially when connecting a dichotomous observation (1 or 0) at Time 1 with a dichotomous observation at Time 2 and there are lots of ties. The best way to cope with this is to impose some sort of arbitrary (but not capricious) ordering of the tied observations, e.g., by I.D. number. In Figure 1, for instance, there is no particular reason for the two people tied at a score of 18 at Time 1 to have the line going to the score of 17 at Time 2 be above the line going to the score of 15 at Time 2. [It doesn't really matter in this case, because they both changed, one "losing" one point and the other "losing" two points.]

4. Either quit right there and interpret the results accordingly (Figure 1 is actually an excellent "descriptive statistic" for summarizing the change from pretest to posttest for those 18 people) or proceed to the next step.

5. Calculate an over-all measure of change. What measure? Aye, there's the rub. Intuitively it should be a function of the number of horizontal lines and the extent to which the lines cross. For ordinal and interval measurements the slant of the diagonal lines might also be of interest (with lines slanting upward indicative of "gain" and with lines slanting downward indicative of "loss"). But what function? Let me take a stab at it, using the data in Figure 1:

The percentage of horizontal lines (no change) in that figure is equal to 1 out of 18, or 5.6%. [Unless your eyes are better than mine, it's a bit hard to find the horizontal line for the 15th person, who "went" from 13 to 13, but there it is.] The percentage of upward slanting lines (gains), if I've counted correctly, is equal to 6 out of 18, or 33.3%. The percentage of downward slanting lines (losses) is equal to 11 out of 18, or 61.1%. A person who cares about over-all change for this dataset, and for most such datasets, is likely to be interested in one or more of those percentages. [I love percentages (see Knapp, 2010).]

Statistical inference from sample to population

Up to now I've said nothing about sampling (people, items, etc.). You have to have a defensible statistic before you can determine its sampling distribution and, in turn, talk about significance tests or confidence intervals. If the statistic is a percentage, its sampling distribution (binomial) is well known, as is its approximation (normal) for large samples and for sample percentages that are not close to either 0 or 100. The formulas for testing hypotheses about population percentages and for getting confidence intervals for population percentages are usually expressed in terms of proportions rather than percentages, but the conversion from percentage to proportion is easy (drop the % sign and move the decimal point two places to the left). Caution: concentrate on only one percentage. For the Campbell and Kenny data, for instance, don't test hypotheses for all of the 5.6%, the 33.3%, and the 61.1%, since that would be redundant (they are not independent; they add to 100).

If you wanted to go a little further, you could carry out McNemar's (1947) test of the statistical significance of dichotomous change, which involves setting up a 2x2 contingency table and concentrating on the frequencies in the "off-diagonal" (1,0) and (0,1) cells, where, for example, (1,0) indicates a change from yes to no, and (0,1) indicates a change from no to yes. But I wouldn't bother. Any significance test or any confidence interval assumes that the sample has been drawn at random, and you know how rare that is!

Some closing remarks, and a few more references

I'm with Cronbach and Furby. Forget about the various methods for measuring change that have been suggested by various people. But if you would like to find out more about what some experts say about the measurement of change, I recommend the article by Rogosa, Brandt, and Zimowski (1982), which reads

very well [if you avoid some of the complicated mathematics]; and the book by Hedeker and Gibbons (2006). That book was cited in an interesting May 10, 2007 post on the Daily Kos website entitled "Statistics 101: Measuring change".

Most of the research on the measurement of change has been devoted to the determination of whether or not, or to what extent, change has taken place. There are a few researchers, however, who turn the problem around by claiming in certain situations that change HAS taken place and the problem is to determine if a particular measuring instrument is "sensitive", or "responsive", or has the capacity to detect such change. If you care about that (I don't), you might want to read the letter to the editor of Physical Therapy by Fritz (1999), the response to that letter, and/or some of the articles cited in the exchange.

References

Bell, S. (1999). A beginner's guide to uncertainty of measurement. National Physical Laboratory, Teddington, Middlesex, United Kingdom, TW11 0LW.

Campbell, D.T., & Kenny, D.A. (1999). A primer on regression artifacts. New York: Guilford.

Collins, L.M. (1996). Is reliability obsolete? A commentary on "Are simple gain scores obsolete?". Applied Psychological Measurement, 20, 289-292.

Cronbach, L.J., & Furby, L. (1970). How we should measure "change"...Or should we? Psychological Bulletin, 74, 68-80.

Davies, A.E. (1900). The concept of change. The Philosophical Review, 9, 502-517.

Fritz, J.M. (1999). Sensitivity to change. Physical Therapy, 79, 420-422.

Harris, C. W. (Ed.) (1963). Problems in measuring change. Madison, WI: University of Wisconsin Press.

Hedeker, D., & Gibbons, R.D. (2006) Longitudinal data analysis. Hoboken, NJ: Wiley.

Humphreys, L. (1996). Linear dependence of gain scores on their components imposes constraints on their use and interpretation: A commentary on "Are simple gain scores obsolete?". Applied Psychological Measurement, 20, 293-294.

Kane, M. (2011). The errors of our ways. Journal of Educational Measurement, 48, 12-30.

Knapp, T.R. (1980). The (un)reliability of change scores in counseling research. Measurement and Evaluation in Guidance, 11, 149-157.

Knapp, T.R. (1984). A response to Williams and Zimmerman. Measurement and Evaluation in Guidance, 16, 183-184.

Knapp, T.R. (1990). Treating ordinal scales as interval scales. Nursing Research, 39, 121-123.

Knapp, T.R. (1993). Treating ordinal scales as ordinal scales. Nursing Research, 42, 184-186.

Knapp, T.R. (2010). Percentages: The most useful statistics ever invented. Unpublished monograph. Available free of charge at www.tomswebpage.net.

Knapp, T.R. (2013). The reliability of measuring instruments. Unpublished monograph. Available free of charge at www.tomswebpage.net.

Knapp, T.R., & Schafer, W.D. (2009). From gain score t to ANCOVA F (and vice versa). Practical Assessment, Research, and Evaluation (PARE), 14 (6).

Linn, R.L., & Slinde, J.A. (1977). The determination of the significance of change between pretesting and posttesting periods. Review of Educational Research, 47, 121-150.

Marcus-Roberts, H., & Roberts, F. (1987). Meaningless statistics. Journal of Educational Statistics, 12, 383-394.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157.

O'Connor, E.F., Jr. (1972). Extending classical test theory to the measurement of change. Review of Educational Research, 42, 73-97.

Rogosa, D.R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 90, 726-748.

Stanley, J.C. (1964). Measurement in today's schools (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Stevens, S.S. (1946). On the theory of scales of measurement. Science, 103, 677-680.

Wakita,T., Ueshima,N., & Noguchi, H. (2012). Psychological distance between categories in the Likert Scale: Comparing different numbers of options. Educational and Psychological Measurement, 72, 533-546.

Williams, R.H., & Zimmerman, D.W. (1984). A critique of Knapp's "The (un)reliability of change scores in counseling research". Measurement and Evaluation in Guidance, 16, 179-182.

Williams, R.H., & Zimmerman, D.W. (1996) Are simple gain scores obsolete? Applied Psychological Measurement, 20, 59-69.

Zimmerman, D.W., & Williams, R.H. (1982). Gain scores can be highly reliable. Journal of Educational Measurement, 19, 149-154.

Figure 1. "Change" from pretest to posttest for 18 people