Assessing Validity and Reliability of Likert and Visual Analog(-ue) Scales

Thomas R. Knapp

Introduction

Consider the following scales for measuring pain:

It hurts:	Strongly disagree	Disagree	Can't tell	Agree	Strongly agree
	(1)	(2)	(3)	(4)	(5)

How bad is the pain?: _____
no pain excruciating

How much would you be willing to pay in order to alleviate the pain?_____

The first two examples, or slight variations thereof, are used a lot in research on pain. The third is not. In what follows I would like to discuss how one might go about assessing (testing, determining) the validity and the reliability of measuring instruments of the first kind (a traditional Likert Scale [LS]) and measuring instruments of the second kind (a traditional Visual Analog Scale [VAS]) for measuring the presence or severity of pain and for measuring some other constructs. I will close the paper with a few brief remarks regarding the third example and how its validity and reliability might be assessed.

The sequence of steps

1. Although you might not agree, I think you should start out by addressing content validity (expert judgment, if you will) as you contemplate how you would like to measure pain (or attitude toward legalizing marijuana, or whatever the construct of interest might be). If a Likert-type scale seems to make sense to you, do the pain experts also think so? If they do, how many scale points should you have? Five, as in the above example, and as was the case for the original scale developed by Rensis Likert (1932)? Why an odd number such as five? In order to provide a "neutral", or "no opinion" choice? Might not too many respondents cop out by selecting that choice? Shouldn't you have an even number of scale points (how about just two?) so that respondents have to take a stand one way or the other?

The same sorts of considerations hold for the "more continuous" VAS, originally developed by Freyd (1923). (He called it a Graphic Rating Scale. Unlike Likert, his name was not attached to it by subsequent users. Sad.) How long should it be? (100 millimeters is conventional.) How should the endpoints read? Should there be intermediate descriptors underneath the scale between the two

endpoints? Should it be presented to the respondents horizontally (as above) or vertically? Why might that matter?

2. After you are reasonably satisfied with your choice of scale type (LS or VAS) and its specific properties, you should carry out some sort of pilot study in which you gather evidence regarding feasibility (how willing and capable are subjects to respond?), "face" validity (does it appear to them to be measuring pain, attitude toward marijuana, or whatever?), and tentative reliability (administer it twice to the same sample of people, with a small amount of time in-between administrations, say 5 minutes or thereabouts). This step is crucial in order to "get the bugs out" of the instrument before its further use. But the actual results, e.g., whether the pilot subjects express high pain or low pain, favorable attitudes or unfavorable attitudes, etc., should be of little or no interest, and certainly do not warrant publication.

3. If and when any revisions are made on the basis of the pilot study, the next step is the most difficult. It entails getting hard data regarding the reliability and/or the validity of the LS or the VAS. For a random sample drawn from the same population from which a sample will be drawn in the main study, a formal test-retest assessment should be carried out (again with a short interval between test and retest), and if there exists an instrument that serves as a "gold standard" it should also be administered and the results compared with the scale that is under consideration.

Likert Scales

As far as the reliability of a LS is concerned, you might be interested in evidence for either or both of the scale's "relative reliability" and its "absolute reliability". The former is more conventional; just get the correlation between score at Time 1 and score at Time 2. Ah, but what particular correlation? The Pearson product-moment correlation coefficient? Probably not; it is appropriate only for interval-level scales. (The LS is an ordinal scale.) You could construct a cxc contingency table, where c is the number of categories (scale points) and see if most of the frequencies lie in the upper-right and lower-left portions of the table. That would require a large number of respondents if c is more than 3 or so, in order to "fill up" the c^2 cells; otherwise the table would look rather anemic. If further summary of the results is thought to be necessary, either Guttman's (1946) reliability coefficient or Goodman and Kruskal's (1979) gamma (sometimes called the index of order association) would be good choices for such a table, and would serve as the reliability coefficient (for that sample on that occasion). If the number of observations is fairly small and c is fairly large, you could calculate the Spearman rank correlation between score at Time 1 and score at Time 2, since you shouldn't have too many ties, which can often wreak havoc.

[Exercise for the reader: When using the Spearman rank correlation in determining the relationship between two ordinal variables X and Y, we get the difference between the rank on X and the rank on Y for each observation. For ordinal variables in general, subtraction is a "no-no". (You can't subtract a "strongly agree" from an "undecided", for example.) Shouldn't a rank-difference also be a "no-no"? I think it should, but people do it all the time, especially when they're concerned about whether or not a particular variable is continuous enough, linear enough, or normal enough in order for the Pearson r to be defensible.]

The matter of absolute reliability is easier to assess. Just calculate the % agreement between score at Time 1 and score at Time 2.

If there is a gold standard to which you would like to compare the scale under consideration, the (relative) correlation between scale and standard (a validity coefficient) needs to be calculated. The choice of type of validity coefficient, like the choice of type of reliability coefficient, is difficult. It all depends upon the scale type of the standard. If it is also ordinal, with d scale points, a cxd table would display the data nicely, and Goodman and Kruskal's gamma could serve as the validity coefficient (again, for that sample on that occasion). (N.B.: If a gold standard does exist, serious thought should be given to forgoing the new instrument entirely, unless the LS or VAS under consideration would be briefer but equally reliable and content valid.)

Visual Analog Scales

The process for the assessment of the reliability and validity of a VAS is essentially the same as that for a LS. As indicated above, the principal difference between the two is that a VAS is "more continuous" than a LS, but neither possesses a meaningful unit of measurement. For a VAS there is a surrogate unit of measurement (usually the millimeter), but it wouldn't make any sense to say that a particular patient has X millimeters of pain. (Would it?) For a LS you can't even say 1 what or 2 what,..., since there isn't a surrogate unit.

Having to treat a VAS as an ordinal scale is admittedly disappointing, particularly if it necessitates slicing up the scale into two or more (but not 101) pieces and losing some potentially important information. But let's face it. Most respondents will probably concentrate on the verbal descriptors along the bottom of the scale anyhow, so why not help them along? (If there are no descriptors except for the endpoints, you might consider collapsing the scale into those two categories.)

Statistical inference

For the sample selected for the LS or VAS reliability and validity study, should you carry out a significance test for the reliability coefficient and the validity coefficient? Certainly not a traditional test of the null hypothesis of a zero

relationship. Whether or not a reliability or a validity coefficient is significantly greater than zero is not the point (they darn well better be). You might want to test a "null" hypothesis of a specific non-zero relationship (e.g., one that has been found for some relevant norm group), but the better analysis strategy would be to put a confidence interval around the sample reliability coefficient and the sample validity coefficient. (If you have a non-random sample it should be treated just like a population, i.e., descriptive statistics only.)

The article by Kraemer (1975) explains how to test a hypothesis about, and how to construct a confidence interval for, the Spearman rank correlation coefficient, rho. A similar article by Woods (2007; corrected in 2008) treats estimation for both Spearman's rho and Goodman and Kruskal's gamma. That would take care of Likert Scales nicely. If the raw data for Visual Analog Scales are converted into either ranks or ordered categories, inferences regarding their reliability and validity coefficients could be handled in the same manner.

Combining scores on Likert Scales and Visual Analog Scales

The preceding discussion was concerned with a single-item LS or VAS. Many researchers are interested in combining scores on two or more of such scales in order to get a "total score". (Some people argue that it is also important to distinguish between a Likert *item* and a Likert *scale*, with the latter consisting of a composite of two or more of the former. I disagree; a single Likert item is itself a scale; so is a single VAS.) The problems involved in assessing the validity and reliability of such scores are several magnitudes more difficult than for assessing the validity and reliability of a single LS or a single VAS.

Consider first the case of two Likert-type items, e.g., the following:

The use of marijuana for non-medicinal purposes is widespread.

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

The use of marijuana for non-medicinal purposes should be legalized.

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
(1)	(2)	(3)	(4)	(5)

All combinations of responses are possible and undoubtedly likely. A respondent could disagree, for example, that such use is widespread, but agree that it should be legalized. Another respondent might agree that such use is widespread, but disagree that is should be legalized. How to combine the responses to those two items in order to get a total score? See next paragraph. (Note: Some people, e.g., some "conservative" statisticians, would argue that scores on those two items should never be combined; they should always be analyzed as two separate items.)

The usual way the scores are combined is to merely add the score on Item 1 to the score on Item 2, and in the process of so doing to "reverse score", if and when necessary, so that "high" total scores are indicative of an over-all favorable attitude and "low" total scores are indicative of an over-all unfavorable attitude. The respondent who chose "2" (disagree) for Item 1 and "4" (agree) for Item 2 would get a total score of 4 (i.e., a "reversed" 2) + 4 (i.e., a "regular" 4) = 8, since he(she) appears to hold a generally favorable attitude toward marijuana use. But would you like to treat that respondent the same as a respondent who chose "5" for the first item and "3" for the second item? They both would get a total score of 8. See how complicated this is? Hold on; it gets even worse!

Suppose you now have total scores for all respondents. How do you summarize the data? The usual way is to start by making a frequency distribution of those total scores. That should be fairly straightforward. Scores can range from 2 to 10, whether or not there is any reverse-scoring (do you see why?), so an "ungrouped" frequency distribution should give you a pretty good idea of what's going on. But if you want to summarize the data even further, e.g., by getting measures of central tendency, variability, skewness, and kurtosis, you have some tough choices to make. For example, is it the mean, the median, or the mode that is the most appropriate measure of central tendency for such data? The mean is the most conventional, but should be reserved for interval scales and for scales that have an actual unit of measurement. (Individual Likert scales and combinations of Likert scales are neither: Ordinal in, ordinal out.) The median should therefore be fine, although with an even number of respondents that can get tricky (for example, would you really like to report a median of something like 6.5 for this marijuana example?).

Getting an indication of the variability of those total scores is unbelievably technically complicated. Both variance and standard deviation should be ruled out because of non-intervality. (If you insist on one or both of those, what do you use in the denominator of the formula... n or n-1?) How about the range (the actual range, not the possible range)? No, because of the same non-intervality property. All other measures of variability that involve subtraction are also ruled out. That leaves "eyeballing" the frequency distribution for variability, which is not a bad idea, come to think of it.

I won't even get into problems involved in assessing skewness and kurtosis, which should probably be restricted to interval-level variables in any event. (You can "eyeball" the frequency distribution for those characteristics just like you can for variability, which also isn't a bad idea.)

The disadvantages of combining scores on two VASs are the same as those for combining scores on two LSs. And for three or more items things don't get any better.

What some others have to say about the validity and the reliability of a LS or VAS

The foregoing (do you know the difference between "forgoing" and "foregoing"?) discussion consists largely of my own personal opinions. (You probably already have me pegged, correctly, as a "conservative" statistician.) Before I turn to my most controversial suggestion of replacing almost all Likert Scales and almost all Visual Analog Scales with interval scales, I would like to call your attention to authors who have written about how to assess the reliability and/or the validity of a LS or a VAS, or who have reported their reliabilities or validities in substantive investigations. Some of their views are similar to mine. Others are diametrically opposed.

1. Aitken (1969)

According to Google, this "old" article has been cited 1196 times! It's that good, and has a brief but excellent section on the reliability and validity of a VAS. (But it is very hard to get a hold of. Thank God for helpful librarians like Shirley Ricker at the University of Rochester.)

2. Price, et al. (1983).

As the title of their article indicates, Price, et al. claim that in their study they have found the VAS to be not only valid for measuring pain but also a ratio-level variable. (I don't agree. But read the article and see what you think.)

3. Wewers and Lowe (1990)

This is a very nice summary of just about everything you might want to know concerning the VAS, written by two of my former colleagues at Ohio State (Mary Ellen Wewers and Nancy Lowe). There are fine sections on assessing the reliability and the validity of a VAS. They don't care much for the test-retest approach to the assessment of the reliability of a VAS, but I think that is really the only option. The parallel forms approach is not viable (what constitutes a parallel item to a given single-item VAS?) and things like Cronbach's alpha are no good because they require multiple items that are gathered together in a composite. It comes down to a matter of the amount of time between test and retest. It must be short enough so that the construct being measured hasn't changed, but it must be long enough so that the respondents don't merely "parrot back" at Time 2 whatever they indicated at Time 1; i.e., it must be a "Goldilocks" interval.

4. Von Korff, et al. (1993)

These authors developed what they call a "Quadruple Visual Analog Scale" for measuring pain. It consists of four items, each having "No pain " and "worst possible pain" as the two endpoints, with the numbers 0 through 10 equally spaced beneath each item. The respondents are asked to indicate the amount of

pain (1) now, (2) typical, (3) best, and (4) worst; and then to add across the four items. Interesting, but wrong (in my opinion).

5. Bijur, Silver, and Gallagher (2001)

This article was a report of an actual test-retest (and re-retest...) reliability study of the VAS for measuring acute pain. Respondents were asked to record their pain levels in pairs one minute apart thirty times in a two-hour period. The authors found the VAS to be highly reliable. (Not surprising. If I were asked 60 times in two hours to indicate how much pain I had, I would pick a spot on the VAS and keep repeating it, just to get rid of the researchers!)

6. Owen and Froman (2005)

Although the main purpose of their article was to dissuade researchers from unnecessarily collapsing a continuous scale (especially age) into two or more discrete categories, the authors made some interesting comments regarding Likert Scales. Here are a couple of them:

"...equal appearing interval measurements (e.g., Likert-type scales...)" (p. 496)

"There is little improvement to be gained from trying to increase the response format from seven or nine options to, say, 100. Individual items usually lack adequate reliability, and widening the response format gives an appearance of greater precision, but in truth does not boost the item's reliability... However, when individual items are aggregated to a total (sum or mean) scale score, the continuous score that results usually delivers far greater precision." (p. 499)

A Likert scale might be an "equal appearing interval measurement", but it's not interval-level. And I agree with the first part of the second quote (it sounds like a dig at Visual Analog Scales), but not with the second part. Adding across ordinal items does not result in a defensible continuous score. As the old adage goes, "you can't make a silk purse out of a sow's ear".

7. Davey, et al. (2007)

There is a misconception in the measurement literature that a single item is necessarily unreliable and invalid. Not so, as Davey, et al. found in their use of a one-item LS and a one-item VAS to measure anxiety. Both were found to be reliable and valid. (Nice study.)

8. Hawker, et al. (2011)

This article is a general review of pain scales in general. The first part of the article is devoted to the VAS (which the authors call "a continuous scale"; ouch!). They have this to say about its reliability and validity:

"Reliability. Test–retest reliability has been shown to be good, but higher among literate (r = 0.94, P< 0.001) than illiterate patients (r= 0.71, P < 0.001) before and after attending a rheumatology outpatient clinic [citation].

Validity. In the absence of a gold standard for pain, criterion validity cannot be evaluated. For construct validity, in patients with a variety of rheumatic diseases, the pain VAS has been shown to be highly correlated with a 5-point verbal descriptive scale (“nil,” “mild,” “moderate,”“severe,” and “very severe”) and a numeric rating scale (with response options from “no pain” to “unbearable pain”), with correlations ranging from 0.71–0.78 and.0.62–0.91, respectively) [citation]. The correlation between vertical and horizontal orientations of the VAS is 0.99 [citation] " (page s241)

That's a lot of information packed into two short paragraphs. One study doesn't make for a thorough evaluation of the reliability of a VAS; and as I have indicated above, those significance tests aren't appropriate. The claim about the absence of a gold standard is probably warranted. But I find a correlation of .99 between a vertical VAS and a horizontal VAS hard to believe. (Same people at the same sitting? You can look up the reference if you care.)

9. Vautier (2011)

Although it starts out with some fine comments about basic considerations for the use of the VAS, Vautier's article is a very technical discussion of multiple Visual Analog Scales used for the determination of reliability and construct validity in the measurement of change. The references that are cited are excellent.

10. Franchignoni, Salaffi, and Tesio (2012)

This recent article is a very negative critique of the VAS. Example: "The VAS appears to be a very simple metric ruler, but in fact it's not a true linear ruler from either a pragmatic or a theoretical standpoint. " (page 798). (Right on!) In a couple of indirect references to validity, the authors go on to argue that most people can't discriminate among the 101 possible points for a VAS. They cite Miller's (1956) famous 7 + or - 2 rule), and they compare the VAS unfavorably with a 7-point Likert scale.

<u>Are Likert Scales and Visual Analog Scales really different from one another?</u>

In the previous paragraph I referred to 101 points for a VAS and 7 points for an LS. The two approaches differ methodologically only in the number of points (choices, categories) from which a respondent makes a selection. There are Visual Analog Scales that aren't really visual, and there are Likert Scales that are very visual. An example of the former is the second scale at the beginning of this paper. The only thing "visual" about that is the 100-millimeter line. As examples of the latter, consider the pictorial Oucher (Beyer, et al., 2005) and the pictorial

Defense and Veterans Pain Rating Scale (Pain Management Task Force, 2010) which consist of actual pictures of faces of children (Beyer) or drawings of faces of soldiers (Pain Management Task Force) expressing varying degrees of pain. . Both instruments are actually amalgams of Likert-type scales and Visual Analog Scales, since they also have 0-10 scales juxtaposed near the faces.

I once had the pleasant experience of co-authoring an article about the Oucher with Judy Beyer. (Our article is cited in theirs.) The instrument now exists in forms for each of four ethnic groups (African-American, Caucasian, Hispanic, and Asian), with a boy and girl version of each..

Back to the third item at the beginning of this paper

I am not an economist. I took only the introductory course in college, but I was fortunate to have held a bridging fellowship to the program in Public Policy at the University of Rochester when I was a faculty member there, and I find the way economists look at measurement and statistics problems to be fascinating. (Economics is actually not the study of supply and demand. It is the study of the optimization of utility, subject to budget constraints.)

What has all of that to do with Item #3? Plenty. If you are serious about measuring amount of pain, strength of an attitude, or any other such construct, try to do it in a financial context. The dollar is a great unit of measurement. And how would you assess the reliability and validity? Easy; use Pearson r for both. You might have to make a transformation if the scatter plot between test scores and retest scores, or between scores on the scale and scores on the gold standard, is non-linear, but that's a small price to pay for a higher level of measurement.

Afterthought

Oh, I forgot three other sources. If you're seriously interested in understanding levels of measurement you must start with the classic article by Stevens (1946). Next, you need to read Marcus-Roberts and Roberts (1987) regarding why traditional statistics are inappropriate for ordinal scales. Finally, turn to Agresti (2010). This fine book contains all you'll ever need to know about handling ordinal scales. Agresti says little or nothing about validity and reliability per se, but since most measures of those characteristics involve correlation coefficients of some sort, his suggestions for determining relationships between two ordinal variables should be followed.

References

Agresti, A. (2010). Analysis of ordinal categorical data (2nd. ed.). New York: Wiley.

Aitken, R. C. B. (1969). Measurement of feeling using visual analogue scales. Proceedings of the Royal Society of Medicine, 62, 989-993.

Beyer, J.E., Turner, S.B., Jones, L., Young, L., Onikul, R., & Bohaty, B. (2005). The alternate forms reliability of the Oucher pain scale. Pain Management Nursing, 6 (1), 10-17.

Bijur, P.E., Silver, W., & Gallagher, E.J. (2001). Reliability of the Visual Analog Scale for measurement of acute pain. Academic Emergency Medicine, 8 (12), 1153-1157.

Davey, H.M., Barratt, A.L., Butow, P.N., & Deeks, J.J. (2007). A one-item question with a Likert or Visual Analog Scale adequately measured current anxiety. Journal of Clinical Epidemiology, 60, 356-360.

Franchignoni, F., Salaffi, F., & Tesio, L. (2012). How should we use the visual analogue scale (VAS) in rehabilitation outcomes? I: How much of what? The seductive VAS numbers are not true measures. Journal of Rehabilitation Medicine, 44, 798-799.

Freyd, M. (1923). The graphic rating scale. Journal of Educational Psychology , 14 , 83-102.

Goodman, L.A., & Kruskal, W.H. (1979). Measures of association for cross classifications. New York: Springer-Verlag.

Guttman, L. (1946). The test-retest reliability of qualitative data. Psychometrika, 11 (2), 81-95.

Hawker, G.A., Mian, S., Kendzerska, T., & French, M. (2011). Measures of adult pain. Arthritis Care & Research, 63, S11, S240-S252.

Kraemer, H.C. (1975). On estimation and hypothesis testing problems for correlation coefficients. Psychometrika, 40 (4), 473-485.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22, 5-55.

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

Miller, G.A. (1956). The magical number seven, plus or minus two: Limits on our capacity for processing information. Psychological Review, 63, 81-97.

Owen, S.V., & Froman, R.D. (2005). Why carve up your continuous data? Research in Nursing & Health, 28, 496-503.

Pain Management Task Force (2010). Providing a Standardized DoD and VHA Vision and Approach to Pain Management to Optimize the Care for Warriors and their Families. Office of the Army Surgeon General.

Price, D.D., McGrath, P.A., Rafii, I.A., & Buckingham, B. (1983). The validation of Visual Analogue Scales as ratio scale measures for chronic and experimental Pain, 17, 45-56.

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

Vautier, S. (2011). Measuring change with multiple Visual Analogue Scales: Application to tense arousal. European Journal of Psychological Assessment, 27, 111-120.

Von Korff, M,, Deyo, R.A, Cherkin, D., & Barlow, S.F. (1993). Back pain in primary care: Outcomes at 1 year. Spine, 18, 855-862.

Wewers, M.E., & Lowe, N.K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. Research in Nursing & Health, 13, 227-236.

Woods, C.M. (2007; 2008). Confidence intervals for gamma-family measures of ordinal association. Psychological Methods, 12 (2), 185-204.