

# NOTE J: ORDINAL, NOMINAL AND LIKERT SCALE VARIABLES

## ORDINAL ISSUES

9/4/07 to 9/9/07 On sci.stat.edu

Norman Cliff, Jeffrey Long and others have been developing statistical methods based on ordinal methods (mostly Kendall's tau as far as I can see), and they argue that ordinal questions (and they suggest that nearly all questions in Psychology, for example, are ordinal) should be answered with ordinal methods

See for example

<[http://www.amazon.com/Ordinal-Methods-Behavioral-Data-Analysis/dp/0805813330/ref=sr\\_1\\_3/103-8728016-1667817?ie=UTF8&s=books&qid=1188899201&sr=1-3](http://www.amazon.com/Ordinal-Methods-Behavioral-Data-Analysis/dp/0805813330/ref=sr_1_3/103-8728016-1667817?ie=UTF8&s=books&qid=1188899201&sr=1-3)>

[http://www.leaonline.com/doi/abs/10.1207/s15327906mbr3103\\_4?journalCode=mbr](http://www.leaonline.com/doi/abs/10.1207/s15327906mbr3103_4?journalCode=mbr)

I wondered what people here thought of these methods? Are they worth the trouble?

Lance

Are they worth the trouble?" ...you haven't said what you are comparing them to. The alternatives seem to be to do something based on: un-ordered categories; ordered Categories; ordered lists of items; partially ordered lists of items; ordered categories with numerical limits; observations of numerical values; observations of numerical values within some of them only known to fall within certain limits; etc.

Basically you should try to make as much use as possible of all the information you have, without introducing information you don't have.

However, perhaps your question should really be "what type of data is it best to collect?" since it is the type of data (characteristics of the data) collected that should determine your choice of method. There have been some recent moves in certain areas to replace questions that effectively say "put these in your order of preference" with "how much would you be prepared to pay to have the use of each of these", thus trying to push things towards using a relative utility scale.

David Jones

Hmm. I was thinking more along the lines of Lord's "Numbers don't know where they come from". Does it matter if you treat ordinal data as interval data and analyze using ordinary MR, or is it really necessary (and the conclusions will be different or more reliable) if one goes to the enormous trouble of converting them to Cliff's dominance scores and then analyzing using his ordinal regression method? Further, there is little software for, and fewer options for, ordinal regression than for ordinary MR - so the trouble is multiplied.

Lance

The reason that people usually give for doing ordinal-level analyses is something like "the intervals on this rating scale have not been shown to be equal, so ...". Well, the intervals are almost certainly not equal, but that's not the issue. The issue is whether they are close enough to equal that treating them as equal -- i.e., doing interval-level analyses -- will not lead to wrong substantive conclusions. That's a judgement call that depends on the measure and the questions being asked. In my experience, ordinal- and interval-level analyses of measures whose intervals do not appear to be grossly unequal have almost always led to similar substantive conclusions. But the only way to know for sure, and to convince someone who is picking at your methodology because they dislike your conclusions, is to do both kinds of analyses.

Many people use "ordinal" when they should say "ordered categories", which confuses the ordinal-interval distinction with the continuous- discrete distinction. "Ordinal" does not imply "discrete". Often, the problems that people encounter with ordered categorical data are not due to the categories (or, more properly, the category boundaries) being unequally spaced, but are due to the fact that the categories are wide, creating large tie blocks of cases that are then treated as if they were identical.

Ray Koopman

Thanks for the thoughtful reply. I think Cliff et al. are targeting genuine ordinal data (first second, etc., and ranks, etc) because they claim that is the most common level of measurement in psychology. I am quite dubious. I do agree about your other points.

Lance

No, the most common type of variable in psychology has values that are ordered categories not ranks. The variables are often assumed to be not severely discrepant from interval level.

Confusion often results from failure to distinguish between A) variables with a few ordered values, e.g., agreement items used in a Likert scale and B) ranks where there are (almost) as many values as there are cases.

Another part of the discussion that is missed is that very commonly variables are part of a set designed as items in a summative scale and are not intended to operate alone as representing a construct.

When one is uncertain whether the variable might have values that are severely discrepant from interval level, there are two things among those that one might do. (It is rare for the values of a summative scale based on a few items to be severely discrepant from interval level.)

1) Run your analysis with the scale scores as is and as strict ranks. Or 2) use the methods developed by the group at Leiden which are available, for example, in the SPSS CATEGORIES module. These have ways of doing the analysis as nominal, ordinal, and interval so one can compare the results.

Art Kendall

Nice post, Ray. I'll just add that I think there is another issue about use of indirect measures of various things in Psychology. For example, in areas that use response time (RT) as the primary measure, RT differences between two or more conditions are often

used as an indirect measure of something else (e.g., inhibition). RT itself has ratio scale properties\*. Does this mean that the (indirect) measure of inhibition also has those properties? Probably not, because all kinds of other things besides inhibition also affect RT. Perhaps this is what Cliff et al. are getting at?

\* However, one could argue that RTs below a certain level represent anticipations, not reactions to a stimulus. In this case, RT would be interval, not ratio.

Bruce Weaver

Thanks, Bruce. My interpretation of what Cliff et al have said is that they are trying to make a general point -- that our measures almost always lack demonstrated interval-scale properties, that the questions we ask are often intrinsically ordinal, and that we should therefore routinely do ordinal-level analyses -- and that, although your take on RT is friendly to their position, they haven't said enough about such particular cases to warrant a conclusion about whether or not they would agree with you about RT.

But you have pointed up a problem with the traditional N-O-I-R hierarchy. RT (and other measures of duration, as well as frequency counts, GSR, GSP, other physiological measures, etc) has a real zero, but even if there is some construct which we can agree has a zero that corresponds to zero reaction time, RT still isn't necessarily a ratio measure of that construct unless the function relating RT to the construct is a straight line. In general, a measure is a ratio measure of a construct if the function relating it to the construct is a straight line that passes through (0,0), and an interval measure if the function is a straight line that does not pass through (0,0), but what do we call it if the function passes through (0,0) but is only monotonic, not linear? We need a label (ordinal-with-a-real-zero?), and general recognition that such things can exist.

Ray Koopman

## **LIKERT SCALE ISSUES**

The following are excerpts from messages on SEMNET (2005-2007) about dealing with nominal measurements using Likert scales, which are inherently ordinal. They illustrate the complexity of dealing with such measurements and computer models that require all data to be in the form of continuous (floating point or integer) variables. The responders are all experts who have had to deal with this problem, and represent only a sample of a wider concern about measurements in the social sciences.

---

The Likert scale was introduced by Rensis Likert in, "A Technique for the Measurement of Attitudes," Archives of Psychology, No.140, 1932, p.55

Michael Haenlein

---

Michael--

You may also wish to look at his thesis at Columbia University. Rensis Likert described his approach to measurement in an appendix. You might be able to get the thesis through Dissertation Abstracts International.

--Ed Rigdon

-----  
Professor Haenlein,

This is still not precisely what you are looking for, and I imagine you already have this reference, but you can find portions of the Appendix of the paper you seek here:

Likert, R. (1967). The method of constructing an attitude scale. In Martin Fishbein (Ed.), *Readings in attitude theory and measurement* (pp. 90-95). New York: Wiley.

Note that the page location in Volume 140 of Archives of Psychology according to the above reference is pp. 44-53.

Dave Wagner

-----  
I didn't mention anything about "science" in my email on this issue ...maybe someone else? However, what I would (and have said before) is that science may be quantitative or non-quantitative - and the use of numbers and invoked relations (whether additive, ordinal, or simply "categorical") is a matter for deep consideration in relation to the kinds of statements one wishes to make, or claim, about a phenomenon. For practical purposes within the social sciences, precision of "measurement" may be fine at just an ordered-category level. For others, the precise value of the temperature of a thermocouple, or the precise timing of the oscillations of a crystal within a computer chip may mean the difference between life and death in say the Space Shuttle.

Paul Barrett

-----  
I think you (Claudia Ramos) mean "Likert" scale variables (named after Rensis Likert)? Probably you are working with Likert-like variables (true Likert scales have gone thru very intensive psychometric assessment)... although they are arguably ordinal in nature, most of the time we treat these variables as interval or better. Although I do not use PLS, you can probably just treat them as regular interval or ratio-level data.

William D. Marelich

-----  
Dear all:

Another issue in the SEM world of marketing research (my world) is the discussion about the nature of the scales. Most of times, marketing researchers use Likert scales as a measure of the observable items. This scale is clearly ordinal, but it is a common practice to use it as a numeric scale. Even in the most relevant journals as Journal of Marketing or Journal of Marketing Research, when a latent concept is defined by observable items,

these items are measured in a Likert scale but they are treated as not discrete. Recently, I have read some papers of Karl Jöreskog where he said that Likert scale (from strongly agree to strongly disagree) is ordinal and researchers have to deal with this scale in SEM as a discrete scale. This means that the means, variances and covariances do not have meaning. He purposes to use the polychoric correlations and the WLS estimation method.

Jose Antonio Martinez Garcia

---

Hello all:

I have little information on Likert Scala. I want to share it.

Although type of Likert scale is ordinal, If (1) the scale is at least five levels (its original had five levels) , and ( 2) if the item (question) is related to respondent directly (that is you can not ask question about his company etc.) about his/her, then you can use this type of scale which was developed by Likert as interval scale.

Gulhayet Golbaby

---

Hi Jose,

I don't exactly share Gulhayat's optimism about treating certain types of ordinal data as interval without further empirical examination. From item response theory, it is easy to see that even large ordinal scales can be radically nonlinear. I would, instead, figure out the purpose of the use and the quality of the data before moving forward. For example, if your goal is to measure model fit to test a theory, or model comparison, some of the nuances of the model may be more trivial. Your major issue with ordinal data in this circumstance is likely to be the impact of multivariate nonnormality. This may be adjustable with chi-square correction, bootstrapping, or other techniques. However, if you are interested in conducting mean comparisons, your data cannot be transformed out of multivariate nonnormality, or other reasons that directly relate to understanding the scale of the item where ML extraction won't suffice, then it would be best to consider a polychoric technique for ordinal data (I'd recommend WLS-MV – mean and variance adjusted, given some recent literature comparing it to WLS). It may be beneficial to run both a ML model and a nonparametric extraction and compare the two sets of results. Recall that many criteria with which SEM uses frequently are only appropriate in ML, as they have not been tested with simulation in nonparametric instances.

Jason C. Cole

---

Hi, Jose,

As we may all have seen, this issue is often dealt with in journals in various ways. But I think we have to notice that when Likert scales are "used", in most cases it is noted that they are "treated" (or "regarded") as interval, which does not mean to change the nature of the scale but is permissible in certain contexts. I share with Jason's view in the last post that the most important thing is to know the purpose of the use of the data. I have noticed that quite a few articles cited Tabachnick et al. (1983) to support the use of Likert scales

as interval (Tabachnick et al. was updated in 2001). I quote their key arguments as follows:

"Continuous variables are those that are measured on some scale that changes values smoothly rather than in steps. Continuous variables can take on any value within the range of the scale. Precision is limited only by the measuring instrument, not by the nature of the scale itself. Some examples of continuous variables are times as measured on an old-fashioned analog clock face, annual income, age, temperature, distance, and scores on the Graduate Record Exam."

"Discrete variables are those that can take on a finite number of values (usually a fairly small number) and there is no smooth transition from one value or category to the next. Examples include time as measured by a digital clock, geographical area (e.g., categories based on continent), categories of religious affiliation, and type of community (rural or urban). Discrete variables may be used in multivariate analyses if there are a large number of categories and the categories represent some attribute that is changing in a quantitative way. For instance, a variable that represents numerous age categories by letting, say, 1 stand for 0-4 years, 2 stand for 5-9 years, 3 stand for 10-14 years, and so on up through the normal age span, would be useful because there are a lot of categories and the numbers designate a quantitative attribute (increasing age). But if the same numbers are used to designate categories of religious affiliation, they are probably not in appropriate form for analysis, because religions ordinarily do not fall along a quantitative continuum."

"It should be apparent that the distinction between continuous and discrete variables may be impossible to make in the limit. If you add enough digits to the digital clock, it becomes for all practical purposes a continuous measurement, while time as measured by the analogue device can also be read in discrete categories, say, hours. Continuous measurements may be rendered discrete simply by specifying cutoffs on the continuous scales."

(Tabachnick, Barbara G. and Fidell, Linda S., 1983, Using Multivariate Statistics, Harper & Row, Publishers, New York, pp.9-10.)

Charles C. Cui

---

First, never use articles published in journals, even in the most prestigious journals, as guides to good methods procedure. There are many reasons why articles are selected for publication. Many are published in spite of their methods practice, not because of it.

Second, there is an extensive literature here. Check references in

Rigdon and Ferguson (1991) and Babakus, Ferguson and Joreskog (1989 or 1987), both JMR. Some argue that numbers are numbers, and don't know where they came from. In many cases it makes little practical difference whether you treat the results as ordinal or interval.

Things have changed a little bit since then. The difficulties in correctly modeling ordinal data, including the needed sample size, have come back down to practical levels, so the

cost may now be justified in terms of the potentially small gains from applying correct procedure.

Ed Rigdon

---

Hi all,

The key argument in Ferrando (1995) is this. If you ask some items of a personality test in a Likert scale and after some time you ask the same questions in a continuous scale, correcting them measuring the millimeters since the right limit to the place the person draw the cross, you do not find differences between them.

The fact is that we can describe a continuous line as a Likert scale with infinite number of choices. And that's the reason why in an Exploratory Factorial Analysis with 9 choices we have bigger eigenvalues than with 5, and in a continuous than with 9...

Sorry for my english and sorry for talk about an article written in spanish.

Luis Manuel Lozano

---

What you appear to be saying is that there is an exact linear correspondence between the Likert scale integers and the continuous measures. E.g. on a 1000 unit "continuous scale", a Likert score of 1, 2, 3, 4 and 5 would correspond say to scores of 100, 300, 500, 700, 900? This is a remarkable finding as it demonstrates the capacity of humans to make perfect linear judgments in self-reports of personality type items such as "I like going to parties".

However, if you are saying that the rescaling of the 1000-unit magnitude scores into Likert integer categories yields a one-to-one correspondence, then OK. But, now each individual may or may not be exhibiting non-linear judgment processes which are conveniently "rescaled" into the Likert equal-interval range. For example, scores between 1 and 200 would be assigned a Likert score of 1, 201-400, 2 etc. Plenty of room here for non-linearity/non-additively.

The issue seems to be whether individuals show evidence of non-additively when using a "continuous" response format, and not whether once rescaled into a 1-5 metric, linearity is "revealed" by the coarse classification. A most interesting issue.

Paul Barrett

---

I tell my students that

1. Likert scales may give erroneous answers BUT;
2. They are very widely used.
3. If they really want to know the answer - interval analysis of Likert scales are good for pilot work but cannot give definitive answers.

4. If the interval analyses give very large EFFECT SIZES, as opposed to very low p-values the effect is certainly there.
5. If the objective is making a decision such as chose 'product 1' rather than 'product 2' then they are unlikely to make a mistake based on the interval analysis.
6. If they are interested in the MAGNITUDE of the effect, then interval analyses are flawed

Diana Kornbut

---

Ed and all,

It is not as if assumptions involving interval measurement not being met, ordinal measurement is, therefore, "correct." If intervals on an ordinal scale that result from differences on underlying latent variable (and who would ever imagine a latent variable to be ordinal rather than interval?) do not reflect those latent interval differences, then analyses of variables as ordinal can be misleading, wrong. We may not know how to recover and use all the information that is inherent in quasi-interval measures, e.g., Likert scales, but that does not mean we should simply ignore that information either.

To ignore it is akin to dichotomizing continuous variables.

I note, too, that the definition of "continuous" is sometimes fairly loose. Continuous is a conceptual matter, but our measurements are invariably categorical; it is just a matter of scale. Temperature is conceived as continuous; our measurement technology simply limits the precision of our measurement. Intelligence is generally conceived as continuous, but our measures of it permit only whole points of difference, with the scale being effectively limited to about 80 IQ points (60-140). Depression is also conceived as continuous, but the BDI permits only whole points of difference, and the number of scale points is effectively limited to about 40. And so on down. Social class, whatever that is, can be conceived as continuous, but most of our indicators have only a few points of distinction. Age is continuous, but it is common to find it expressed in terms of whole years, i.e., categorical.

So, if someone does a study of high school students and treats their ages in an analysis in terms of years ranging from 15-18, does the limited number of possible points of age make the variable categorical?

But age is often a measured variable used as a proxy for some underlying variable related to, but not caused by age. "Maturity" might be one example. If the ages 15-18 were taken as proxies for maturity, the relationship between age and maturity would be considerably less than perfect, and the intervals between adjacent years might well not be equal. The problems are not so different from those involved in Likert scales. So, should the data be treated simply as ordinal or might it be dealt with as interval?

My recommendation is that investigators should more often analyze their data under different assumptions of all kinds and then try to understand and explain any differences in results. We learn nothing about our assumptions without challenging (violating) them.

Obviously, more research is needed.

Lee Sechrest

---

Hi Ed, Jose et al.,

I too have heard many people comment that the use of specialized, theoretically correct methods for dealing with categorical data in SEM are not really worth the hassle. I won't take a firm stance on this issue here, but it useful to discuss it, as reminds us of a few core assumptions about SEM that are not always explicitly stated in appreciable terms.

All of the popular (and essentially equivalent) mathematical frameworks for classical SEM assume that the good ol' covariance matrix actually preserves all of the statistical information contained in the raw data set itself.

That's why we can judge a model's consistency with "the data" even though we have more conveniently fitted it to the second moments. Thus, the cluster of equivalent frameworks based on the notion of the covariance matrix as a sufficient statistic -- Bentler-Weeks, RAM, Jöreskog-Keesling-Wiley -- are the "pure" covariance structure models. We can refer to them as classical SEM. They are special cases of broader causal modeling frameworks that include alternative link functions for categorical variables (such as the framework implemented in Mplus).

Are the "pure" covariance structure models really suitable for most research situations? I would venture that in the majority of cases, the covariance matrix is not really a sufficient statistic for modeling, even though it is usually treated as such. As Jose pointed out, Jöreskog (1994) stated quite clearly that a covariance matrix of observed categorical variables was meaningless. Special types of correlations (e.g., polychorics) and estimators (e.g., WLS) have been suggested as ways around this problem in the case where alternative link functions are not available. But these methods have many restrictive assumptions as well. And we still of course have the problem of dealing with outcome variables that are truly categorical, especially in health research (e.g., death, disease, etc.). Classical SEM is not really appropriate in this last situation.

To what degree the use of conventional SEM with categorical variables (i.e., just maximum-likelihood estimation using an input covariance matrix) has compromised scientific conclusions over the years is difficult to determine. Fitting to alternative input matrices (e.g., polychorics with WLS) may not be offering anything better, since in many cases we are likely violating a new set of assumptions that still have yet to be fully explored in simulations... so better the devil (estimator) you know???

As to the usefulness of some of the newer alternatives to classical SEM, I think they have much to offer in the way of interpretability, which may gain them acceptance in a wider variety of research areas. For example, Mplus logit link allows interpretation in terms of odds ratios, which are less mystifying than the usual structural coefficients. I was once asked by some colleagues to interpret some path analysis (with measured variables) results where some of the DVs were binary. The coefficients had been estimated in LISREL using ML and a covariance matrix. I think I said something like, "Oh, those are the numbers which best solve the covariance structure equations." However, I recently

ran the same models in Mplus with logit links, the results of which were more illuminating and might actually be useable for policy purposes.

I don't mean to imply that classical SEM and its implementation in the packages (old friends) we have all grown up with are becoming obsolete. But I think the era has come where we need to more and more carefully consider the nature of our data, and the research questions we wish to answer in our modeling endeavors. Hence, the question, "Are categorical variable methods really worth the trouble?", is one that deserves careful consideration in every application.

Cameron McIntosh

---

To anybody interested,

I've been thinking about the ordinal/interval issue a little bit for the past few days. I am sympathetic to all the positions. Paul Barrett's comment about "science" particularly got me thinking, however. Does science really require something more than a "practical" application? And what is that more? I might learn if I took the time to read Paul's web. But not having done so, I will venture the opinion, no. Note, however, that my position equates "science" with basic research and "practical" with applied research; if that's not what Paul means, my comments are a bit beside the point.

As a practical example, a Canadian girl, Lisa Neve, is in prison for her entire life partly on the basis of her score on a twenty item psychopathy scale. When she was sentenced (I think about ten years ago), I wondered how the scale would stand up under the current SEM methods available then that might include the usual tests of validity, however defined. Some of the recent comments on SEM come from people, I suspect, with similar clinical concerns. In contrast, in "scientific" research, the magnitude of a particular slope is not often of much scientific interest. All that really matters is whether it's significantly different from zero and has the right sign. If the value were important, we would focus much more on unstandardized slopes, and quantities, such as Cohen's D, would be consigned to the dustbin of methodological history. (If you think this position is uniquely mine, you should read the early sociological methodological literature by Blalock and Duncan and ask yourself why economists do not use standardized slopes.) In other words, I don't think that the many violations that SEM researchers engage in produce misleading results. (Although to take the opposite position, I did argue some years back that two-dimensional structure of the abortion scale [due, I believe, to Alice Rossi] was an artifact of applying classical factor analysis to dichotomous items.

The interval/ordinal debate also reminds of another example. About fifteen years ago, my former partner, JM, was consulting for a physiotherapist who was studying recovery of shoulder injuries suffered by professional and amateur football players. The physiotherapist had measures on each player, repeated within each occasion, on three occasions. He used two instruments to measure each player's shoulder strength: a "strain gauge," a simple mechanical device that the trainer could administer on the football field, and a cybex, the machine we see in workout gyms. From an engineering point of view, the cybex was vastly superior to the strain gauge and, therefore, was regarded as the "gold standard" for measuring shoulder strength. Yet, when JM analyzed the data, she

found that the strain gauge scores had the same stability as the cybex scores and predicted recovery equally well. Most of you can probably figure out an explanation for this result. The bulk of the measurement error, both across football players and between occasions within the same football player were resided in the player. The mechanics of the instrument (including the ability of the trainer to administer the measure and read off the results) probably contributed only a trivial amount to the variance of the error term. In other words, the (low) unreliability that was observed resided in the player not in the instrument.

What's the relevance of this story for the debate? As I see it, the strain gauge and cybex correspond to "classical" and more appropriate methods for the analysis of ordinal Likert items. The analogue to the human factor is the invariably tenuous connection between most sets of Likert items and the constructs that we have in mind. Given this weak connection, the question of whether the relation is really linear, though interesting, is beside the point.

Finally, to end on a really radical position (that I don't hold; otherwise, why am I here?) What's the point of doing SEM? If we have good data, we don't need it, and, if we have bad data, it won't save it. So do we really need it? Maybe we ought to go back to contingency tables, regression, t-test, ANOVA, etc.

Mike Gillespie

---

Hello Mike

It all depends upon the kinds of statements you wish to make about the causes of a phenomenon. We can do a lot with "it happens, it doesn't" through to "more of this results in more of that" through to "a 5.57 unit increase in X results in a 3.022 increase in Y". The problem for the social sciences is that it invariably ends up using methods, which provide statements of the latter form, whilst working with constructs and variables, which remain at the level of precision of the middle kind of statement.

As Michell states, whether a variable exhibits a quantitative structure, or not, is a hypothesis in its own right which is required to be tested. Simply using a methodology to construct a latent linear variable is not the same as empirically determining that such a variable is indeed "quantitatively structured" in the "world". Until that hypothesis is tested, we remain in a fog of uncertainty about the measurement precision (and structure) of our variables; and although we use techniques which rely upon equal-interval/ratio-level measurement and which would permit us to use precise terminology of magnitude effects, we are invariably uncomfortable with this and usually moderate our results using ordinal "qualifier" terms of effects. E.g. People scoring higher on X tend to do Y etc.

Does it matter - well; we are back at "it all depends upon the kinds of statement you want to make about cause and effect". Richard Lynn for example has this very problem - of asserting that females have lower intelligence than males, on the basis of using scaled ability test scores, which yield maybe a 3-5 point IQ difference across large groups of each gender. Yet, there is no empirical evidence that IQ is a quantitative variable. So, what exactly does a 3-5 point difference in ordinal scores mean? [Note the use of "precise" estimates, 0.33 standard deviation units of a 15-unit standard deviation for a

scaled test score which may not be additive at all). That there are discrepancies between the scores of males and females is not in question, it is the precision adopted by the investigator that speaks of even a 0.2 IQ point magnitude, or a 0.21sd "advantage" which remains invalid until IQ as a variable is shown to be quantitatively structured (i.e possessing additive magnitude relations using a standard unit of measurement).

See for example:

Lynn, R. and Irwing, P. (2004) Title Sex differences on the progressive matrices: A meta-analysis. *Intelligence*. Vol 32(5) 2004, 481-498.

-Abstract-

"A meta-analysis is presented of 57 studies of sex differences in general population samples on the Standard and Advanced Progressive Matrices (SPM and APM, respectively). Results showed that there is no difference among children aged 6-14 years, but that males obtain higher means from the age of 15 through to old age. Among adults, the male advantage is 0.33d equivalent to 5 IQ points. These results disconfirm the frequent assertion that there are no sex differences on the progressive matrices and support a developmental theory that a male advantage appears from the age of 15 years. A meta-analysis of 15 studies of child samples on the Colored Progressive Matrices showed that among children aged 5-11 years boys have an advantage of 0.21d equivalent to 3.2 IQ points."

Would we have thought any differently about the result if the difference was reported as a difference in the group medians or interquartile ranges of the rank scores on the test (Ravens) - and not using a scaled "IQ" score at all? Perhaps - perhaps not.

Paul Barrett

-----  
In my survey data, there are two types of measuring scale; 1-5 (1=strongly agree, 2= agree, 3= neutral, 4= disagree and 5= strongly disagree) and 1-3 (1= very high, 2= modest and 3= very low) in Likert scale. For analysis, I need to convert 1-3 scale equivalent to 1-5 scale. Can anyone kindly suggest me the possible ways to do that. Also I have used a reverse scale (1-5); how a reverse scale can be made equivalent to the original 1-5 scale?

Azharul karim

-----  
You can't really do either one. You can't turn a 3 response category scale into a 5 response category scale because you don't have the information. But I agree with Cam in thinking that you probably do not need to do this.

Mathematically, it is easy to flip a reversed scale:

new score = 6 - old score

But that will probably not make the reversed scale equivalent. Reverse-worded items tend to have their own factor structure, as respondents reply in a fundamentally different way to negatively worded items than to positively worded items. It is as if there is a

"negativity / positivity" factor that shapes responses, along with the substantive factor that you are trying to capture.

--Ed Rigdon

-----  
You can standardize both types of item.

Then if you want to construct scales using mixed item types it is a straightforward process of addition.

Martin G. Evans

+++++

The issue here is, 1) If the question is worded in a positive sense, then we have a positive Likert scale. 2) If the question is worded in a negative sense, then we have a negative Likert scale. 3) These imply two separate variables (factors). 4) The issue is, can these two be combined into a single Likert scale representing a single variable (factor)?

-----  
+++++

Brent,

You had a separate factor for items that were reverse scored? And the model with two factors - direct scored items and reverse scored items - had a non-significant chi-square. What was the correlation between the factors? It seems you have two factors. What happens when you assign all of the items to one factor?

Stan Mulaik

-----  
Professor Mulaik and Professor Rigdon,

Thank-you for taking time to reply. If I split the indicators so that the ones that were reverse scored to load on a separate factor, the chi-square is not significant. So, if I understand your replies correctly, it could be that the scale is likely to be unidimensional, but the reverse scoring made it seem like it was not unidimensional. Does that sound right?

Brent Coker.

-----  
Brent—

Well, if there is another factor involved, then the scale is NOT unidimensional. But that's the nature of reverse-worded items. They suggest that there are very likely two factors at work--a substantive factor and a positive / negative factor. Even if you model all items loading on the substantive factor, and the reverse-worded items additionally loading on a second, methods factor, that is still two factors (even if it is also a remarkable demonstration).

In my opinion, you are better off just using the reverse-worded items to make sure that your respondents were paying attention. Then discard the items before beginning your factor analysis--like the bay leaf or bouquet garnish in a sauce, they are important, but you don't want them turning up at table.

--Ed Rigdon

-----

I was wondering if anyone knows what the proper reliability test is for ordinal measures? Do I still use Cronbach's Alpha?

---

Thanks in advance for all of your help!  
Joelle

-----

Hi Joelle,

How many ordinal item you have? I don't see why you can't use Cronbach's alpha. In fact, in a typical Likert-type scale, a participant gives answer on one of several categories. In my own opinion, these categories consist of an ordinal response, such as 1=Strongly Disagree, 2=Disagree, 3=Neither, 4=Agree, 5=Strongly Agree. To make the scale score more continuous, researchers take the sum/mean of several items.

HTH,  
Yung-jui

-----

Yung-jui,

I disagree with you. Data consisting of a ranking or ordering of measurements cannot assess the distance amongst the order values. For instance the distance between 1 and 2 maybe shorter than between 3 and 4. This is the nature of ordinal measures. Having said all that, I believe that for ordinal values, reliability can be assessed in terms of the weighted kappa (see Cohen 1968).

Andrea

Reference

Cohen J (1968) Weighted Kappa: nominal scale agreement with provision for scaled disagreement on partial credit. Psychol Bull 70:213-220

-----

Hi Andrea,

But how can you know whether the distance between 'Strongly Disagree' and 'Disagree' is the same as that between 'Strongly Agree' and 'Agree'? Yes, we assume they are the

same, but are they actually? Besides, we only have 1,2,3,4,5, i.e. five intervals, for a single item. My point is that if given only one item (Likert-type 5 point), the response can be seen as ordinal as well.

Yung-jui  
-----

Hi Yung-jui

But how can you know whether the distance between 'Strongly Disagree' and 'Disagree' is the same as that between 'Strongly Agree' and 'Agree'? Yes, we assume they are the same, but are they actually?

I don't understand your point -- obviously it is an "assumption" that renders ordinal and cardinal values different.

Besides, we only have 1,2,3,4,5, i.e. five intervals, for a single item. My point is that if given only one item (Likert-type 5 point), the response can be seen as ordinal as well.

Why? Suppose you need to assess the quality of five products. If you use an ordinal scale the result will be that you will have an order of preference having products ranked from nr 1 to nr 5. You will never be able to know however whether or not two products have the same quality level, whereas if you use a 5 point Likert type scale you will. That doesn't imply that Likert type scales are "sinless" ... but that's another story.

Andrea  
-----

Dear all,

my personal opinion is that all scales are ordinal, infinitesimal discrete. When you ask in a questionnaire for an opinion or when you want to measure an attitude, it is better to assign numbers as Daniel Yang points out. I think for most people (this is not very scientific, is only an opinion) the distance between 4 and 5 is the same than between 1 and 2. But for most people may be the distance between "much" and "enough" is not the same than "little" and "nothing".

Jose  
-----

Hi Andrea

Many thanks - I wasn't aware of this approach. It looks like it may have some application to the issues I am dealing with. In particular concerns in the literature that Asian respondents avoid using extreme values on Likert-type scales. I will read in more depth at a later date when I revisit this project although one concern over the use of the approach in an SEM context is the requirement to pre-standardize the data.

For an excellent discussion on issues related to cross-cultural research see:

Schaffer, Bryan and Christine M Riordan (2003), "A Review of Cross-Cultural Methodologies for Organizational Research: A Best-Practices Approach," *Organizational Research Methods*, 6 (2), 169-215.

See specifically p192 for discussion on how certain cultural groups differ in their response sets.

Murray-Garcia JL, Selby JV, Schmittiel J, Grumbach K, Quesenberry CP Jr. Racial and ethnic differences in a patient survey: patients' values, ratings, and reports regarding physician primary care performance in a large health maintenance organization. *Med Care*. 2000;38:300-10.

Thanks Again,

Greg

++++  
Since there are many marketers and survey users on this list.

Please tell me your views, preferably with references on whether there is any systematic difference between these 2 forms of a Likert item.

Appologies if this is of not interest to you

FORM 1

How confident are you about your responses to the test questions?

Very confident

Confident

Moderately confident

Unconfident

Very unconfident

FORM 2

Show the extent of your agreement with the following statement.

I am confident that my responses to the test questions are correct

Strongly agree

Agree

Neither agree nor disagree

Disagree

Strongly disagree

I have data on this, but am interested to know whether list members can

predict the results

- A. no difference
- B. more positive responses in form 1
- C. more positive responses in form 2

Best

Diana

Professor Diana Kornbrot

-----

Diana--

(Sorry, no references.)

The language of FORM 1 seems more direct, so I would expect more \*accurate\* answers from FORM 1.

However, FORM 2 seems more suited to acquiescence bias, so I expect responses to FORM 2 to indicate a higher level of respondent confidence. Therefore, I choose option C, assuming "highly confident" is at the high end of the scale, and assuming sufficient statistical power.

Obviously, if power is low, I favor option A.

--Ed Rigdon

Edward E. Rigdon, Professor and Chair,

-----

Diana,

I would second what Ed Rigdon argued, in principle.

However, I would not be so sure whether it will be pure acquiescence bias.

I could imagine that with option B less categories are actually used (in particular I would be surprised if the middle category would not be underutilised; after analysing such scales by IRT these neither x nor y categories almost never worked properly). So, I would expect a distribution

that is closer to normal with form A than with form B but I am not sure whether

the means differ much (significance depends on the sample sizes, of course).

When will you let us know what the empirical findings reveal?

Thomas

-----

Obviously, the more confident people are the less normal the distribution. That is why comparisons of group means of Likert ITEMS is ALWAYS in breach of ANOVA assumptions, unless both groups are overall neutral. If one group makes more positive responses it is likely to be negatively skewed, if another group makes more negative responses it is likely to be positively skewed. So the assumptions of Mann-Whitney and other rank based tests are also violated. The appropriate analysis would be based on cumulative probabilities either logit or normal transformed. Then you have the 'minor' problem of explaining to public.

Best

Diana

-----  
Hi, Diana! This is really interesting! Kudos to Ed, Thomas, and other for courageously offering predictions.

Diana and All, here's a thought about this and a couple of questions. Suppose we recast your ANOVA as a simple multiple regression problem:

$$Y = b_0 + b_1X + e,$$

where Y is the Likert rating, X is a dummy coded group identifier coding membership in one of two groups, b<sub>0</sub> is an intercept, b<sub>1</sub> is a slope coefficient, and e is the difference between Y and predicted Y. I've dropped subscripts because email is not amenable to them. Let's say you're going to use OLS to fit this model with the typical linear regression assumptions, e's are IID ~ N(0, var(e)), and so on.

Suppose also that in one of the groups, the Y's look positively skewed, and in the other, they appear negatively skewed. The question is, what's more important here to your model assumptions, the distributions of the Y's, or the distribution of the e's? IMHO it's the latter.

Having asked this, however, it's clear to me that one should be prepared to treat data from scales like this as ordinal.

A second question: Diane and All, was Likert's original scale specifically an agree-disagree scale? I seem to recall that it was, although I wasn't around at the time. :-)

Cheers,

Lynd

-----  
>A second question: Diane and All, was Likert's original scale  
>specifically an agree-disagree scale? I seem to recall that it was,  
>although I wasn't around at the time. :-)

It was a five point scale, anchored as

- 5: Strongly approve
- 4: Approve
- 3: Undecided
- 2: Disapprove
- 1: Strongly disapprove

At least that's what's written in his paper:

"Likert, Rensis (1932), "A technique for the measurement of attitudes," Archives of psychology, 140, 55"

Michael

-----  
As I recall (apparently, fallibly), one reason for the adoption of Likert's method for measuring attitudes was that it was easier to use than Thurstone's paired comparisons procedure. But Likert's and Thurstone's views on what an attitude is, also differed.

Cheers,

Lynd

-----  
Thanks for this.

But I am extremely unhappy about casting ordinal data into methods such as ANOVA or regression that assume the data are metric (interval or ratio).

This approach is fundamentally flawed because the distance between adjacent points on the scale are not equal.

I am not interested in average scores, they are meaningless. I am only interested in the proportion of people which choose each alternative. So any model I use would focus on those proportions.

Best

Diana

-----  
This is very interesting, good for Likert

To me 'approval' is appropriate, like confidence.

Agreement is very different. I might disagree with 'I approve of x' because I strongly approve, rather than merely approve

Best

Diana

-----  
Hi, Diana. I get your point and I agree with you about the metrics. In certainly doesn't make sense to average ordinal (or nominal) data.

The point I was trying to make (and a small one it is, at that) is that in models like ANOVA and regression, it's not the distribution of Y's that matters so much as the distribution of residuals ( $Y - \hat{Y}$ ) and what you'd like to be able to assume about them.

Also, as I think about it, how rating scale categories are (or are not) labeled is probably one of the most important considerations. Even without labels, or with strictly numeric labels and anchored endpoints, there's no guarantee that respondents produce interval or better data.

Cheers,

Lynd

-----  
{1} Paul Barrett

tel: +64 (0)9-373-7599 x82143 Mob: 021-415625

Adjunct Professor of Psychometrics, University of Auckland, NZ

Adjunct Assoc. Prof. of Psychology, University of Canterbury, NZ  
email: p.barrett@xtra.co.nz  
paul.barrett@auckland.ac.nz  
paul.barrett@canterbury.ac.nz  
web: www.pbarrett.net

(2) William D. Marelich, Ph.D.  
Associate Professor  
Dept. of Psychology  
CSU Fullerton  
wmarelich@gmail.com OR  
wmarelich@fullerton.edu

(3) Jose Antonio Martínez García  
Universidad Politécnica de Cartagena  
Grupo de Investigación en Marketing  
martinezjose1987@yahoo.es

(4) Jason C. Cole, PhD  
Senior Research Scientist & President  
Consulting Measurement Group, Inc.  
Tel: 866 STATS 99 (ex. 5)  
Fax: 818 905 7768  
7071 Warner Ave. #F-400  
Huntington Beach, CA 92647  
E-mail: [jcole@webcmg.com](mailto:jcole@webcmg.com)  
web: <http://www.webcmg.com>  
*The Measurement of Success*

(5) Dr. Charles C. Cui  
Senior Lecturer in International Management & Marketing  
Manchester Business School  
The University of Manchester  
Booth Street West  
Manchester M15 6PB  
United Kingdom  
Tel: 0161-306 3461  
Fax: 0161-306 3167  
Charles.C.Cui@mbs.ac.uk  
[www.mbs.ac.uk](http://www.mbs.ac.uk)

(6) Edward E. Rigdon, Professor and Chair,  
Department of Marketing  
Georgia State University  
P.O. Box 3991  
Atlanta, GA 30302-3991

(express: 35 Broad St., Suite 1300, zip 30303)  
phone (404) 651-4180 fax (404) 651-4198

(7) Luis Manuel Lozano  
Facultad de Psicología  
Universidad de Oviedo  
Plaza Feijoo s/n  
33003 Oviedo

(8) Cameron N. McIntosh, MA  
Analyst / Analyste  
Health Analysis and Measurement Group / Groupe d'analyse et de mesure de la santé  
Statistics Canada / Statistique Canada  
24-Q R.H. Coats Building  
Ottawa, ON  
K1A 0T6  
Phone: (613) 951-3725  
Fax: (613) 951-3959

(9) Professor Diana Kornbrot  
Evaluation Co-ordinator, Blended Learning Unit  
University of Hertfordshire  
College Lane, Hatfield, Hertfordshire AL10 9AB, UK  
Blended Learning Unit  
voice +44 (0) 170 728 1315  
fax +44 (0) 170 728 1320  
Psychology  
voice +44 (0) 170 728 4626  
fax +44 (0) 170 728 5073  
email: d.e.kornbrot@herts.ac.uk

(10) [sechrest@EMAIL.ARIZONA.EDU](mailto:sechrest@EMAIL.ARIZONA.EDU)

(11) [Michael.Gillespie@LIU.EDU](mailto:Michael.Gillespie@LIU.EDU)

(13) Azharul karim  
Postgraduate student  
University of Melbourne  
Australia

(14) Martin G. Evans  
Professor Emeritus, Rotman School of Management, University of Toronto.

(15) Michael Haenlein  
Professor of Marketing  
ESCP-EAP European School of Management  
79, Avenue de la République | 75011 Paris | France

(16) Dave Wagner  
Ph.D. Student (ABD)  
Touro University International  
[dwagner@tourou.edu](mailto:dwagner@tourou.edu)

(17) Brent Coker  
Brent.Coker@VUW.AC.NZ

(18 )Joelle Ferron, MSW, LICSW  
Doctoral Student  
School of Social Work, UNC Chapel Hill

( 19 )Daniel Yang <yungjui@GATE.SINICA.EDU.TW> escribió:

( 20 ) Andrea Vocino  
Bowater School of Management and Marketing  
Deakin University  
Burwood VIC 3125 Australia

(21) Jose Antonio Martínez García  
Universidad Politécnica de Cartagena  
Grupo de Investigación en Marketing  
[www.upct.es/~gim](http://www.upct.es/~gim)  
Página personal: [www.upct.es/~gim/inv\\_jose](http://www.upct.es/~gim/inv_jose)  
Email: martinezjose1987@yahoo.es

(22) Thomas.Salzberger@WU-WIEN.AC.AT

(23) Lynd Bacon <ldb@LBA.COM>