Safe and Supportive Schools
Engagement | Safety | Environment

Evaluating the Reliability of Surveys and Assessments

WEBINAR QUESTION AND ANSWER SUMMARY

On October 25 and 26, 2011, the Safe and Supportive Schools Technical Assistance Center (Center) hosted a Webinar titled "Evaluating the Reliability of Surveys and Assessments." During the session, the presenter, Dr. Lorin Mueller, Principal Research Scientist at the American Institutes for Research, received several questions from the audience. Since the presenter could not answer all of the questions during the event, the Center has prepared the following Webinar Question and Answer Summary with responses to each question. For additional information, please email or call the Center (sssta@air.org; 1-800-258-8413).

Please note the content of this summary was prepared under a contract from the U.S. Department of Education, Office of Safe and Healthy Students to the American Institutes for Research (AIR). This Q/A summary does not necessarily represent the policy or views of the U.S. Department of Education, nor does it imply endorsement by the U.S. Department of Education.

Q1. What can we do to ensure the validity of our survey items?

Validity is a complex issue. The joint Standards (AERA, APA, & NCME, 1999) talk about three forms of validity evidence: content, criterion-related, and construct. Content validity means that the content of the survey supports the inferences you intend to make. Content validity goes beyond face validity (superficially appearing to measure the construct of interest); you can have content valid items that only an expert understands represent school climate. But in general, you should see a clear link between the content of the questionnaire, the construct you want to measure, and the intended use of the instrument. In general, this step is the first in establishing validity.

Criterion-related validity evidence is generated when your survey items are correlated in expected ways to expected outcomes. For example, if schools with higher levels of climate for safety have fewer bullying incidents, and schools with higher climate for academic achievement have higher test scores and more students entering college, these correlations support the validity of the measure. This step is generally next in the validation process.

Construct validity is an extension of content and criterion related validity to form the argument that the measure shows a theoretically meaningful pattern of results (correlations, for example) across predictors, criteria, and other measures of a similar construct. In most cases, construct validity is gathered through years of study and research.

Note that the key issue is that validity is a measure of how appropriate the inferences you make based on instrument scores are. So validity isn't a property of the survey items themselves – but on the way you use the instrument to inform your decisions. It's a good idea to think about whether the items support what you're using them for.

From a practical perspective, the easiest way to ensure validity is to use an established, research-based measure and use it in ways recommended by the scale creators. If you wish to create your own scale, start by thoroughly describing what you intend to measure, all of the conditions that make a difference, and then construct a measure by writing clear questions relating to those conditions.

Q2. When you say that it's important that scores are stable from day to day, what do you mean by stable? How much change can be tolerated (i.e., how much would need to occur for the scores to be unstable)?

It depends on the particular method being used to assess stability. When looking at correlations over time for a given observation, correlations over time generally can range from .6 to about .9 depending on what you want to do with the data.

If you're looking at two occasions and they're correlated at about .6, you can raise your reliability if you use the average of those two occasions. So it will be above .6, correlation will be above .6, should be about .75 or so according to the Spearman-Brown prophecy formula (which you can find on the internet or in the books I cite) so that would be considered to be reasonable reliability if you're doing things like research. So referring back to whether boys experience a different climate than girls, if as part of a survey you ask a student "are you a boy or are you a girl," they might answer a little bit differently on different occasions, either because they make a mistake bubbling in a box or they're goofing around. Sometimes you get unreliable data there, but if you're seeing correlations at about the .6 level, it's probably okay to move on from there.

With self-reports of gender you'd hope for higher reliability but for assessments of climate at the classroom level or at the school level for an individual student, those perceptions can change day-to-day because events influence our perceptions, and so we want to make sure that we get a lot of observations to make it more stable. So generally I think we're talking about .6 if you're going to use multiple data points and going up from there if you're only going to use one.

Again, this will depend a lot on what you intend to do with the data. For example, if you've got two observations for each respondent and you averaged those, that will increase your reliability substantially and you'll have a more stable estimate of what you expect to see.

Q3. When you say it is important to link the time element to what you're trying to investigate, can you give an example of how that might be operationalized?

Let's say you're interested in how a bullying incident affects performance on a standardized test. Well, you wouldn't want to find a kid who had been bullied once a couple of months ago and count that the same as someone who had been bullied earlier that day. So my recommendation is just to do some thinking about whether you're measuring something that you expect to last a long time or change relatively quickly, and make sure that your measurement strategy matches the time element – as an example, that your questions reflect the appropriate time element.

Q4. Can you explain a bit more about the "Reliability of EFA is no better than rational scoring (e.g., a priori scales)" What is rational scoring… is that the same as confirmatory factor analysis?

Rational scoring is simply defining your scales in advance and using classical statistics to evaluate the quality. You could use rational scoring to create a confirmatory factor analysis but that is a more complex procedure. The whole point I was trying to make is that a lot of folks try to use EFA to show that they are being rigorous in their scale development, but this technique is flawed for climate scales (because of clustered/grouped samples) and because the results of an EFA aren't necessarily more reproducible than results based on simply creating theoretically meaningful scales. For climate, multi-level confirmatory factor analysis with a diverse sample would be appropriate, but very difficult in practice to conduct.

Q5. Our school climate scales are a mixture of self-reported perceptions and behaviors, mostly Likert scales and "counts" (categorical response options, 0-1, 2-3, etc). Is it ok to use alphas to assess the reliability of our scales when both Likert and counts are included?

This is not recommended. You don't know whether or not those behavior counts are equally spaced intervals: getting from zero to one may be pretty easy, getting from 15 to 16 may be pretty difficult; you don't know what the distribution of those counts will be.

Pattern analyses are probably more appropriate for mixed-type scales but don't get focused on the numbers. Instead you need to be able to demonstrate that things relate to one another in a sensible way. That also affects how you interpret the scales — you would not want to add a behavior count in with Likert type scales unless you can show that they're comparable in the way that they're distributed across the items.

It's a pretty complicated question. Again I don't recommend alpha for those mixed type scales; I do recommend pattern analyses.

If you absolutely must convert mixed scales to a single score, use cross-tabs to decide how each of the frequency measures should be scored against the Likert-scale. So for example if a count of zero is consistent with respondents who generally select "strongly disagree" to Likert-type items, then you can assign that the same score as "strongly disagree." Likewise if a count of three is consistent with the midpoint of the Likert scale, then three can be assigned the scale midpoint. Counts of one and two can be assigned the same score as "disagree."

Q6. I have heard of surveys that ask the student respondents at the end whether they answered the questions honestly. If a student is not answering them honestly, I don't think they are going to admit that, are they? I've wondered about what you would do with that data. What are your thoughts?

If somebody says no, you might want to throw that case out because they either didn't read the question or they're admitting to you that you asked them sensitive questions that they didn't feel comfortable responding to. If respondents say yes, there are no guarantees that they aren't lying. At least you're removing the people that are telling you that they're lying.

Q7. If our data change in a subsequent year, is there a way to know for sure that the change is due to the implementation of Evidence Based Programs in our school, or maybe due to something else?

That's a very difficult question. It gets to the idea of establishing the stability of what you're measuring over time. It's helpful if you can establish that stability by doing measurements throughout the year, pre- and post- intervention. Also if you look at a model like generalizability theory, where you can really get an idea of how different sources of variation might affect the reliability of your instrument, that's going to give you a better idea of whether or not things will change much from year-to-year, or from person-to-person, in order to identify what really is the range that you'd expect to see an average measure fall within.

Once you establish the standard error of measurement (again you can find this term and formula in the books I mentioned or on the internet), you can determine how sure you are that the change you observe reflects something real rather than measurement error. This process is much like a Z- or t-test.

Q8. How do you reconcile stability with interventions meant to affect school climate? Wouldn't you expect non-stable responses if you're trying to improve school climate?

Yes. In fact, when we talk about stability estimates, we want to make sure that we're measuring the construct over an interval where we don't expect things to change. So if you're planning an intervention and you're going to do a time-one and time-two assessment, relatively quickly after you do the time one assessment, select a sub-sample of your original sample and re-administer the survey. It generally doesn't take a lot of effort to do this and you can explain it by saying "We're trying to get an estimate of the reliability here. We've selected you to participate in this special study." You might have to provide them with some incentive. The idea is to get 100-200 responses out of 1000 and look at how they correlate to their earlier ratings. In those cases, you can demonstrate stability over time because it happens before your intervention is complete and it should demonstrate some level of reliability.

Similarly if you gather generalizability data as described in the previous question, you can determine whether changes are an issue of stability or program changes.

Q9. When I'm reviewing articles, what am I looking for in terms of the reliability coefficient scores? How do I determine what is robust, what is okay, what is suspicious or what is weak?

This is a complicated question because to me it depends on the design of the research in question. In general, you want to see reliabilities above .7 for research, .8 for diagnostic/feedback purposes, and .9 for high-stakes decisions. But those values are just guidelines and much depends on what is realistic for the practical measurement constraints. Also, consider that coefficient alpha is a lower bound estimate of reliability – the actual reliability in cases where internal consistency is an appropriate measure of reliability may be much higher. And as I mentioned alpha is not necessarily appropriate in all cases.

There are a few things to consider when evaluating the estimated reliability of a measure (and I say estimated because many design and sampling issues affect the number). I'll put this in terms of alpha since it is the most commonly reported, but the thinking can be applied to any classical reliability estimate (such as an ICC). First, how diverse is the construct being measured? If it's very diverse, alpha will be low unless there are dozens if not hundreds of items (such as in knowledge tests). Second, how diverse is the sample? If the sample is not very diverse, alpha will be underestimated. Third, is this a single-shot measure or will the measure be averaged across many raters/occasions? If you look at school climate measures, you generally see very low reliability for an individual climate response, but since they are averaged across about 20 students, the teacher-level average can be very reliable.

You should also be aware that reliability is different than classification consistency. Reliability relates to how differences in scores are reproducible; this is known as "relative" reliability because it just looks at the reliability of your score versus the average. Classification consistency is the reliability of the scale when you apply a fixed cut score. This estimate may be lower than reliability, because depending on where the cut score is set, lots of people might find on the wrong side of the cut score.

Lastly, if you know anything about Item Response Theory (IRT), you can get a much more precise measure of reliability for various points on the scale, and this is generally expressed as a conditional standard error of measurement (meaning the SEM at a particular point on the scale).

Q10. Would confirmatory or exploratory analysis be better for behavior change?

That depends on how your scales are set up. There are models called latent growth models which might look at scales over time and which correlate, for example, behavior change over time. Let's say you've got time one, time two, and time three; to determine the growth trajectory you would use a particular type of confirmatory factor analysis model. So I would probably point you in that direction.

Q11. What is the advantage of using ICC over alphas? And what's an example of when you would use ICC instead of alpha?

I recently finished a report for the Institute of Education Sciences looking at pre-service teacher knowledge linked to perceptions of program focus at about 100 universities. When we looked at students within schools and how much they agreed about the content of their school's curriculum, what we found was that the alphas were very low for our subscales because they were asking diverse questions and the questions were pretty broad so we couldn't, for an individual student, identify the big differences between what the curriculum was.

Part of that is because there were 100 different schools that we had surveyed, and we had about 30 people per school. Within a school, we were able to demonstrate that they agreed on what was involved in their curriculum. The alphas didn't look good because the total was suppressed by the fact that within a given school there was a lot of agreement. When we looked at the ICCs, the ICCs were much better. We were able to say for a given school, "We have about

.85 reliability on these different instruments even though the alphas of the instruments were only about .50 for an individual student." This was a case where, with so many grouped observations and where the group sizes were so large, the variability of that total variance in the alpha formula was suppressed and we had to go to something like an ICC to really show that our measures were reliable in the way they were used.

Now there are cases where you've got grouped observations that might not be a problem. For example, if you've got a parent survey and on occasion both parents might respond to this survey, that's a case where you might be able to ignore it. But then again, if you think as a matter of course that both parents respond to the survey, then you might want to make sure that you include that in your reliability estimate.

Again, this is going to push you more towards a generalizability theory model, and you'll be looking at the ICC as part of it. But an ICC is really a general way of describing reliability and an ICC is a special case of generalizability theory, and coefficient alpha is a special case of an ICC.

Q12. Are there concerns about reliability for online surveys?

There are different concerns for reliability of online surveys than there are for the reliability of pencil and paper surveys. Different people are going to fill them out more poorly or more accurately depending upon their particular background.

Some people aren't as familiar with computers and aren't able to navigate as well. In general, there are different sources of reliability. Obviously getting people to sit down face to face and get almost everybody to fill out your survey, that's going to help.

With online surveys, we know response rates tend to be a little bit lower and that may have an impact on your reliability especially if you want to calculate a group average for a teacher, for example. If you only have ten observations, your reliability is going to be low.