Welcome to Today’s Webinar!

Evaluating the Reliability of Surveys and Assessments

This event will start at 11:00 am EDT.

Welcome to Today's Webinar

Audio Information
Dial: 800-779-3152

Conference ID: 1935128

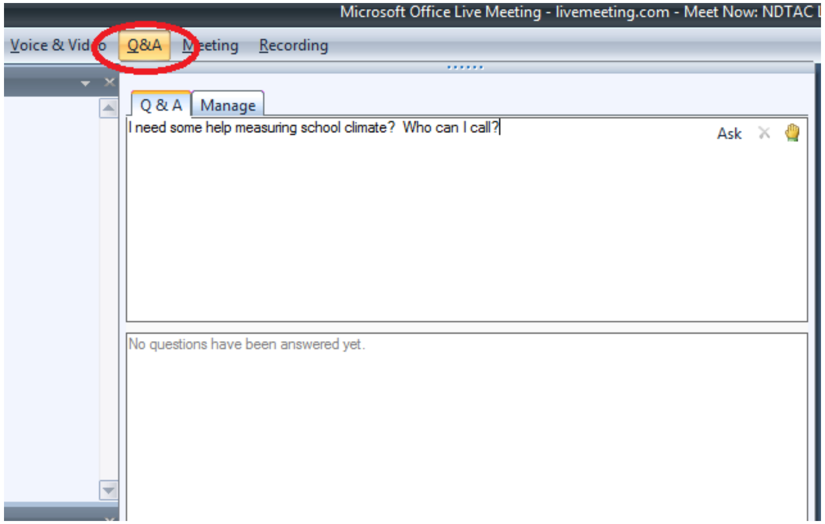
If you have technical difficulties logging into the web-based portion of the event, please contact Live Meeting Customer Support at 1 (866) 493-2825.

If you have any questions about the Live Meeting technology or the Webinar, please contact SSSTA at sssta@air.org.

Safe and Supportive Schools
Engagement | Safety | Environment

Questions, Event Evaluation & Contact Information

Q&A

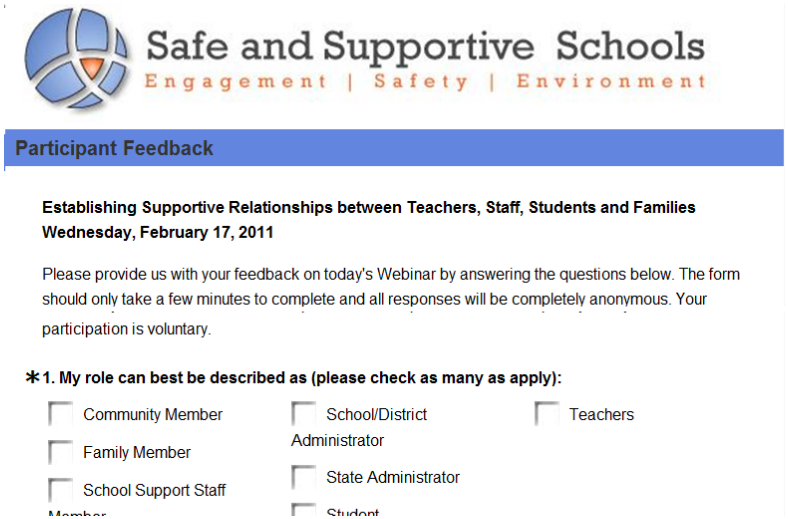

Microsoft Office Live Meeting - livemeeting.com - Meet Now: NDTAC L
Voice & Video
Q&A
Meeting
Recording
Q & A
Manage
Ask

I need some help measuring school climate? Who can I call?

No questions have been answered yet.

If you have a question for the presenters, please type it in the Q & A Pane or email sssta@air.org during the Webinar.

Evaluation



Safe and Supportive Schools
Engagement | Safety | Environment

Participant Feedback

Establishing Supportive Relationships between Teachers, Staff, Students and Families
Wednesday, February 17, 2011

Please provide us with your feedback on today's Webinar by answering the questions below. The form should only take a few minutes to complete and all responses will be completely anonymous. Your participation is voluntary.

***1. My role can best be described as (please check as many as apply):**

□ Community Member
□ Family Member
□ School Support Staff Member
□ School/District Administrator
□ State Administrator
□ Student
□ Teachers

An event evaluation will appear as the last slide in the presentation. Please input your answers *directly* into the slide. All answers are *completely anonymous* and are not visible to other participants.

For assistance during the Webinar,
please contact the Safe and Supportive Schools Technical Assistance Center at sssta@air.org.

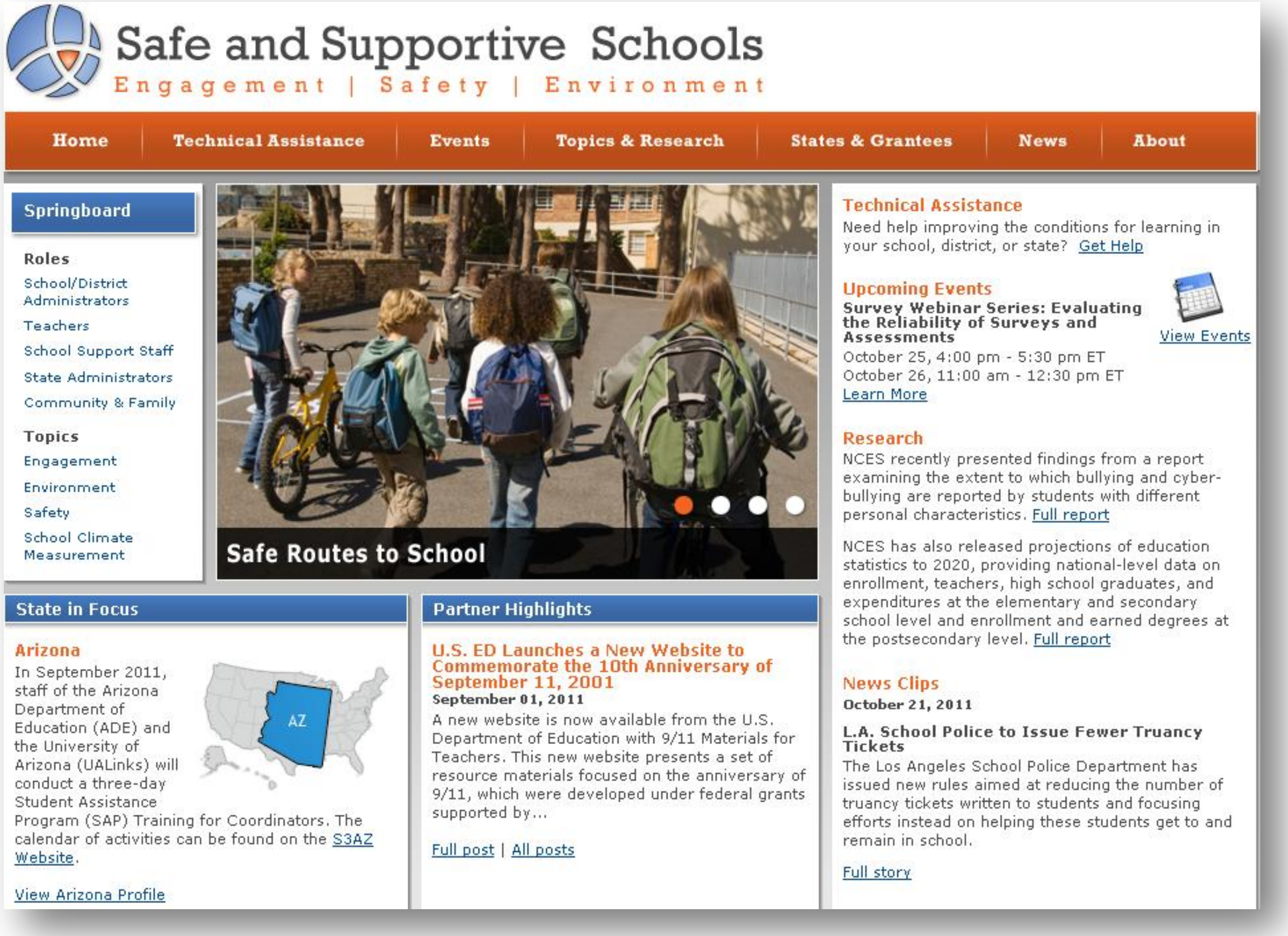
The Safe and Supportive Schools Technical Assistance Center

- Funded by the U.S. Department of Education's Office of Safe and Healthy Students.
- Provides training and support to states, including 11 grantees funded under the Safe and Supportive Schools Program and other state administrators; administrators of districts and schools; teachers; support staff at schools; communities and families; and students.
- Goal is to improve schools' conditions for learning through measurement and program implementation, so that all students have the opportunity to realize academic success in safe and supportive environments.

*The content of this presentation 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 presentation does not necessarily represent the policy or views of the U.S. Department of Education, nor do they imply endorsement by the U.S. Department of Education.

Safe and Supportive Schools Website

http://safesupportiveschools.ed.gov



Safe and Supportive Schools

Engagement | Safety | Environment

Springboard

Roles

- School/District Administrators
- Teachers
- School Support Staff
- State Administrators
- Community & Family

Topics

- Engagement
- Environment
- Safety
- School Climate Measurement

Safe Routes to School

Technical Assistance

Need help improving the conditions for learning in your school, district, or state? Get Help

Upcoming Events

Survey Webinar Series: Evaluating the Reliability of Surveys and Assessments

October 25, 4:00 pm - 5:30 pm ET
October 26, 11:00 am - 12:30 pm ET
Learn More

View Events

Research

NCES recently presented findings from a report examining the extent to which bullying and cyber-bullying are reported by students with different personal characteristics. Full report

NCES has also released projections of education statistics to 2020, providing national-level data on enrollment, teachers, high school graduates, and expenditures at the elementary and secondary school level and enrollment and earned degrees at the postsecondary level. Full report

News Clips

October 21, 2011

L.A. School Police to Issue Fewer Truancy Tickets

The Los Angeles School Police Department has issued new rules aimed at reducing the number of truancy tickets written to students and focusing efforts instead on helping these students get to and remain in school.

Full story

State in Focus

Arizona

In September 2011, staff of the Arizona Department of Education (ADE) and the University of Arizona (UALinks) will conduct a three-day Student Assistance Program (SAP) Training for Coordinators. The calendar of activities can be found on the S3AZ Website.

View Arizona Profile

Partner Highlights

U.S. ED Launches a New Website to Commemorate the 10th Anniversary of September 11, 2001

September 01, 2011

A new website is now available from the U.S. Department of Education with 9/11 Materials for Teachers. This new website presents a set of resource materials focused on the anniversary of 9/11, which were developed under federal grants supported by...

Full post | All posts

Safe and Supportive Schools
Engagement | Safety | Environment

Polling Question #1

Which of the following best describes your current role?

- ❑ State Education Personnel
- ❑ District or School Administrator
- ❑ Teacher or School Support Staff
- ❑ Community or Family Representative
- ❑ Student
- ❑ Researcher
- ❑ Other

Polling Question #2

Which of the following best describes the primary reason you chose to participate in today’s session?

- ❑ Learn what is measurement reliability and why it is important
- ❑ Learn more about generally how to evaluate my data to determine if it is reliable
- ❑ Learn about more advanced methods to conduct a reliability assessment
- ❑ Learn ways to improve reliability
- ❑ More than one of the above

Evaluating the Reliability of Surveys and Other Assessments

Dr. Lorin Mueller, American Institutes for Research
Washington, DC

Safe and Supportive Schools
Engagement | Safety | Environment

Session Overview

1 Purpose and definitions/key concepts

2 Statistical methods for assessing reliability

3 Common problems and how to resolve them

Notes Before We Begin

- **This is not a tutorial.**
  - The goal is to expand your thinking about measuring reliability in school climate surveys and other assessments.
  - There is plenty of practical guidance for these methods on the internet.
- **What do I mean by other assessments?**
  - When measuring school climate, you might want to relate it to other things: teacher evaluations, achievement levels, demographic characteristics.
  - Examples: Do boys rate climate differently than girls? Do certain teachers within the school foster a more supportive environment than others? Does that correlate to teacher evaluations?
  - Establishing reliability on these other measures is **just as important** as on the climate measures themselves.

Purpose of Reliability

Why do we want to demonstrate reliability in surveys?

1. **Better understand the thing we want to measure; sometimes *facets* don't correlate the way we expect them to.**

   *Facets - pieces of an instrument, like observations on different occasions or survey items

2. **To make better decisions based on the data we obtain.**

3. **Show that our results aren't erroneous/spurious.**

4. **Identify and correct/remove erroneous data.**

Definition & Key Concepts

- **Scale**: An item or set of items designed to measure something (a construct)
- **Item**: A single question on a scale
- **Item/scale types**
  - **Continuous/ratio**: true “zero” point, equally spaced units, measures can fall between units, e.g., height or normalized achievement test scores
  - **Interval (rating)**: equally spaced units, whole units only, e.g., “Likert-type” agreement, Olympic ratings
  - **Categorical (dichotomous/polytomous, ordinal, nominal)**: mutually exclusive categorical units, e.g., gender, ethnicity, grade level, symptoms/behavior checklist

*In most cases, we treat continuous & interval the same.

Scale Types: More Examples

- **Continuous/ratio**
  - Average standardized test score
  - Difference from the average classroom score on 10 school climate items (can be positive or negative)
- **Interval**
  - Agreement scales – “My teacher sets high expectations for achievement.” 1 = Strongly disagree, 3 = Neither agree nor disagree, 5 = Strongly agree
- **Categorical**
  - Education level of teacher; Achievement level of student (ordinal)
  - Gender/ethnicity of student (nominal)

Polling Question #3

Which of the following best describes your experience in conducting reliability assessments with survey items?

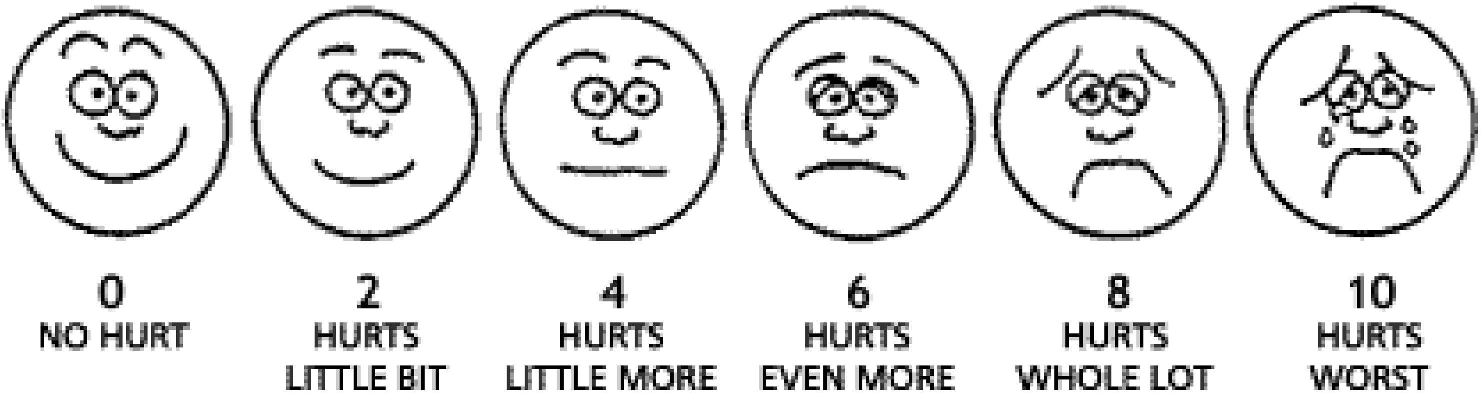
❑ We have not had much experience conducting analyses to determine reliability of survey items.

❑ We have experience generating alphas but nothing more advanced.

❑ We have experience conducting more advanced analyses such as factor analyses or HLM.

❑ We have experience with a range of analysis methods but want to learn more about improving the reliability of our perception-based survey items.

Interval Scale: My Favorite Example



- Above are examples of the “faces scale.” The faces represent different levels of satisfaction (left) or pain (right).
- This is an ordinal scale, but we usually treat it like it’s interval.
- We know the difference between 0-2 is not the same as the difference between 8-10. That lowers reliability, but not by much, and it makes the scale much easier to use.

What is Reliability?

- **Reliability**
  - Reproducibility of the scores over occasions, items, raters
  - Internal consistency, interrater agreement, test-retest/multiple occasions
  - E.g., the extent to which students within a classroom rate climate the same no matter which student is asked, what day you ask them, or which questions you ask
- **Reliability is a necessary condition for *validity***
- **Validity**
  - The extent to which the measure supports the intended inferences
  - Evidence: content, criterion-related, “construct”
  - E.g., Does a positive school climate correlate to important outcomes? Does changing school climate improve those outcomes? Does this measure of school climate correlate to other measures? Broader: Is school climate real?

Reliability Doesn’t Confer Validity

- **Reliability is necessary for validity, but…**
  - Not all reliable measures are valid & sometimes you may not be able to demonstrate reliability.
- **For example:**
  - What if students all rate climate as poor because the teacher is demanding? Or positive because the teacher gives high marks?
  - What if students respond to the questionnaire with answers they think the teacher or school wants them to give?
  - What if school climate doesn’t correlate to anything?
- **For validity, you need to collect evidence against these criticisms (other than reliability evidence).**

Three Facets of Reliability

- **Across items / internal consistency**
  - Answers the question “How much would scores change if I had selected a different set of items?”
  - **Important to have good coverage of the concept.**
- **Across raters / interrater agreement**
  - Answers the question “How much would scores change if there were different students in the classroom?” (or different observers)
  - **Important to establish common experience.**
- **Across occasions / stability**
  - Answers the question “How much would scores change if I measured the person at a different time?”
  - **Important that responses are stable from day to day.**

Assessing Reliability

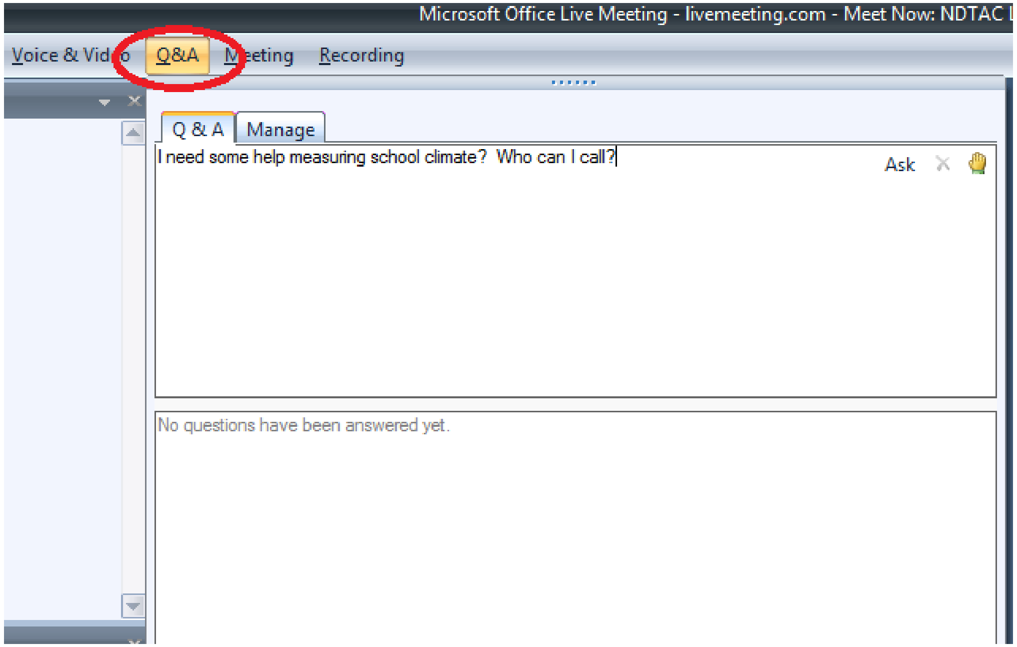
- **The most common method is coefficient alpha, a.k.a. “Cronbach’s alpha”.**
- **For situations where respondents are rating the same target, use interrater agreement.**
  - E.g., Students/parents in a classroom rate their teacher or school
- **When you are concerned that data might change, use a stability measure – correlation of measurements over time.**
- **If data are not continuous/interval, you will be limited in your approaches.**
  - Contingency tables for multiple variables / occasions; patterns (e.g. pairs of yes/no questions)
  - Agreement indices for multiple raters

Assessing Reliability Matrix

	Measurement Facets		
Data Type	**Items**	**Occasions**	**Raters**
Continuous / Interval	Coefficient alpha, factor analysis	Correlation over time, Generalizability	Intra-class correlation (ICC)
Categorical	Cross-classification / contingency tables	Cross-classification / contingency tables	% agreement, Cohen's kappa
Mixed	Pattern analysis		

Questions?

If you have a question for the presenter, please type it in the Q & A Pane or email sssta@air.org.



Microsoft Office Live Meeting - livemeeting.com - Meet Now: NDTAC I

I need some help measuring school climate? Who can I call?

No questions have been answered yet.

Coefficient Alpha

- **Internal consistency estimate**
- **Appropriate for continuous and interval data**
- **Available in most standard statistical packages**
- **Established ranges**
  - .70 ok for research, e.g., correlating climate to classroom achievement
  - .80 for diagnostic purposes, e.g., giving a teacher ways to improve classroom climate
  - .90 for high stakes decisions, e.g., negative sanctions for poor climate scores

Limitations of Coefficient Alpha

- **Can be high with lots of items**
- **Can overestimate reliability for temporal constructs (e.g., mood) and underestimate reliability for diverse constructs**
  - Important to link the time element to what you're trying to investigate.
  - If you want to know how mood on a given day influences perceptions of climate, alpha is OK.
  - If you want to know how mood influences achievement over the year, a one-day measurement will not suffice, nor will alpha.
- **Overused with categorical data**
- **Often incorrectly interpreted as validity evidence**

Coefficient Alpha - Formula

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{Y_i}^{2}}{\sigma_{x}^{2}}\right)$$

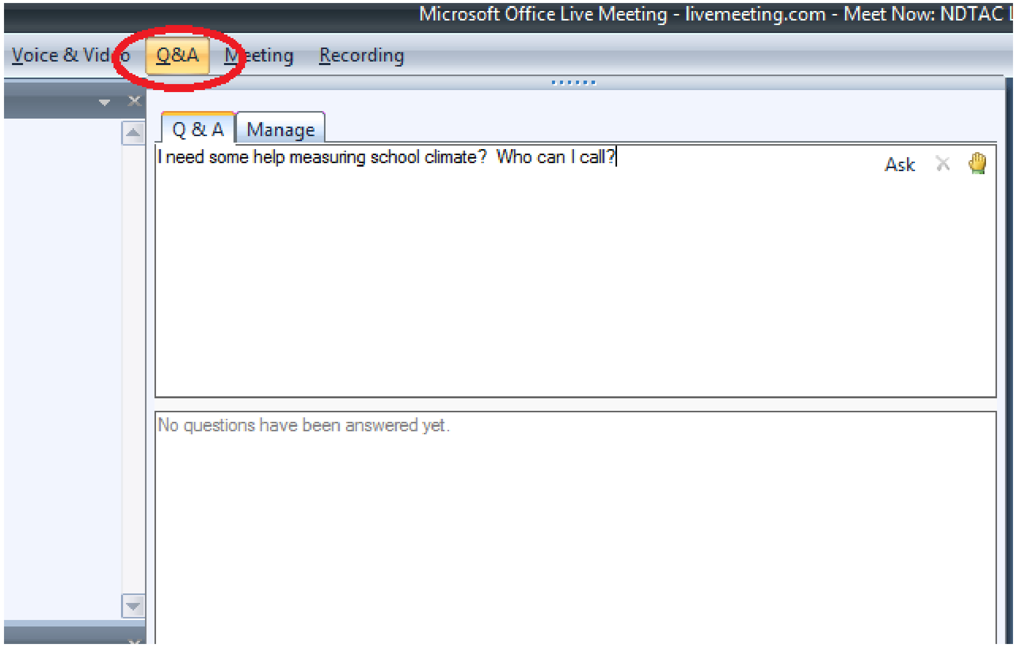
- K = number of items; σYi2 is the variance attributable to an item Yi ; σX2 is the total variance.
- Thus, as the variance due to items goes up, alpha goes down.
- It assumes variance not attributable to items is attributable to persons.

Exploratory Factor Analysis (EFA)

- **Seeing a lot of this in reliability studies now that EFA software is built in to standard statistical packages, guidance on the internet**
- **BUT:**
  - In most cases, you probably have a pretty good conceptual model that supports confirmatory factor analysis (CFA) – slightly different research question
  - Can get spurious results with small sample sizes, unhelpful results regardless
  - Reliability of EFA no better than rational scoring (e.g., *a priori* scales)
- **Especially bad with climate measures, since you can't usually model the sources of variation at each level: student, classroom, school**
- **For school climate measures, multi-level CFA is better: makes you specify how things *should* relate in advance, then tells you if you're wrong**

Questions?

If you have a question for the presenter, please type it in the Q & A Pane or email sssta@air.org.


Microsoft Office Live Meeting - livemeeting.com - Meet Now: NDTAC I
Voice & Video
Q&A
Meeting
Recording
Q & A
Manage
Ask

I need some help measuring school climate? Who can I call?

No questions have been answered yet.

Generalizability Theory

- **Extension of alpha method to account for multiple sources of variance**
  - Items, occasions of measurement, raters/judges
  - Each of these are “facets” of measurement and can have their own variance associated with them
  - Can estimate reliability in more complex designs or design data collection to meet reliability goals (D studies)
  - Coefficient alpha is a special case of Generalizability Theory
- **Think of this as the psychometric equivalent of HLM - can have “randomized” or “fixed” facets**

Random vs. Fixed Facets

- **Like in Hierarchical Linear Modeling (HLM), an advanced multi-level statistical technique, you can have random or fixed facets.**
- **A random facet is one that may change over time – like you might include slightly different items on a survey, or different students might rate the teacher.**
- **A fixed facet is one that you expect won't change.**
  - Example: You have a small set of evaluators visit classrooms to observe teachers. At the end of the year, you make decisions about teachers based on the evaluators' observations.
  - For this year, evaluators are a fixed facet.
  - For upcoming years, if evaluators might change, they are a random facet.
- **This approach to reliability can be very sophisticated and *informative*.**

Intraclass Correlation Coefficients

- Intraclass correlation coefficients (ICCs) are the proportion of variance attributable to a facet of measurement.
- Each ICC is a special case of Generalizability theory.
- Generally we use the ICC in cases where we have a lot of observations of a single target – such as students in a classroom rating a teacher: How reliable is that rating?
- You can compute the reliability for a single measurement or the classroom average (much higher).
- Generally interpret similar to alpha.

ICC Formulas

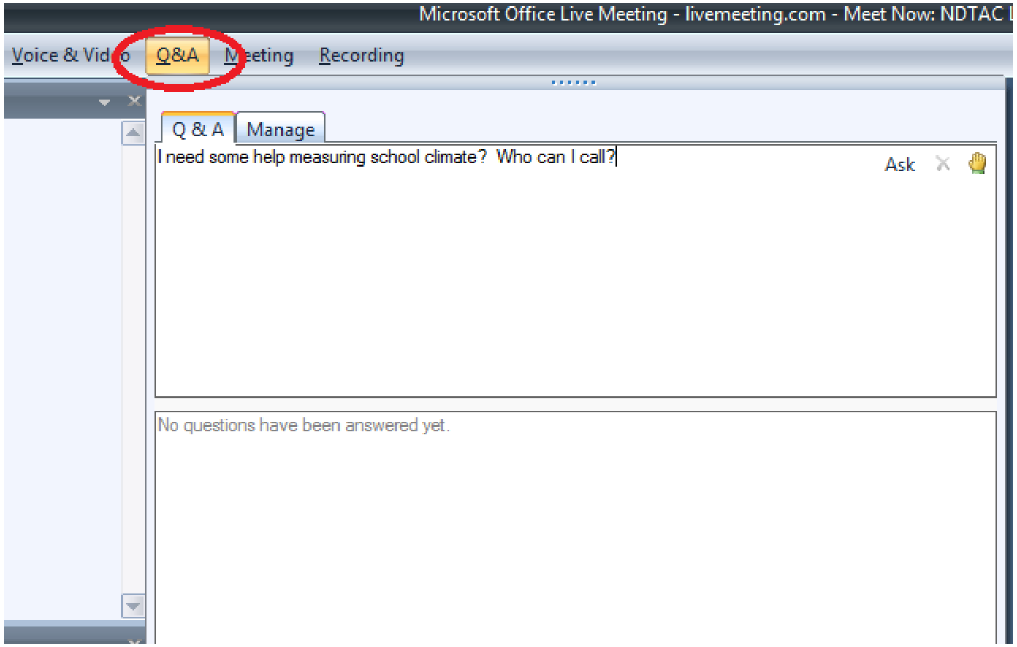
$$ICC1 = \frac{MS_{bg} - MS_{wg}}{MS_{bg} + (n-1) * MS_{wg}}$$

$$ICC2 = \frac{MS_{bg} - MS_{wg}}{MS_{bg}}$$

- MS_{bg} = mean square between groups (ANOVA)
- MS_{wg} = mean square within groups
- n = average group size

Questions?

If you have a question for the presenter, please type it in the Q & A Pane or email sssta@air.org.


Microsoft Office Live Meeting - livemeeting.com - Meet Now: NDTAC l
Voice & Video
Q&A
Meeting
Recording
Q & A
Manage
Ask

I need some help measuring school climate? Who can I call?

No questions have been answered yet.

Categorical Data

- **Most commonly use cross-tabulations (cross-classification, contingency) to check for the consistency of reported codes across variables**
  - Multiple measures for contingency tables: Sensitivity, specificity, hits, misses
  - Can be applied across items within a measure or across time with the same item
- **Applied three ways**
  - If the items ask for the same information, check the percentage correspondence (sex, gender).
  - If the items ask for related information, check the **association** (school bullying incidents, reported police visits).
  - Identify unlikely response combinations (Student reports high achievement levels, no academic extracurricular activities).

2 x 2 Contingency Table

	Self-Reported Gender		
Sex from Database	**Girl**	**Boy**	**Total**
Female	26	**1**	27
Male	**3**	14	17
Total	**29**	**15**	**44**

Categorical Data - Agreement

- **Most common indices are percentage agreement.**
  - % Exact is almost always reported.
  - % Adjacent is often included for multiple ordered levels.
- **Contingency tables**
- **Best measure depends on how disagreement affects the decisions you want to make.**
- **There are more sophisticated indices, but little agreement over which is best (ironically).**

Pattern Analysis

- **Examine records for unlikely patterns of responses**
- **Can use this method for continuous/interval, categorical, or mixed data**
- **Methods**
  - Contingency tables
  - Group means/outlier analysis
  - Unlikely strings of the same response (A, A, A, A…)
  - Inconsistent responses to reverse-coded items
  - “Honesty” scales – items no one should endorse
- **“Erasure analysis” used to detect cheating in Atlanta and DC, in conjunction with changes in group means**

Tough Questions

- **What do my data look like?**
- **Are all of your observations independent? If not, you might want to look at agreement indices and multi-level models.**
- **How am I going to use the data?**
  - At what level? (group, individual)
  - How many observations?
  - What variables are critical to my work?
- **What is the best thing to do with seemingly unreliable responses?**

Handling Unreliable Responses

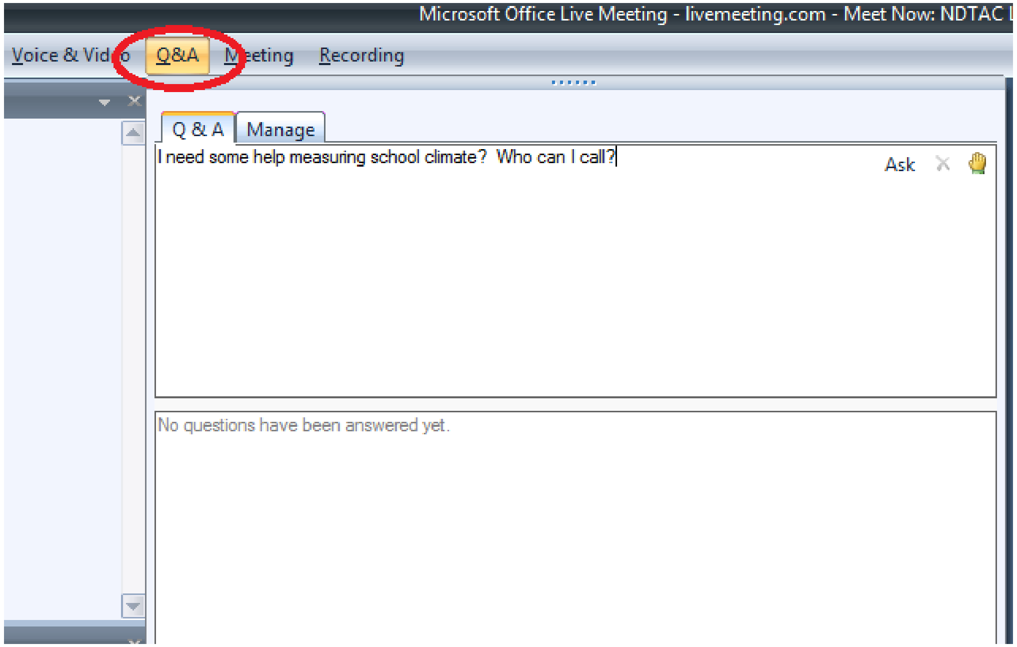
- **Follow-up to clarify if possible** – new survey software can detect inconsistent responses and ask for a confirmation; build into interview protocols
- **Recode or correct variable** – good if there is a lot of evidence it was a mistake (e.g., misreporting gender in a longitudinal survey)
- **Delete offending variable** – good if the variable isn't critical and you want to retain the other variables; fill in "human" in ethnicity
- **Delete offending case** – good if the response makes the entire case questionable; pattern analysis
- **Ignore it** – acceptable if it is not expected to influence your results

Improving Reliability

- Make the questions as simple as possible.
- If you are relying on demographics, put them up front and ask questions in different ways and explain why you’re asking.
- Try to build in a re-administration of some of the survey questions over a period where you would not expect things to change.
- Know when responses might change and how that will affect your results. Counterbalance conditions that have a strong impact on your measurement (e.g., time of year, rating source, item referent).

Questions?

If you have a question for the presenter, please type it in the Q & A Pane or email sssta@air.org.


Microsoft Office Live Meeting - livemeeting.com - Meet Now: NDTAC
Voice & Video
Q&A
Meeting
Recording
Q & A
Manage
Ask

I need some help measuring school climate? Who can I call?

No questions have been answered yet.

Safe and Supportive Schools
Engagement | Safety | Environment

Citations

1. Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), Multilevel Theory, Research, and Methods in Organizations (pp. 349-381). San Francisco, CA: Jossey-Bass, Inc.
2. Hockenberry M.J., & Wilson D. (2009). Wong’s Essentials of Pediatric Nursing (8th Ed.). St. Louis: Mosby.
3. Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York, NY: McGraw-Hill.
4. Shavelson, R. J., & Webb, N. M. (1991). Generalizability Theory: A Primer. Newbury Park, CA: Sage Publications.

Suggested Readings

1. Abelson, R. P. (1995). *Statistics as a Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.
2. Allen, M. J., & Yen, W.M. (2001). Introduction to Measurement Theory Waveland Pr Inc.
3. Crocker, L., & Algina, A. (2006). *Introduction to Classical and Modern Test Theory*. Wadsworth Pub Co.
4. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE.
5. Lord, F. M., & Novick, M. R. (2008). *Statistical Theories of Mental Test Scores.* Addison-Wesley Publishing Company, Inc.
6. Nunnally, J., & Berstein, I. (1994). Psychometric Theory. McGraw-Hill.
7. Shavelson, R. J., & Webb, N M. (1991). *Generalizability Theory: A Primer*. Thousand Oaks, CA:SAGE.
8. Shavelson, R. J, Rowley, G.L., & Webb, N. M. (1989). Generalizability theory. *American Psychology*, *44*, 922-932.
9. Suen, H. K. (1990). Principles of Test Theories. New York: Routledge.