VM²ED
Validity Evidence for Measurement
in Mathematics Education

Evidence Types Guidebook

Brianna Bentley[1], Timothy Folger[2], Jonathan Bostic[2], Erin Krupa[1], Katie Burkett[1], and David Stokes[1]

[1]Department of STEM Education, North Carolina State University

[2]School of Inclusive Teacher Education, Bowling Green State University

If you would like to reference this material, then please use the following citation:

Bentley, B., Folger, T., Bostic, J., Krupa, E. E., Burkett, K., & Stokes D. (2024). *Evidence types guidebook.* Validity Evidence for Measurement in Mathematics Education. https://www.mathedmeasures.org

This material is based upon work supported by the National Science Foundation under Grant No. (DRL #1920619 & #1920621). Any opinions, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Test Content

From *Standards for Educational & Psychological Testing*:

"Test content refers to the themes, wording, and format of the items, tasks, or questions on a test. Administration and scoring may also be relevant to content-based evidence...Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on content can also come from expert judgments of the relationship between parts of the test and the construct" (AERA et al., 2014, p. 14).

Alignment with frameworks/standards/theory/learning trajectory: Quality and strength of connection between an item or item set and a desired framework, set of standards, or learning trajectory.

Construct Definition: Constructs are broad concepts or topics for a study. Constructs can be conceptually defined in that they have meaning in theoretical terms. They can be abstract and do not necessarily need to be directly observable. Examples of constructs include intelligence or life satisfaction (Lavrakas, 2008).

Data from experts: Information from experts in the field that has been collected, observed, generated or created to support original research (University of Leeds Library, 2017).

Fairness of content: Fairness is broadly defined as equitable treatment of all test takers during the testing process, relative absence of measurement bias, equitable access to the constructs being measured, and justifiable validity of test score interpretation for the intended purpose(s). Given that test fairness is closely related to the interpretations and uses of test scores as well as the claims made from those interpretations and uses, it is critical to obtain and weigh validity evidence to support or refute the score interpretations, their uses, and the potential socio-political consequences in order to evaluate fairness (Banerjee, 2016).

Field Work: Practical work conducted by a researcher in the natural environment, rather than a laboratory or office ("Fieldwork Definition," n.d.).

Literature Review: A literature review is a comprehensive summary of previous research on a topic. The literature review surveys scholarly articles, books, and other sources relevant to a particular area of research. The review should describe, summarize, objectively evaluate and clarify the previous research in the field (American Psychological Association, 2020).

Participant-generated content: Data that have been created by the participant of the research study for use by the researcher. These data may support test content in providing baseline support for connections to a desired construct.

Revision Process: During the analysis of initial results, if researchers discover that the assessment does not effectively measure the outcomes as desired, then the assessment method is revised. revisions entail rewording questions, replacing questions, and changing question formats (Assessment of Student Learning, n.d.).

Standard Setting: A process where levels of achievement or proficiency are defined and the corresponding cutscores are established. A cutscore classifies individuals below the score into one level and those above into the next, higher level. The rationale and procedures used for creating the cutscores should be clearly described (Bejar, 2008).

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org.10.1037/0000165-000

Assessment of Student Learning. (n.d.). The Assessment Process. Retrieved July 09, 2020, from https://www.missouristate.edu/assessment/the-assessment-process.htm

Banerjee, H. L. (2016). Test Fairness in Second Language Assessment. *Studies in Applied Linguistics & TESOL, 16*(1), 54-59.

Bejar, I. I. (2008). Standard setting: What is it? Why is it important?. *R&D Connections*, *7*, 1-6.

Fieldwork: Definition of Fieldwork by Oxford Dictionary on Lexico.com also meaning of Fieldwork. (n.d.). Retrieved July 09, 2020, from https://www.lexico.com/en/definition/fieldwork

Lavrakas, P. J. (2008). Construct. Retrieved July 09, 2020, from https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n91.xml

University of Leeds Library. (2017, May 12). Research data management explained. Retrieved July 09, 2020, from https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

Response Processes

<u>From *Standards for Educational & Psychological Testing*:</u>

“Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers.” (AERA et al., 2014, p. 15)

Cognitive interview: Interview procedure meant to explore a participant’s comprehension of an item or task (Leighton, 2017).

Error related to response patterns/CTT: Classical Test Theory (CTT) predicts the true score of a test-taker by accounting for and calculating an error score (Novick, 1966). Standard error, as seen in CTT, may highlight ideas about potential noise in the data that could in turn, highlight respondents' performance or engagement (Crocker & Algina, 2006).

Eye tracking/physiological data: Physiological data involves the empirical observation of variables pertaining to the functioning of systems in the human body. Eye tracking data are collected by measuring the movements of the eye, including fixation and duration. It is inferred that eye movement is a proxy for attention and in turn, evidence of participants' responses. These data provide developers with evidence of comprehension processes that take place while reading (Paulson & Henry, 2002).

fMRI (Functional Magnetic Resonance Image): Neuroimaging that measures brain activity by detecting changes in blood oxygenation and flow. fMRI data could be used to measure changes in brain activity in response to an assessment item (Byars et al., 2002).

Focus groups: Interview technique of collecting data by orchestrating discourse with a small group of individuals. Participants of the focus group discuss feelings or thoughts regarding various elements of a test item. Item developers can use these data to examine the psychological processes of test-takers (Padilla & Benitez, 2014).

Generalizability-theory (G-theory) related evidence: Generalizability theory provides a conceptual and statistical framework to model multiple sources of error in assessment data. This enables assessment developers to quantify and address inconsistencies in observed scores that could develop over replications of the assessment. G-theory evidence related to response processes includes variance in how test-takers respond to items, and rater-error (Lane, 2019; Brennan, 2001).

Log data: Log data provides a record of actions taken while working through an item or task (Stadler et al., 2020).

Predicted response patterns/processes based on Learning Trajectories: Learning Trajectories refers to the progression of student thinking during the learning of mathematics. This coincides with progress levels that can be used to make generalizations of student thinking (Confrey et al., 2019).

Rater agreement/reliability: The consistency of scores assigned by two or more independent raters of the same performance. Evaluation of rater reliability is needed to promote valid score interpretations and uses. Construct-irrelevant variance may occur as a result of poor rater agreement/reliability. Sample quantitative forms of rater agreement/reliability include using ICC, rwg, Kappa, and percent agreement (Lane, 2019).

Rater training and calibration: Developing rater agreement by establishing consistency among raters. Construct-relevant variance is affected by aspects of rater training such as training materials, training procedures, and rater calibration (Lane, 2019).

Sorting tasks: The process of grouping a set of items into categories based on meaning. Sorting tasks can be used to elicit the knowledge structure of a respondent (Tang and Clariana, 2017).

Studies of respondent's speed of task completion (Response times): Measure of time required to complete a task provide evidence of how respondents engage with the item/set of items in desired or undesired ways. Typically, this is in regards to response times focus on the relationship between response time and cognitive demand of the item (Padilla & Benitez, 2014).

Think alouds: Interview procedure used to measure problem-solving processes by asking the participant to articulate their thoughts in response to an item or task. Think alouds provide item developers the opportunity to examine the cognitive processes of test-takers (Leighton, 2017; Padilla & Benitez, 2014).

Written work: Any written work documented by the test-taker that can be used to identify solution strategies and provide evidence of their thinking. This includes "scratch-work" as well as written justifications (Watson, 1995).

References

Brennan, R. L. (2001). *Generalizability Theory.* New York, NY: SPRINGER.

Byars, A. W., Holland, S. K., Strawsburg, R. H., Bommer, W., Dunn, R. S., Schmithorst, V. J., & Plante, E. (2002). Practical aspects of conducting large-scale functional magnetic resonance imaging studies in children. *Journal of Child Neurology, 17*(12), 885-889.

Confrey, J., Toutkoushian, E., & Shah, M. (2019). A validation argument from soup to nuts: Assessing progress on learning trajectories for middle-school mathematics. *Applied Measurement in Education: Argument-Based Validation in Practice: Examples from Mathematics Education, 32*(1), 23-42.

Crocker, L. & Algina, J. (2006). *Introduction to classical and modern test theory.* Mason, OH: Thomson Wadsworth

Lane, S. (2019). Modeling rater response processes in evaluating score meaning. *Journal of Educational Measurement, 56*(3), 653-663.

Leighton, J. P., & Ohio Library and Information Network. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18.

Padilla, J., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144.

Paulson, E. J., & Henry, J. (2002). Does the degrees of reading power assessment reflect the reading process? an eye-movement examination. *Journal of Adolescent & Adult Literacy, 46*(3), 234-244.

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior, 111*, 106442.

Tang, H., & Clariana, R. (2016). Leveraging a sorting task as a measure of knowledge structure in bilingual settings. *Technology, Knowledge and Learning, 22*(1), 23-35.

Watson, A. (1995). Evidence for pupils' mathematical achievements. *For the Learning of Mathematics, 15*(1), 16-28.

Internal Structure

From *Standards for Educational & Psychological Testing*:

"Analyses of the internal structure of a test can indicate the degree to which the relationships among the items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p.16). "If the rationale for a test score interpretation for a given use depends on the premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided" (AERA et al., 2014, p. 26).

Bayesian Network Models: Graphical model denoting the dependence structure of a decomposed joint probability distribution. The graph of the model may depict a causal schematic. The model may allow for the prediction of events given observed information. Evidence may include how the node probabilities were derived (Horný, 2014).

Cluster analysis: Grouping method for items based on a similarity measure, like distance, or a statistical mixture model. Within the context of internal validity, derived clusters should correspond to the construct groups and/or indicate sufficient groupings. Evidence may include model selection. (Hastie, Tibshirani, Friedman, 2009; Johnson & Wichern, 2015; Zelterman, 2015)

Factor analysis: models a set of observable random variables (to be collected) as weighted sums (linear combinations) of fewer, unobservable (latent) variables/constructs/factors. Variable relationships are specified through loadings on the set of common factors. These loadings indicate the strength of the relationship between the variables and the factors. Variables representative of a given factor should have large (absolute value) loadings on that factor. For valid constructs, various methods should tend to agree (Hancock, Stapleton, Mueller, 2019; Hastie, Tibshirani, & Friedman, 2009; Johnson & Wichern, 2015; Zelterman, 2015).

> **Factor Analysis - Bifactor:** This rotation method utilizes a "bi-factor" rotation in the factor analysis. The bi-factor rotation constrains the loadings for each variable to be primarily on 2 factors. An example use of this model would be to validate a posited construct set that explains p characteristics with$m$ factors consisting of one general factor and $m - 1$sub-group factors. For instance 6 items, represented by 3 factors: 1 general factor and 2 sub-group factors. Evidence would show that each of the six factors would primarily relate to (have high loadings on) the general factor and/or one subgroup factor (Jennrich & Bentler, 2011).
>
> **Factor Analysis - CFA (Confirmatory Factor Analysis):** The number of factors is determined *a priori* (i.e. based on theory) (Hancock, Stapleton, Mueller, 2019; Kline, 2015).
>
> **Factor Analysis - EFA/ESEM (Exploratory Factor Analysis/Exploratory Structural Equation Modeling):** The number of factors is derived from the data (as opposed to a theory-based, or hypothesized from the literature-based, prior specification of the factors/constructs) (Hancock, Stapleton, Mueller, 2019; Kline, 2015).

Factor Analysis - MTMM (Multi-trait Multi-method matrix): Uses factor analysis to assess relationships between multiple constructs measured in different ways (Shen, 2017).

Factor Analysis - Parallel Analysis: Method for selecting the number of factors to retain during an exploratory factor analysis. The ordered eigenvalues of the observed data correlation matrix are compared to quantiles obtained from the aggregated (by order) correlation matrix eigenvalues of randomly generated datasets of the same size. The comparison is usually visualized through a plot, with order decreasing from left to right. The number of observed/sample eigenvalues larger than the generated average or specified quantile eigenvalues is the number of factors to retain. This method may be more accurate than the scree plot, or K1 rule (Hayton, Allen, & Scarpello, 2004).

Factor Analysis - PAF (Principal Axis Factoring): This method uses an iterative procedure to determine the loadings and other estimates (Johnson & Wichern, 2015).

Factor Analysis - PCA (Principal Component Analysis): Method where the estimates (loadings, etc.) are derived from the principal components solution (Johnson & Wichern, 2015).

Item difficulty: Item difficulty indices allow “researchers to compare the predicted order of item difficulty with the actual order of item difficulty in a data set” (Boone, 2016, p. 4). “An item exhibiting difficulty higher than the ability level of the respondent will have a lower probability of being correctly answered than an item of difficulty below the ability level of the respondent” (Boone, 2016, p. 3).

IRT (Item Response Theory): Is a collection of methods, based on the belief that a given instrument is an indicator of latent (unobservable) ability. Used to validate the placing of individuals along a given construct continuum. Can be used as evidence of an item's relationship to a latent ability measure, where certain items are more informative (should have higher weight) than others (Yang, 2014). In IRT the individual item is the unit of interest, as opposed to classical test theory the entire instrument is the unit of focus in deriving scores. (De Ayala, 2013)

LCA (Latent Class Analysis): Models different group representations through probabilities of an outcome given a classification. In this model, factor analysis loadings representing correlations are replaced with conditional probabilities (given a class/factor) (Hagenaars & McCutcheon, 2002; Hancock, Stapleton, Mueller, 2019).

LPA (Latent Profile Analysis): Similar to latent class analysis where the measurements are continuous as opposed to binary. Used to identify response patterns across groups (Hancock, Stapleton, Mueller, 2019).

Multidimensional scaling: Dimension reduction & visualization method that seeks to preserve distances or similarities in a smaller dimension than that of the original items or variables. Two dimensional representations may allow for graphical displays of an expected or specified construct group, or some other grouping of interest. Validity evidence might also include the rationale for the chosen measures of similarity/dissimilarity (Hastie, Tibshirani, Friedman, 2009; Johnson & Wichern, 2015).

Rasch modeling: A logistic regression model focusing on the probability of success as a function of the difference between a respondent's ability and item difficulty. Item difficulty is considered in score comparisons. Can be used to validate the range and representation of items along a construct. Such a model might indicate the validity or comparability of alternate test forms with respect to a particular concept (Boone, 2016). This method is derived from Rasch (1966).

TETRAD: Software focusing on causal models (Philosophy Department of Carnegie Mellon University, 2024).

References

Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how?. *CBE—Life Sciences Education*, *15*(4), rm4.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge University Press.

Hancock, G. (Ed.), Stapleton, L. (Ed.), Mueller, R. (Ed.). (2019). The Reviewer's Guide to Quantitative Methods in the Social Sciences. New York: Routledge, https://doi-org.prox.lib.ncsu.edu/10.4324/9781315755649

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, *7*(2), 191-205.

Hastie, T., R. Tibshirani, and J. Friedman. "The elements of statistical learning (2nd), ed." (2009).

Horný, M. (2014). Bayesian networks: A Technical report. *no*, *5*, 15.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*(4), 537-549.

Johnson, R. A., & Wichern, D. W. (2015). Applied multivariate statistical analysis. *Statistics*, *6215*(10), 10.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Philosophy Department of Carnegie Mellon University. (2024). *Tetrad* [Computer software]. Carnegie Mellon University. https://www.cmu.edu/dietrich/philosophy/tetrad/index.html

Rasch, G. (1966). An item analysis which takes individual differences into account. *British journal of mathematical and statistical psychology*, *19*(1), 49-57.

Shen, F. (2017). Multitrait‐Multimethod Matrix. *The international encyclopedia of communication research methods*, 1-6.

Yang, F. M. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171.

Zelterman, D. (2015). *Applied multivariate statistics with R*. Cham: Springer.

Relations to Other Variables

From *Standards for Educational & Psychological Testing*:

Relations to other variables may provide evidence, for example, that indicates how "...test scores [may or may not be] influenced by ancillary variables such as [individual or group characteristic]" (AERA et al., 2014, p.12). In addition: "Inferences about processes involved in performance can… be developed by analyzing the relationship among parts of [a] test… and between other variables" (AERA et al., 2014, p.15). "... Analyses of the relationship of test scores to other variables external to the test provide [an] important source of validity evidence" (AERA et al., 2014, p.16).

Alignment with expert opinion of test user (e.g. teacher, therapist): The idea that measured associations coincide with expert opinions, perhaps as a validation of exploratory methods, hypotheses or interpretations (AERA et al., 2014).

Convergent Association or Divergent Association: Associations/measurements related to convergent and discriminant validity. "Convergent validity refers to the degree to which two theoretically related measures of constructs are in fact related whereas discriminant validity is the degree to which two measures that are supposed to be unrelated are not interrelated in reality" (Shen, 2017, p.1). As a point of clarity, scores, outcomes, and interpretations are generally

Correlation analysis: Methods for assessing the associations between variables, or groups of variables. May indicate if certain variables tend to increase together, or if a variable tends to increase while another decreases, etc. May indicate associations between constructs. Correlations are between -1 and 1. Strong relationships are close to 1 or -1. Measures close to zero indicate weak relationships. Statistical independence implies zero correlation (Johnson & Wichern, 2015).

Discriminant validity: "... Discriminant validity is the degree to which two quantities that are supposed to be unrelated are not interrelated in reality." (Shen, 2017).

Discrimination Power: Discrimination power reflects the relationship between item performance and respondent ability. "Items are most effective when they discriminate well between students with high or low total scores" (Jorion et al., 2015; p. 458).

Hierarchical Linear Modeling: Models relationships regarding parameters of interest at various levels according to a nested (or hierarchical) structure (i.e. classroom level means within school level means). Variable relationships can be assessed by considering and testing whether populations differ when parameters depend on the variables, or interactions of interest (Raudenbush & Bryk, 2002).

Multi-trait Multi-method matrix (MTMM): A matrix including multiple constructs measured in various ways. Values may be measures of correlation (Shen, 2017).

Statistical Testing (e.g., t-test, regression, and chi-square): Statistical tests are used in hypothesis testing. They can be used to determine whether a predictor variable has a statistically significant relationship with an outcome variable.

Structural Equation Model: Various methods for multivariate data. Models may specify variable relationships. (Kline, 2015). Schreiber et. al (2006) quote Ullman (2001) as saying "*SEM* has been described as a combination of exploratory factor analysis and multiple regression" (p.324).

Treatment/control study: A research design where treatment effects of one group that received the treatment are compared to a control group that did not receive the treatment (Shavelson, 1996).

Triangulation/crystallization with qualitative data: is used in various contexts, including "... as a means of addressing qualitative/quantitative differences" (Tobin & Begley, 2004, p.392). Crystallization extends triangulation to agreements between more sources of information (Tobin & Begley, 2004).

References

Hastie, T., R. Tibshirani, and J. Friedman. "The elements of statistical learning (2nd), ed." (2009).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Johnson, R. A., & Wichern, D. W. (2015). Applied multivariate statistical analysis. *Statistics*, *6215*(10), 10.

Johnson, R. L., & Morgan, G. B. (2016). *Survey scales: A guide to development, analysis, and reporting*. Guilford Publications.

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, *104*(4), 454-496.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research*, *99*(6), 323-338.

Shavelson, R. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Boston, MA: Allyn & Bacon.

Shen, F. (2017). Multitrait‐Multimethod Matrix. *The international encyclopedia of communication research methods*, 1-6.

Tobin, G. A., & Begley, C. M. (2004). Methodological rigour within a qualitative framework. *Journal of advanced nursing*, *48*(4), 388-396.

Zelterman, D. (2015). *Applied multivariate statistics with R* (p. 54). Cham: Springer.

Consequences of Testing

From *Standards for Educational & Psychological Testing*:

“Some consequences of test use follow directly from the interpretation of test scores for uses intended by the test developer. The validation process involves gathering evidence to evaluate the soundness of these proposed interpretations for their intended uses. Other consequences may also be part of a claim that extends beyond the interpretation of use of scores intended by the test developer. For example, a test of student achievement might provide data for a system intended to identify and improve lower-performing schools… Still other consequences are unintended, and are often negative. For example, school district or statewide educational testing on selected subjects may lead teachers to focus on those subjects at the expense of others… Unintended consequences merit close examination” (AERA et al., 2014, p. 19).

Appropriate cut score: If it is necessary to establish a cut score, then measurement error and adverse impact should be considered when establishing a cut score. The cut score should reflect the intended interpretation and use of test scores proposed by the test developers (Cascio & Aguinis, 2005).

Bias as one consequence of testing: Bias as a consequence of testing pertains to adverse social consequences that can be attributed to a source of test invalidity, such as construct-irrelevant variance (Messick, 1989).

Cost-benefit analysis: Systematic approach evaluating the value, or future benefits, against the cost and/or negative consequences of an assessment program (Kane, 2013).

Documentation of unintended behavior changes based on test use:

Behavioral Changes in Teachers: Unintended behavioral changes in teachers may refer to morale, stress, and ethical behavior (Mehrens, 1998).

Behavioral Changes in Students: Unintended behavioral changes in students may refer to self-concept and academic confidence (Mehrens, 1998).

Explicit intended uses and interpretations and warn against inappropriate uses: Validation logically starts by explicitly stating the proposed interpretation and use of test scores. Additionally, developers should describe expected consequences and caution against inappropriate uses (AERA et al., 2014). The evaluation of the inferences and assumptions inherent to the proposed interpretation and use presents a defense against inappropriate interpretations and uses of test scores (Kane, 2013).

Impact of assessment is similar under clinical and practical implementations: Intended consequences of testing should be clearly outlined during validation. When it is stated or implied that a proposed interpretation and use will result in a certain outcome, there should exist evidence forming a basis for expecting that outcome. Those who mandate test use should measure the impact of the assessment and minimize any negative consequences (AERA et al., 2014).

Item functioning such as DIF - unknown subgroups had to know: Statistical techniques useful for identifying variances in performance by subgroups. Differences in performance can be analyzed at both test and item levels (Osterlind & Everson, 2009).

Motivational consequences:

Teacher & Administration Motivation: Teacher and administration motivation refers to the effort applied to improve student learning as a result of the assessment (Lane & Stone, 2002).

Student Motivation: Measures of student motivation often focus on the effort put forth by students while completing the assessment. This effort is influenced by factors such as teacher attitude and the stakes for the individual student. Measures of student motivation may also focus on effort put forth by the student throughout the academic year (Lane & Stone, 2002).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing.

Cascio, W. F., & Aguinis, H. (2005). Test development and use: New twists on old questions. *Human Resource Management, 44*(3), 219-235.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 23-30.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13), 13.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Second ed.). Thousand Oaks, Calif: SAGE.

Reliability

<u>From Standards for Educational & Psychological Testing:</u>

"The term reliability has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effort on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported" (AERA et al., 2014, p. 33).

<u>Alternate form:</u> Alternate-form reliability is the consistency of test results between two different – but potentially equivalent – forms of a test. Alternate-form reliability is used when it is necessary to have two forms of the same tests. To determine alternate form reliability two forms of the same test are administered to students and students' scores are correlated on the two test forms. The resulting coefficient is called the alternate form coefficient of reliability. Alternate form reliability is needed whenever two test forms are being used to measure the same thing. Ideally, the administration of the two forms should be done in a short time span (APA Dictionary of Psychology, n.d.).

<u>Generalizability theory- D-studies:</u> G-theory recognizes that an assessment might be adapted for particular decisions and so distinguishes a generalizability (G) study from a decision (D) study. A D-study typically selects only some facets for a particular purpose, thereby narrowing the score interpretation to a universe of generalization. A different generalizability (reliability) coefficient can then be calculated for each particular use of the assessment. By employing simulated D studies, it is therefore possible to examine how the generalizability coefficients would change under different circumstances, and consequently determine the ideal conditions under which our measurements would be the most reliable (Shavelson & Webb, 2006).

<u>Generalizability theory- G-studies:</u> Generalizability (G) theory is a statistical theory for evaluating the dependability (or reliability) of behavioral measurements. In a G-study, the universe of admissible observations is defined as broadly as possible (items, occasions, raters if appropriate, etc.) to provide variance component estimates to a wide variety of decision makers. G-theory estimates the components of observed-score variance contributed by the object of measurement, the facets, and their combinations. In this way, the theory isolates different sources of score variation in measurements (Shavelson & Webb, 2006).

<u>Inter-rater reliability- Kappa:</u> Kappa is a way of measuring agreement or reliability, correcting for how often ratings might agree by chance. Cohen's kappa, which works for two raters, and Fleiss' kappa, an adaptation that works for any fixed number of raters, improve upon the joint probability in that they take into account the amount of agreement that could be expected to occur through chance. Kappa is similar to a correlation coefficient in that it cannot go above +1.0 or below -1.0. Because it is used as a measure of agreement, only positive values would be expected in most situations; negative values would indicate systematic disagreement. Kappa can only achieve very high values when both the agreement is good and the rate of the target condition is near 50% because it includes the base rate in the calculation of joint probabilities (Gwet, 2014).

Inter-rater reliability- Percent agreement: The degree of agreement among raters. It is a score of how much homogeneity or consensus exists in the ratings given by various judges. Greater agreement is typically signaled by agreement value closer to 1 and little agreement is closer to a value of 0 (Gwet, 2014).

Internal consistency or alternatives- Alpha: A statistic calculated from the pairwise correlations between items. Internal consistency ranges between negative infinity and one. Coefficient alpha will be negative whenever there is greater within-subject variability than between-subject variability (Deng & Chan, 2017).

Internal consistency or alternatives- IRT Reliability: Reliability in IRT is defined as a function that is conditional on the scores of the measured latent construct. Precision of measurement differs across the latent construct continuum and can be generalized to the whole target population. In IRT, measurement precision is often depicted by the information curves. These curves can be treated as a function of the latent factor conditional on the item parameters. They can be calculated for an individual item (item information curve) or for the whole test (test information curve). The test information curve can be used to evaluate the performance of the test. During test development, you want to make sure that the selected items can provide adequate precision across the interested range of the latent construct continuum (An & Yung, 2014).

Internal consistency or alternatives- Omega: Omega determines the reliability of a test score. Coefficient omega is computed using the item factor loadings and uniqueness from a factor analysis and is a more general form of reliability when compared to coefficient alpha.

Internal consistency or alternatives- Raykov: Raykov derived a method for composite reliability through structural equation modeling and a percentile bootstrap confidence interval. This does not assume items are unidimensional (Raykov and Shrout, 2002; Raykov, 1997, 1998).

Item Remainder Correlations: Item remainder correlations are the Pearson correlations between scores on an individual item and the sum of the scores of the remaining items representing the same dimension. These measures can be used to eliminate items based on the correlations and some elimination criteria (Spector, 1992).

Kuder-Richardson formula 20: The Kuder-Richardson formula 20 (KR-20) is a special case of Chronbach's alpha where the item responses are dichotomous (Cortina, 1993).

Standard Error of Measurement (SEM) measurement: The standard deviation of errors of measurement that is associated with the test scores for a specified group of test takers. Standard Error of Measurement is directly related to a test's reliability. The larger the SEM, the lower the test's reliability. If test reliability = 0, the SEM will equal the standard deviation of the observed test scores. If test reliability = 1.00, the SEM is zero (Glen, 2018).

Sensitivity analysis: Sensitivity analysis is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be divided and allocated to different sources of uncertainty in its inputs (Saltelli, 2002).

Test - Retest: Test-Retest Reliability (sometimes called retest reliability) measures test consistency — the reliability of a test measured over time. In other words, give the same test twice to the same people at different times to see if the scores are the same. For example, test on a Monday, then again the following Monday. The two scores are then correlated. Test-retest reliability coefficients (also called coefficients of stability) vary between 0 and 1, where a correlation of .9(90%) would indicate a very high correlation (good reliability) and a value of 10% a very low one (poor reliability). For measuring reliability for two tests, use the Pearson Correlation Coefficient. One disadvantage: it overestimates the true relationship for small samples (under 15). If you have more than two tests, use Intraclass Correlation. This can also be used for two tests, and has the advantage that it does not overestimate relationships for small samples. However, it is more challenging to calculate, compared to the simplicity of Pearson's (Glen, 2017).

Yule's Y: Yule's Y is used to measure the association between two binary variables, often binary attributes of an object or individual. It is a coefficient of correlation more aptly termed the "coefficient of colligation" (Yule, 1912, p. 593) and is a "normalized variant of the odds ratio, defined in a way that … ranges[s] from -1 to +1" (Tan et al., 2004, p. 297). Yule's Y is similar to Cohen's κ, but provides better estimations of inter-rater reliability when marginal distributions are unequal (Wirtz & Caspar, 2002, as cited in Dunekacke et al., 2016).

References

An, X., & Yung, Y. F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *SAS Institute Inc. SAS364-2014*, *10*(4).

APA Dictionary of Psychology. (n.d.). Retrieved July 09, 2020, from https://dictionary.apa.org/alternate-form

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Deng, L., &amp; Chan, W. (2017, April). Testing the Difference Between Reliability Coefficients Alpha and Omega. Retrieved July 09, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5965544/

Dunekacke, S. Jenße, L. Eilerts, K., & Blömeke, S. (2016). Epistemological beliefs of prospective preschool teachers and their relation to knowledge, perception, and planning abilities in the field of mathematics: A process model. *ZDM Mathematics Education, 48*, 125-137. DOI 10.1007/s11858-015-0711-6

Glen, S. (2017, November 14). Test-Retest Reliability / Repeatability. Retrieved July 09, 2020, from https://www.statisticshowto.com/test-retest-reliability/

Glen, S. (2018, August 01). Standard Error of Measurement (SEm): Definition, Meaning. Retrieved July 09, 2020, from https://www.statisticshowto.com/standard-error-of-measurement/

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Item response theory. (2020, March 13). Retrieved July 09, 2020, from https://en.wikipedia.org/wiki/Item_response_theory

Raykov T. (1997). Estimation of composite reliability for congeneric measures. Applied Psychological Measurement, 21, 173-184. doi:10.1177/01466216970212006 [CrossRef] [Google Scholar]

Raykov T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. Applied Psychological Measurement, 22, 369-374. doi:10.1177/014662169802200406

Raykov T., Shrout P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. Structural Equation Modeling, 9, 195-212. doi:10.1207/s15328007sem0902_3

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk analysis*, *22*(3), 579-590.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. *Handbook of complementary methods in education research*, *309*, 322.

Spector, P. (1992). Conducting the item analysis. *Summated rating scale construction. An Introduction*, 29-46.

Tan, P. N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, *29*(4), 293-313.

What is Trustworthiness in Qualitative Research? (2020, April 09). Retrieved July 09, 2020, from https://www.statisticssolutions.com/what-is-trustworthiness-in-qualitative-research/

Writz, M., & Caspar, F. (2002). Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen: Hogrefe.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *75*(6), 579-652.