Measurement in Educational Research

by Frances Chumney

Without measurement, forward progress would not be possible in educational or other research contexts. Measurement, the process by which a variable is operationalized for the purpose of describing that variable in a quantitative manner (Hills, 1981; Kane, 2001), serves as the link between questions and answers. While the process of measurement bridges the gap between research objectives/questions and the data that is collected, measurement techniques (i.e., analytical approaches) bridge the gap between data and conclusions. Sound research is founded on three underlying concepts which serve as pillars to its structure: theoretical relevance, a sound process of measurement, and appropriate analytic methods. Theoretical relevance is important in that it ties new research to existing research and should guide the development of research questions and hypotheses. However, theoretical relevance is only a part of measurement insofar as it leads to the identification of key variables to be considered; theoretical relevance not be addressed in this paper. The importance of the measurement process and analytic methods cannot be emphasized strongly enough. Without a sound process, there is no confidence in the data it yields (Krebs, 1987); without sound analyses, there is no confidence in the results, interpretations, or conclusions. The purposes of this paper are to present an overview of what the measurement process should look like from the perspective of measurement theory, highlight the consequences of practices that are not consistent with this approach, discuss traditional approaches to analytic methods, and outline the changes made possible by modern techniques.

THE MEASUREMENT PROCESS

Measurement is the driving force behind the design and implementation of data collection. For example, suppose a researcher wishes to investigate sex differences in reading

ability at the end of fifth grade. Before collecting data to analyze and draw conclusions, the researcher must decide how to operationalize "reading ability," and the best methods for assigning numeric values to the reading ability and sex variables. This is the form that the process of measurement takes in practice, and is essential to the development of sound research. The process of measurement is also embedded within school-based settings (Delandshere & Petrosky, 1998). Imagine a second grade teacher who has guided students through a science unit on the life cycle of monarch butterflies and must turn in grades for report cards at the end of the quarter. To evaluate what students have learned, the teacher selects the end-of-unit exam that came with the curriculum s/he used for the science unit. The teacher may not be thinking about the process of measurement, but they have made the decision to operationalize student learning about the material according to the operational definition utilized by the publisher in developing the unit test items. The purpose of these examples is not to provide an encompassing account of the way measurement is used in research or school-based settings, but rather to illustrate the consistent role measurement plays in data collection across settings. Even though the purpose of measurement in these two examples seems different, both the researcher and the teacher have the same overarching goal of understanding a specific phenomenon, and their processes are driven by the theory of measurement.

ASSUMPTIONS OF MEASUREMENT THEORY

Ultimately, the purposes of measurement are to clarify the construct(s) of interest to be measured, identify how it will be measured, and determine how those measurements will be quantified. The process of measurement – as based on measurement theory – provides a framework for the design of data collection efforts by imposing assumptions of operationalization, reliability, validity, and utility. Operationalization, reliability, validity, and utility are discussed here as separate entities. It is important to note, however, that the four

concepts do not impact measurement independently. In fact, operationalization is inherent to the concepts of reliability, validity, and utility which, taken together, form the foundation on which analytic evaluations are conducted, results are interpreted, and conclusions are drawn. Furthermore, these four assumptions are key factors in the process of scale/instrument selection. Ideally, any instrument used in either a research or classroom setting should operationalize the variable(s) of interest in the same way, have a history of reliability with similar samples, have been identified as having strong properties related to validity, and be appropriate for use (utility) with the target population (Krebs, 1987).

One of the biggest challenges associated with the assumptions of measurement theory is that some researchers do not understand what they are or why they are important. In many cases, it is impossible for even a knowledgeable researcher to find an instrument that is an ideal match for their proposed study. Given the options of developing a new instrument that would fit perfectly or trying to force an existing instrument into the study, most researchers opt for the latter. When potentially-relevant instruments are available, it is rare for a single instrument to be a good match for a research project with regard to operationalization, reliability, validity, and utility. Therefore, evaluation of an instrument for use is typically a matter of finding an instrument that is stronger is some areas and weaker in others, but has the closest overall fit to the specifics of the research endeavor.

Operationalization

Operationalization is the process by which the domain area of interest in defined for the purpose of specifying in a consistent way what it is that will be measured (Krebs, 1987). Using one of the examples introduced above, suppose a researcher wishes to investigate sex differences in reading ability at the end of fifth grade. The first step in the measurement process necessary to address this research agenda is to define what is meant by “sex” and “reading ability.” For the

purposes of this illustration, let's assume that the researcher defines "sex" as the biological and physiological characteristics of the child, and "reading ability" as the ability to read aloud quickly with few errors. Without providing these operational definitions for the variables of interest, the intentions of the research are not as clear. Such definitions enable the investigator to communicate specific details about the constructs they wish to study.

Operationalization in Instrument Selection. The way variables of interest are operationalized play a key role in the selection and evaluation of scales/instruments. Because the operational definition targets the specific aspects of a domain that a researcher intends to measure, it narrows the pool of potential instruments to those intended to measure the same key elements as are included in the operational definition. Thus, as a better, more focused (i.e., more specific, detailed, and tailored for objective evaluation) definition is developed by the research, the pool of suitable instruments shrinks. It is logical – even without research training – that an instrument designed to measure a specific trait will do a better job of measuring that trait than would an instrument designed to measure something else. This logic holds in the context of measure selection, as the utilization of an instrument designed to measure something different from the researcher's targeted definition of the construct will result in a mismatch between the research objective and the data. In this case, the researcher will be unable to draw sound inferences based on any analytical techniques. The importance of a match of operationalization is true in applied educational settings as well; a teacher who wishes to measure reading fluency and defines it as the accuracy with which a child reads a passage aloud with not be able to draw any related conclusions if she has the child complete a comprehension test or only considers the number of words the child was able to read aloud in a set amount of time.

In cases where an instrument is intended to measure a construct different from that of the research project, or when the construct has a different operational definition, it is not appropriate to use the measure at all. If an instrument is not measuring the construct of interest, it is an unnecessary burden on participants and will not produce good data for the researcher. In such cases, the researcher *should* work to find a different instrument or take the time to properly develop and test a new instrument. Unfortunately, researchers often opt to develop, test, and use a new instrument in a single study (sometimes even with a single sample), or modify and use existing scales/instruments without testing the new instrument. For example, if a researcher is interested in evaluating the quality of parent-teacher relationships from both the teacher and parent perspectives, but can only identify an instrument designed to measure the concept from the teacher's point of view, the researcher may decide to change all references to the student/parent to references about the child/teacher and administer it to parents. Without the opportunity to test the performance of this measure with a new sample, the researcher who uses this revised measure is implying that the construct of parent-teacher relationships does not differ between parents and teachers.

Reliability

Reliability is roughly defined as the extent to which an instrument produces similar results in similar conditions, or, the consistency of observations/scores that are derived from some scale or instrument (Thorndike, 1982). In the context of classical test theory, reliability can take different forms, including test-retest, split-half, internal consistency, inter-rater, and parallel/alternate forms (Grimm & Yarnold, 2000), each of which can serve a different purpose, and all of which are based on the idea that an individual's score on an instrument (x) is the sum of their true score (t) and the error (e) associated with the measurement of their true score ($x = t + e$; Grimm & Yarnold, 2000).

Reliability in Instrument Selection. Whenever the concept of instrument reliability is considered, it is important to remember that reliability is not a characteristic of a scale/test (Thorndike, 1982). Evaluating the past performance of an instrument, then, means considering all the reliability evidence available. When reviewing literature that has used an instrument, it is necessary to consider the context of its development and not just how the target variables were operationalized or whether alpha (internal consistency) was found to be high. Operationalization is important, but it is only the first step. It is not possible to evaluate the reliability of an instrument for a particular study until data collection is complete. Therefore, one should pay attention to how the instrument has performed in the past as well as the details of the study design/procedure and the sample with which the instrument was used, regardless of whether the instrument is being selected for a research study or application within an educational setting.

Because reliability is a characteristic of the sample and not the measure, the best a researcher can hope for is that a measure has a history of use with samples that are consistent with the researcher's target population. In most situations, this is not a realistic expectation, and the researcher must settle for an instrument used with samples that may not share many (or any) similar characteristics as the target population. Regardless of whether the samples described in past research approximate the population of interest, it is important for researchers to assess reliability within their own samples. Ideally, in cases where a history of use with a similar target population does not exist, instruments should be tested with small samples of the target population prior to use in a research study. Because of the resource demands such practice impose on the research progress, this step is often ignored. The important consequence of moving too quickly through this type of instrument evaluation is that the data collected using an

instrument is more likely to have lower reliability than instruments which have been properly vetted for the context and population of the study.

Validity

Validity is the extent to which a measure actually measures what it is intended to measure (Grimm & Yarnold, 2000). In the context of classical test theory, validity can take multiple forms, including construct, face, criterion, convergent, discriminant, and ecological (Grimm & Yarnold, 2000). From a theoretical perspective, all forms of validity are important and the "best" instrument to use for a study is the one characterized by high levels of all types of validity given the context in which it will be applied. In practice, however, the type of validity that is important depends on the research question(s) being asked. Validity is a characteristic of the relationship between the construct(s) of interest, the target population, the context of the resulting data collection effort, and the intended use(s) of subject scores on the instrument (Messick, 1989).

Reliability and validity are related in that they serve as means of evaluating the usefulness of a measure on two different and important dimensions. Both reliability and validity are limited by the possibility of change, regardless of whether that change is internal to the participants or controlled by the experimenter in an alteration of testing conditions. Often, one is sacrificed at the expense of the other. For instance, when a measure is developed, inclusion of the items with the highest inter-correlations often increases reliability; whether or not these items are measuring strictly the same construct is another issue. It is not safe to assume that highly correlated items are measuring the same construct because they could be measuring related ones instead; including all correlated items might actually decrease the validity of the measure.

Validity in Instrument Selection. Like operationalization and reliability, validity is important in the context of instrument evaluation prior to use. Regardless of the operational definition provided for the outcome measured by the instrument, it is important that the

instrument be designed to actually measure that outcome. If we think again about the fifth grade reading ability test, consider a situation in which "reading ability" is operationalized as reading fluency. If the researcher(s) responsible for the present study pose research questions or hypotheses that frame reading fluency and the accuracy at which a child reads aloud ($accuracy = [(words\ correct)/(total\ words)]$), then an instrument which measures only the number of words correctly read aloud is not a good choice with regard to content validity because the researcher is interested in both words correct and the total number of words. Conceptual matches between a proposed study and a potential instrument should be evaluated for other forms of validity in a similar fashion.

Because there are different forms of validity, there is not a single index or other evaluation criterion for determining whether an existing measure is appropriate for a proposed study. It is important for researchers to understand the different types of validity, identify those that are of particular importance to the proposed study from a practical and substantive perspective, and place special emphasis on those characteristics of potential instruments. For example, if a researcher operationalizes their construct of interest as being the satisfaction of teachers after completing a professional development training experience and they want to identify teachers along the continuum of satisfaction, then the ideal instrument is one with validity evidence supporting the use of the instrument to identify extremely unsatisfied teachers as well as extremely satisfied teachers, and not just a dichotomous grouping of satisfied *vs.* unsatisfied persons. In cases where an instrument is described as being a measure of one construct but there is no evidence in support of it relating positively to things it should be positively correlated with, the validity of the instrument should be questioned. If the researcher is unable to thoroughly explain the discrepancy, the instrument should not be used.

The gathering and evaluation of validity evidence is a time-consuming and sometimes costly endeavor. For these reasons, it is often overlooked by researchers during the initial stages of development. During this time, researchers often focus on reliability and put aside the issues of validity with the justification that it will be addressed in future research. The problem, then, is that individuals who are taking on the responsibility of measure selection do not always know to pay attention to whether or not validity evidence for an instrument exists.

Utility

Utility is the extent to which an instrument is a feasible means for collecting data, with reference to administration costs and respondent burden (Hills, 1981). It might be argued that utility is not as necessary as satisfying the assumptions of reliability and validity, but if an instrument has low utility, the data collected using it are subject to higher-than-usual levels of item- or instrument-level nonresponse, respondent acquiescence, and other errors that can occur during the data collection process (Biemer & Lyberg, 2003; Groves, 1989).

Utility in Instrument Selection. In the context of instrument selection, one could argue that operationalization, reliability and validity are important because they impact the inferences that can be made and the conclusions drawn from the findings. Utility is different, in that it is likely to have much more impact on the research design and implementation procedures. This is not to say that utility is not important with regard to findings, only to point out that a mismatch between the utility of an instrument and a given study is likely to impact findings most in the form of missing data.

To evaluate the utility of an instrument, one is essentially deciding how usable it is, given the target population and objectives of a research scenario. Some important factors to consider during this process are the length, the instructions, how scores are derived, whether alternative forms are available, and any identified potential problems. The length of an instrument includes

both the number of items and the amount of time it takes to complete. One must decide whether the questions are too many (in combination with other selected instruments) such that they will overly burden respondents, is there time built into the study to allow participants to respond to each item, and does the target population have the capacity to attend to all items? Considering the instructions are an important aspect of instrument evaluation, as an instrument is unlikely to perform well if the directions are unclear or confusing, written at a level higher than the ability of the participants, or if they imply that the instrument will be used in a manner differ from the way it is presented (e.g., the instructions imply that the respondent should circle their response when the current study utilizes an interview format). The types of scores that can be derived from an instrument and the tools required to do so is an important consideration. In the case of some instruments, the researcher calculates the sum or mean score; for other instruments, the researcher must score each item as correct/incorrect and then calculate the proportion of items answered correctly; and still other instruments produce norm-based scores which require special software or the ability to correctly utilize scoring tables. In the context of a repeated-measures design, a researcher might want to consider whether alternate forms of an instrument are available, as alternate forms are designed to measure the same construct with the same level of precision, but serve to reduce the effects associated with multiple exposures to a test. Finally, the researcher should take note of any problems others have had with the instrument. For example, an instrument which includes items noted to make respondents feel uncomfortable might not be the best choice with all target populations. In the context of applied educational settings, utility is extremely important because teachers typically have little time or other resources at their disposal with which to administer assessments, particularly ones that might require one-on-one time for administration.

Of the four assumptions of measurement theory, utility is likely the most commonly met assumption in research. Unlike the importance of operationalization (which may be misunderstood by some researchers), and reliability and validity (concepts which may be entirely overlooked or ignored because a researcher identifies himself as "not good at stats), the utility of an instrument is typically considered because it includes basic characteristics about an instrument and how it is administered without requiring an understanding of statistics or statistics-related concepts such as reliability and validity.

THE ANALYTIC PROCESS

Just as a sound measurement process is necessary to produce good, relevant data to address researcher hypotheses, appropriate methods are needed to analyze those data to ultimately answer the research questions posed. Given the broad range of research questions that one might ask, analytic methods employ a broad range of statistical methods. Regardless of how well the data might be suited for a complex analytic procedure, it is always preferable to use a simpler approach, assuming such a technique allows for appropriate consideration of the context and properties of the data and research questions. For example, if the research question focuses on whether the boy to girl ratio is consistent for third and fourth graders in a single school, perhaps only a chi-square test is necessary or appropriate. In some instances, however, a more sophisticated approach to data analysis is warranted. Historically, analytic techniques in education have been driven by classical test theory (CTT; Gulliksen, 1950; Spearman, 1904). Over recent decades, however, the continued development of computer technology usable for estimation of models encouraged an evolution of sorts within the methodology field, which in turn prompted and encouraged the development of more modern approaches to measurement and analytic techniques.

Classical Test Theory

CTT is an analytic approach which focuses on the test score as the primary unit of analysis and postulates that test scores are a linear combination of an observed score and some amount of error. The goal of CTT is to determine how much of the observed test score variance is due to true score variance versus error variance, and then to reduce error variance as much as possible (Fan, 1998). CTT is concerned with the reliability of instruments, where reliability is defined as consistency and is a measure of the relationship between the proportion of true score variance (factor variance) relative to total variance (i.e., the sum of true score variance and error variance). In the CTT framework, test scores are meaningful only relative to other persons within the same sample. Thus, an individual's test scores cannot be interpreted without reference to sample norms such as criterion-reference norms (i.e., scores have absolute meaning relative to items and tests, and the metric is interpreted without reference to its distribution) and norm-referenced scores (i.e., scores are only meaning relative to the sample, and the metric is based on the distribution of sample scores). A primary advantage of CTT is that its theoretical assumptions are rather weak, which makes the approach easy to adapt across situations (Hambleton & Jones, 1993). Despite this flexibility, a lack of theoretical foundation is also a primary disadvantage of CTT, as it is not appropriate for testing a priori, theory-driven models.

In addition to the limitations of sample-specific scores and a lack of adequacy to test theory-based models, there are three major disadvantages of CTT which contributed to the development of alternative analytic techniques (Reid, 2007). The first is that the person score it estimates is dependent on item characteristics, and the item characteristics are dependent on the person characteristics – a situation of circular dependency (Fan, 1998). Circular dependence is problematic because it implies that an individual's score (or, level of trait or ability tested by a particular instrument) is dependent on the particular items used. The second major disadvantage

of CTT is that estimates of scores at either end of the distribution (high or low) are unstable, even when norms are used. The third major disadvantage is that it is not possible to calculate scores without administering an entire instrument. In standardized testing situations, a very large number of items are sometimes required to achieve the level of specificity in scores that is desirable, but large items lead to respondent fatigue, which impacts cognitive functioning and effects test scores.

Trending Away from Classical Test Theory

Over the past few decades, measurement techniques have started to evolve and new methods have emerged as alternatives to classical test theory, including latent trait modeling, item response theory, dynamic and adaptive testing, and multilevel measurement. Such techniques are often referred to as "modern" measurement approaches. Because these new techniques offer creative ways to address the limitations of classical test theory, they have real potential to forward scientific understanding of educational phenomena. The most significant improvement of these methods over CTT is that they are model-based, which means they can be used to test hypotheses about the relationships that exist between variables.

Modern measurement approaches have changed the face of educational research in significant ways over the past few decades. As technology has improved, estimation of theory-based models has become more possible and flexible. As technology has become more commonplace, it has transitioned from a tool for researchers to use during analyses to a mode of data collection. This advancement is of particular importance in the field of educational testing, as more and more testing is now conducted via computers. The power and flexibility of computers makes it possible for tests to be tailored to the individual while they are completing the test, as is the case with computerized adaptive testing. Modern measurement has also been informed by research advances in other fields. For instance, as our understanding of the

importance of context in which behaviors occur has improved, methods have developed (i.e., multilevel models) to allow consideration of factors other than test scores and participant characteristics.

In practice, the development of modern measurement techniques seems to be driven by need. On the one hand, need-driven progress is still progress, and once a need has been met for one research project, the method can be generalized to other projects. On the other hand, need-driven progress often means that solutions are found and applied to data within the context of substantive research before the technique is fully understood. In the realm of methodological research, it is interesting to see new methods applied to "real" data, and often makes us aware of characteristics of the method that we might not have thought of, or might not previously had considered important to investigate. In the realm of substantive research, though, applying methods to data before they are completely understood can lead to erroneous findings and unsubstantiated conclusions which may lead future research down steep, ugly rabbit holes.

Model-based measurement approaches are still novel in the research world, and will continue to be for some time. This sets the scene with publication opportunity for those individuals with the motivation and training to pursue a deeper understanding of these modern techniques. Unfortunately, the newness of these techniques also makes them attractive to substantive researchers who wish to apply them without understanding their assumptions or the implications of using them. It is not unheard of for a researcher to insist on applying a certain model because it sounds really "cool," and because they have not seen anyone else in their field use that type of model, they believe it will make them "cutting edge" and improve their chances of publication and exposure. This is problematic when the data must be "finessed" to fit with the

requirements of the model/technique, or when the model is not appropriate for the research questions.

At this point in history, the measurement tools available to us have taught many lessons. As a field, educational research has a great deal of momentum and there are promising new developments on the horizon. It is important, though, to not get carried away with the excitement of the moment and forget about the foundations on which quality research is built (i.e., quality of measurement process). As more people receive advanced training in analytic methods, they shoulder the burden of advancing the methods as well as applying the technique. Non-methodologists need to better understand the potential consequences of using techniques for substantive purposes before they are understood methodologically.

CONCLUSION

In order for scientific discovery and process to develop and move forward in any field (including education), analytic techniques must be applied to data collected as the result of a thorough development of a measurement process prior to data collection. In the current educational research climate, sound measurement process practices are frequently ignored, as researchers often settle for "close-enough" instruments or attempt to develop their own in a single study. In school-based settings, there is little evidence to support any claim that the average school personnel understand the process of measurement well enough to produce arguably good data. Such a lack in training is illustrated by the current popularity of training programs focused on paradigms for making data-based instructional decisions (e.g., response to intervention programs). Fortunately, such programs are becoming more widely available and online delivery systems are making them more accessible. Thus, teachers and administrators are beginning to understand the role and importance of data-driven decisions and their potential advantages for students.

The development and incorporation of modern measurement techniques is exciting and promising. Unfortunately, these practices represent only the way in which measurement bridges the gap between data and outcomes. Ultimately, the usefulness of analytic techniques is dependent on the quality of data that is collected and its potential to address the research questions of interest. With consideration of current practices, the future of education research looks bleak, but there is reason for hope. Just as schools are coming to understand the importance of measurement, more and more graduate training programs are encouraging students to complete coursework focused on research design and measurement. This is an important evolutionary process for the research world, as individuals who receive such training are more likely to understand the importance of including methodologists (who understand the intricacies of the measurement process and analytical techniques) in the development phases of a research project instead of consulting them only after the data has been collected and expecting magic to happen.

References

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality.* Hoboken, NJ: John Wiley & Sons.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*(2), 14-24.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.

Grimm, L. G., & Yarnold, P. R. (2000). *Reading and understanding more multivariate statistics.* Washington, DC: American Psychological Association.

Groves, R. M. (1989). *Survey errors and survey costs.* Hoboken, NJ: John Wiley & Sons.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R. K., & Jones, R. w. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practices, 12,* 38-47.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd ed.). Columbus, OH: Charles E. Merrill Publishing Company.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38¸*319-342.

Krebs, D. A. (1987). Measurement theory. *Physical Therapy, 67,* 1834-1839.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed.). New York: Macmillan.

Reid, C. A., Kolakowsky-Hayner, S. A., Lewis, A. N., & Armstrong, A. J. (2007). Modern psychometric methodology: Applications of item response theory. *Rehabilitation Counseling Bulletin, 50*(3), 177-188.

Spearman, C. E. (1904). ‘General intelligence’ objectively determined and measured. *American Journal of Psychology, 5*, 201-293.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston: Houghton Mifflin.