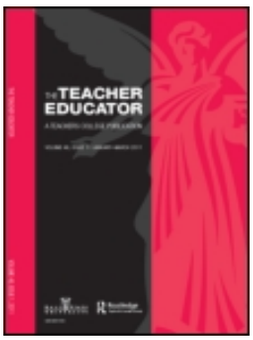

The Teacher Educator

ISSN: 0887-8730 (Print) 1938-8101 (Online) Journal homepage: http://www.tandfonline.com/loi/utte20

A Teacher's Guide to Assessment Concepts and Statistics

Carole Newman & Isadore Newman

To cite this article: Carole Newman & Isadore Newman (2013) A Teacher's Guide to Assessment Concepts and Statistics, The Teacher Educator, 48:2, 87-95, DOI: 10.1080/08878730.2013.771495

To link to this article: https://doi.org/10.1080/08878730.2013.771495

Published online: 17 Apr 2013.

The Teacher Educator, 48:87–95, 2013

ISSN: 0887-8730 print/1938-8101 online
DOI: 10.1080/08878730.2013.771495

Routledge
Taylor & Francis Group

FEATURED ARTICLE

A TEACHER'S GUIDE TO ASSESSMENT CONCEPTS AND STATISTICS

CAROLE NEWMAN

College of Education, Florida International University, and
College of Education, The University of Akron

ISADORE NEWMAN

College of Education and Department of Human Development and Molecular Genetics, College of Medicine, Florida International University, and College of Education, The University of Akron

The concept of teacher accountability assumes teachers will use data-driven decision making to plan and deliver appropriate and effective instruction to their students. In order to do so, teachers must be able to accurately interpret the data that is given to them, and that requires the knowledge of some basic concepts of assessment and statistics. This article will provide the classroom teacher with the basic vocabulary of assessment and clear descriptions of these concepts to facilitate their use of assessment data to develop effective instruction.

Do teacher practitioners *really* need to know anything about statistics and test interpretation? The simple answer is *YES*. In today's world there is no getting away from the importance of assessment in judging a teacher's skill, a school's success, student proficiency and for making funding, salary and even hiring and firing decisions. An incessant call for ''accountability'' requires that we quantify our student outcomes in a number of areas from mastery of content to graduation rate, by using a wide range of assessments—some good and some not so good. But, to use and interpret assessments appropriately requires some basic understanding of statistics and research design. These concepts are not independent; they are necessarily interdependent for practitioners to adequately understand how to use the information generated by assessment in an appropriate formative, diagnostic, and applied way.

You might be thinking, ''How can I possibly go wrong in interpreting the data? If I read it, I will know what it means.'' The problem is that interpreting what we read is not always so clear-cut. Although the appropriate use of good information can help us to improve student learning, misinterpretations can lead us in the wrong direction. To illustrate this point we will begin by telling you two true short stories that may seem find hard to believe, but demonstrate how research was misunderstood by two teachers who were trying to help a student.

Address correspondence to Carole Newman, College of Education, Florida International University, 11200 SW 8th St., Miami, FL 33199, USA. E-mail: cnewman@uakron.edu

In our first story a well-meaning third-grade teacher explained to a parent why she chose to wait in teaching her son to read. The teacher shared a research article with the parent that said there was a *significant correlation* between height and reading readiness. Because the student was short, the teacher believed he was not ready for more intensive reading instruction. She told the mother that she did not want to frustrate the child, and the teacher suggested waiting until he got taller and physically matured, thus more ready to read. This may seem to be obviously incorrect, but it is indicative of how a caring teacher misinterpreted the meaning of *statistical significance to imply causation* (correlation does not mean causation), and thereby hampered this child's learning and academic success.

Our second true story is more related to psychometrics—reliability and validity. In the fourth grade, this same precocious child asked his teacher, ''How do you know that Columbus discovered America?'' The teacher responded by saying. ''It's in our textbook.'' The little boy responded, ''I know it's in the book, but how do you know that's right?'' (This is a *validity* question.) Unfortunately, the teacher thought the child was trying to be a wiseguy, and she removed him from the class. (Many years later this teacher was taking a graduate-level research course, and her former student, who she had removed from class, was the graduate assistant responsible for teaching the module on reliability and validity. However, he was too shy to reintroduce himself and recall the incident to his former teacher.) Again, the teacher in this story assumed that because it was written in a material approved by the district, it must be right. Unfortunately, that is not always the case, and instead of encouraging critical thinking in her students, her response tended to hamper it.

So What do Teachers Really Need to Know?

The call for data-driven decision making and teacher accountability has made it more important than ever that teachers accurately interpret the data they are provided with so they can use them formatively, summatively, and diagnostically to make appropriate academic choices for students. Teachers must be able to look at scores, growth patterns, test reliability, and validity to determine how much weight to give to the data, and to be able to clearly explain results to parents. This requires an understanding of some minimal concepts in statistics, psychometrics, and research design. The following is a brief introduction to some of these basic concepts selected to help educators better understand the literature so they can ask informed questions about what they are reading, and so they can more accurately interpret data, make appropriate inferences, and take meaningful action based upon the data.

These very basic concepts are (a) What statistical vocabulary do I need to know?; (b) What does statistical significance mean?; (c) What is reliability and its relationship to validity?, and (d) What are the differences between criterion- and norm-referenced tests?

Without these minimal understandings teachers cannot accurately interpret test scores and they will be less likely to use the information to make appropriate, data-driven curricular decisions. Understanding of key concepts will allow teachers to form and voice their opinions based on thoughtful and scholarly arguments that can be evaluated objectively.

Interpreting the Language of Statistics

To have a conceptual (not a computational) understanding of statistical concepts, one needs a basic vocabulary. Therefore, our intent is to introduce some of these terms to im-

prove communication among professionals and to increase the usefulness and appropriate application of data.

Grade Equivalence

This measure is very frequently misunderstood by parents and teachers. For example, if Jimmy, who is in the fourth grade, scores 11.5 on a math test, this does not mean Jimmy is capable of doing the same work as a student who is in the fifth month of the eleventh grade. It does mean that Jimmy's score was similar to what would be expected of a student in the fifth month of the eleventh grade who was taking the test on the fourth-grade material. Jimmy's test probably had to do with addition, subtraction, multiplication and division, whereas the eleventh-grade test would most likely include some concepts of algebra, geometry, and other advanced mathematical concepts. In all likelihood, Jimmy would not be familiar with any of these concepts, and he would not be able to respond correctly.

Age Equivalence

This is very similar to the idea of Grade Equivalence. One of the major differences is that instead of reporting a score in terms of a grade level that is divided into 10 intervals (representing each month of school), age equivalence is divided into 12 intervals, thus reporting the age and month.

One of the problems with both Grade and Age Equivalence measures is that the same tests are not given to students at all grade and age levels, but the test calculations are determined by extending a regression line past the tested populations, and then assuming all students will maintain the same rate of growth, regardless of their grade or age. Thus, it assumes a fourth grader learns at the same rate as an eighth grader, on average. We know this is not necessarily true. Therefore, the further the test score is from the grade or age intended, the more questionable these become.

Standard Scores

Test scores from various tests are often inappropriately compared. These tests frequently have different means and standard deviations (*SD*), and therefore, the raw scores are really not comparable. In order to be able to appropriately compare scores from different tests, one needs to convert the raw scores to some common unit of measure. These common units are referred to as *standard scores*. That way, the scores are calculated to have similar relative positions in their distribution of scores and comparisons can be interpreted correctly (Nitko & Brookhart, 2011; Popham, 2001). For example, if IQ Test A has a mean of 100, and IQ Test B has a mean of 50, then a person who scored 100 on Test A and one who scored 50 on Test B are comparable. Both are at the mean (average) of their testing group. By knowing the mean and standard deviation of each test, the standards score allow us to convert scores to a common unit. A *standard deviation* is conceptually a statistical value used to determine how close data points are to a mean value. If the $SD = 0$, then all scores have the same value as the mean. The larger the *SD*, above or below the mean, the further the average of all scores are from the arithmetic mean. The following are seven of the most commonly used standard scores that are frequently reported:

Z-Scores

This is the most commonly used standard score. A Z-score of 0 will always be reported as the mean (''z'' think zero), and the *SD* will always be 1. Z-scores with a range from +3 to −3 account for approximately 99% of all scores. Any Z-score above 0 is always above the mean and a score below 0 is always below the group mean. In the IQ test example above, both of the scores, 100 and 50, would have a Z-score of 0 because they are both at the mean of their groups.

T-Scores

This is conceptually the same as a Z-score, but instead of having a mean of 0 for the distribution of scores, the T-score has a mean of 50, and instead of having a *SD* of 1, it has a *SD* of 10 (''T'' think 10). So any individual who has a T-score of 50 is at the mean of his or her group. Likewise, any score that is 1 *SD* above the mean in both Z and T scores are at the same relative position in the distribution of scores. The transformation of Zs to Ts was done to eliminate the need to deal with negative numbers that have to be used to describe scores below the mean when reporting Z-scores.

Stanines

Another form of standard scores that are frequently reported are *stanines*. As the term implies, these raw scores are divided into nine values. The mean of a stanine distribution is always 5 and the *SD* is always 2, so someone who is at the average of a stanine distribution has a score of 5. This is comparable to a Z-score of 0 and a T-score of 50. In using this calculation, 20% of the scores in a distribution will be at the stanine of 5. The major disadvantage of using stanines is that it assumes that scores are normally distributed. If this is not the case, interpretation is difficult. An additional disadvantage is that although Z- and T-scores actually represent every score in the distribution, stanine scores do not. Scores are grouped into the nine divisions, so it is likely that a stanine score of 5 will contain a range of scores, some of which will be closer to a stanine of 6 and some will be closer to a stanine of 4, for that distribution. Therefore, we lose information concerning differences among the individual scores in the same stanine.

Percentiles

Percentile scores range from 0 to 100 and they specify the percentage of scores that fall below a particular score. That is, a percentile of 40 indicates that 40% of the tested population scored below, and 60% scored higher on that particular test. The point where 50% of the scores fall above and 50% fall below is referred to as the *median*. This is one type of average that is commonly reported and differs from the *mean*. The mean adds and then averages all scores in the distribution, so it is more sensitive to extreme scores at either end, whereas the median is always at the 50th percentile.

Percentile Rank

When using percentiles to report scores, individual scores are given a Percentile Rank. A Percentile Rank of 1 is awarded to all scores that fall between 0 and 1%. Similarly, all

scores that fall between 37 and 38% would have a Percentile Rank of 38, and all scores between 76 and 77% have a Percentile Rank of 77.

Normal Curve Equivalents (NCE)

The NCE was developed primarily to evaluate federal programs, such as Title I Reading and Math. Like the T-score, it also has a mean of 50, but the standard deviation is approximately 21 units. The lack of a more common or convenient number to work with for the standard deviation, such as 1 or 10, makes some interpretations more difficult. NCE values range from 1–99, so it is easy to detect small units of gain. For example, an NCE score of 1 would be comparable to a 1 percentile and a Stanine of 1. An NCE of 15 would be comparable to a 5 percentile rank and a Stanine of 2. An NCE of 50 is comparable to a 50 percentile rank and a Stanine of 5, and an NCE of 99 is comparable to a 99 percentile rank and a Stanine of 9. There are tables in almost every statistics book that report these percentages and show the conversion relationship between Zs, Ts, percentiles, NCEs, and so on. These conversions are easy to calculate because they are all standard scores and therefore have a measure of central tendency that has 50% of the scores above and below the median, and some common unit of variability (*SD*).

Vertical Scale Scores

These are less commonly used, but can be seen in some places such as on the TAKS Texas state proficiency test. When using Vertical Scale Scores, a student's scale score in one grade can be compared to the same student's scale score in another grade as long as the scores are in the same language and subject area. This allows one to look at that student's relative growth across grade levels (Nitko & Brookhart, 2011).

All of the standard scores allow one to make relative comparisons between tests. Regardless of whether you are using a Z-, T-, Stanine, Percentile, or NCE score, you can compare all of the scores in terms of the percentage of people who are above or below a particular value. They all report the relative placement of a score in a distribution of scores.

Correlations

This is the measure of the degree of a relationship between two or more variables. There are many types of correlations, but the one most frequently used and the one that we discuss here is the *correlation coefficient* (r). It measures the degree of linear (straight line) relationship between two variables and the values of r range from a $+1$ to -1, with a 0 indicating no relationship. The higher the absolute value of r, that is, the closer the calculated correlation is to either $+1$ or -1, regardless of the sign, the greater the relationship. So, a correlation of .6 shows less of a relationship than a correlation of $-.8$. The sign only relates to the direction of the relationship, not the magnitude. Therefore, a negative correlation means that as one variable increases the other decreases (think the amount of gas in your car and amount of miles driven), and a positive correlation indicates that as one variable increases the other increases (think about time spent studying and test scores).

All tests of significance are really measures of relationships that exist in a sample and are then inferred to the population from which the sample was drawn. However, even if a

significant relationship (correlation) that is not due to chance is found, it is inappropriate to assume that one variable *caused* the other. For example, there is a significant correlation between the amount of ice cream consumed and the number of drownings that occur. It would be inappropriate to assume that eating ice cream causes people to drown even though a significant relationship exists between the two. It is more likely that people eat more ice cream during the summer when they are more likely to go swimming, and when more people swim, the incidence of drowning is more likely. Even though both events increase similarly, one does not cause the other.

Going back to our original story about the teacher who delayed reading instruction because her student was too short, it is apparent that the teacher inappropriately assumed a causal relationship existed between height and reading readiness. No statistical procedure allows one to assume causal relationships. Causality can only be assumed when a research design allows the researcher to control for alternative explanations. This generally requires having a control group that does not get the treatment and the random assignment of subjects to treatment groups. This is seldom, if ever, possible when research is conducted in natural settings such as a classroom or school.

What Does Statistical Significance Mean?

When something is reported as being ''statistically significant'' it means that the relationship being reported is not likely to be due to chance. Although this statement is absolutely correct, it is meaningless by itself.

To understand statistical significance, one needs to know that when a sample from a population of interest is analyzed, such as a sample of students from a particular school district, the results are assumed to represent what occurs in the population from which the sample was drawn. That is, the sample would have a proportion of students who mirror the school district's demographics on variables such as males and females, socioeconomic status (SES), racial groups, second language learners, and so on, so that the results could be assumed to represent the entire school population. To do this accurately, the sample has to be unbiased in that it fairly represents the whole population. If it does not, any inferences to the larger population, regardless of the level of significance (confidence that it is not due to chance), would be inappropriate. (Random sampling procedures are assumed to produce unbiased samples.) Frequently, samples are not representative of the populations from which they are drawn, and therefore any inferences made from the sample to the population are likely to be inaccurate.

The second thing that needs to be understood about statistical significance is that it is operationally defined by the alpha level (level of confidence) selected. In education and the social sciences, that level is usually .05, .01, or .001. This means that at an alpha level of .05 one can be 95% confident that a significant relationship found in a sample is not due to chance and probably exists in the population. For an alpha level of .01, one would be 99% confident that the relationship is not due to chance, and for .001, one would be 99.9% confident that a significant relationship found in a sample also exists in the populations from which the sample was drawn.

Although the above interpretation of an alpha level is accurate, this does not mean that the results will be replicated from sample to sample. Therefore, finding statistical significance at an alpha level of .05 (95% confidence) does not mean that the results would be replicated 95% of the time. In actuality, it would only be replicated 50% of the

time. Explaining why this is the case goes beyond the scope of this article, but it is important to keep in mind that significance does not mean replicability (Newman, Newman, Brown, & McNeely, 2006, p. 143).

As previously stated, statistical significance does not mean that a relationship is causal. If there is a statistically significant relationship between height and intelligence (or readiness to read), it does not mean that height is causing the difference in intelligence. Similarly, if there is a statistically significant relationship between SES and grade point averages (GPA), we cannot say that differences in GPA are due to differences in SES. It is important to remember that one **CANNOT** assume that correlations (relationships) mean causation. Only controlled research design can determine if causal relationships exist (Campbell & Stanley, 1967; Newman et al., 2006).

What is Reliability and its Relationship to Validity?

When using assessment instruments it is very important that teachers understand the concepts of reliability, validity, and the relationship between the two. Reliability is an estimate of the consistency of an assessment. A *reliable* instrument should have similar results every time it is used. If you are measuring a student's height every day for a week, you should reliably get pretty much the same height each day. Technically, ***reliability*** is an estimate of measurement error. If there is no measurement error, you would get the same answer every time you measure the same concept, and you would have a reliability estimate of 1 ($r = 1.0$). For instance, if we incorrectly assume height is related to intelligence, we could conceivably try to estimate a student's intelligence by measuring the student with a yardstick. Every time the student is measured we would get very similar results, and therefore our measure is highly reliable even though it is not meaningful. In the stories presented at the beginning of this article, both examples would have high reliability estimates because little children are likely to be shorter and have lower reading readiness and reading level scores, and many social studies texts would state as fact that Columbus discovered America. However, that does not mean one can infer that height causes reading readiness or that Columbus actually did discover America. Consequently, having a high reliability estimate is important but not sufficient for determining if results are meaningful. We also need estimates of validity.

Validity is an estimate of how well a test measures what it is supposed to measure. There are at least eight types of validity, but we will only discuss a few. *Face validity* (the weakest form) is estimated by how well a test looks like it is measuring what it is supposed to be measuring. *Expert judge validity*, is when a test measures what it purports to measure based on the judgment of experts. Obviously, this is only as good as the expertise of the judges and their credentials should be documented. Another type of validity is *concurrent*, sometimes called *criterion*, *known group*, or *discriminant validity*. This is based on the correlation between a criterion (maybe a similar test or observed behavior) and the test of interest. The type of validity that may be the most useful for educators is *predictive validity*. This estimates how well the test score obtained by an individual predicts the outcome of interest, such as GPA, graduation, and so on. The final type of validity that we will define is *construct validity*. This type of validity is generally a combination of all of other types of validity, and is generally based upon how well all of the validity estimates support the underlying construct (theory) that the test is supposed to measure. This type of validity frequently uses a statistical procedure called factor analysis to estimate the underlying construct of a test.

When evaluating the usefulness or appropriateness of a test, one needs to look at the reported reliability and validity estimates of the instrument. Unfortunately, test developers often provide huge amounts of data about test reliability estimates obtained (how often it gets the same results) from sophisticated statistical techniques, such as Rasch Modeling, item response theory (IRT) for item development, and Cronbach's alpha, but offer insufficient information about the validity (does it really measure what we want it to measure?). Sometimes, the way they present the Rasch Modeling and IRT data makes it sound like they are presenting validity estimates, but they are not. They are really estimates of measurement error (reliability) and not validity. Although knowing if a test has good reliability estimates is important, it is not as important, nor is it a substitute, for having good validity estimates.

What are the Differences Between Criterion- and Norm-Referenced Tests?

Tests can be divided into two major categories, criterion-referenced and norm-referenced. *Criterion-referenced tests* are developed to determine how well a student can achieve the desired objectives of instruction, standards, or benchmarks. Therefore, on a good criterion-referenced test it is possible that all students can achieve a score of 100% or all can fail, depending on how well they achieved the objectives being tested. *Norm-referenced tests* have a different purpose. They are designed to differentiate between students of varying abilities. The reliability estimates of these tests are frequently used as important indicators of the effectiveness of the test, but the more important validity estimates that indicate how well the test actually tests what it purports to test are often not reported.

State proficiency tests, such as the Florida FCAT, the Texas TAKS, and the Indiana I-STEP+ are constructed to serve as a combination of both criterion- and norm-referenced tests. These tests are developed to both differentiate between students and to assess the minimum level of objective attainment necessary to be considered proficient in a particular area. Because they have multiple purposes that are based on both norm- and criterion-referenced concepts, there is an inherent problem in interpreting the scores they produce and in using them as a basis for an accountability system.

This ''serious design flaw'' was reported by Stroup, Pham, and Alexander (2007) from the University of Texas at Austin in a project report on an innovative mathematics program designed to improve math skills for seventh and eighth graders in an economically disadvantaged middle school. Stroup and colleagues reported on the disparity between the dramatic growth students were demonstrating on district designed benchmark tests of the mathematical concepts taught in the classroom and the marginal increase in scores on the end of year Texas standardized TAKS tests. Dr. Stroup and colleagues found that standardized test scores from the previous year were better predictors of how students scored on the following year's standardized tests than were the benchmark tests they had taken during the year. Stroup et al. said these findings threatened the foundation of high stakes test-based accountability and states the tests were ''virtually useless at measuring the effects of classroom instruction.'' Without going into too much detail, Stroup et al.'s research found that the test developers created test items using IRT, which produces items designed more to rank students than to measure what they have learned. These test items were more like those of IQ tests than content tests. Stroup et al. claimed that in Texas (and probably elsewhere) the district benchmark tests assess benchmark achievement, but the standardized tests are developed by the testing company with dual purposes in mind.

Stroup et al.'s findings were reported widely in the popular press like the *New York Times* and *Huffington Post,* but are being challenged by the testing company as misleading.

Conclusion

Even though we as educators have little, if any, input into the construction of the tests that are being used to judge our competence, we can question these tests and challenge test construction companies, school districts, states, and the federal government to take a closer look at the philosophy used to create items. We need them to be sure that the *actual intention* of the test is to identify student growth in content if they are truly to be a measure of accountability. They have to be designed in a manner that will provide accurate information for a sound accountability system.

Too often today's teachers are presented with test scores and are told the results are a measure of their success in the classroom. Although the results certainly are representative of how well students performed on the test, they may not provide a true picture of what has been accomplished academically in the classroom. We believe that it is important for teachers to understand basic concepts of statistics, research design, and assessment so they can more effectively evaluate the information that comes to them as test results. Having this understanding may help them better interpret the information, make appropriate choices in planning instruction, and be better able to explain to parents what their student's scores really mean.

References

Campbell, D. T., & Stanley, J. C. (1967). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Newman, I., Newman, C., Brown, R., & McNeely, S. (2006). *Conceptual statistics for beginners* (3rd ed.). Lanham, MD: University Press of America.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Boston, MA: Pearson.

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th ed.). Boston, MA: Pearson.

Stroup, W. M., Pham, V. H., & Alexander, C. (2007). *Richardson MathForward Project second year final report: Math TAKS results.* Austin, TX: The University of Texas at Austin.