Review Of Research
Vol.2,Issue.1 ,Oct; 12pp.1-4
ISSN:-2249-894X

Available online at www.reviewofresearch.net

ORIGINAL ARTICLE

The Study Of Psychometric Properties Of Final Tests Questions In Sciences Teachers

Hasan Sabetdivshali , maryam Najafi Moghadamnejhad, fatemeh Khoshnava Foomani And Rahmatollah Kharazmi Rahimabadi

Tonekabon Branch, Islamic Azad University ,Tonekabon, Iran
Department of Educational Sciences - Iranshahr Branch- Islamic Azad University –- Iranshahr _ Sistan&Baluchestan-Iran
School of Nursing and Midwifery, Tehran University- IRAN
CCLD Dehkhoda "Correction Center For Learning Disorders" –Guilan-Iran

Abstract:

Background and aim. *Judgments about the level of academic achievement and efficiency of education are to rely on test results and examination of the specific objectives of education and makes clear. The purpose of this study was to obtain accurate information from basic science to teachers Anzali city Suitable for writing test questions, test principles and preparation of appropriate tests to assess the validity of traits measured.*
Methods and material. *The secretary of the 4 groups 24 textbooks and research papers totaling 480 number plates were examined. In this study, three questions about the quality tests of the fundamental principles of testing, Consistent with the objectives of education and was considered valid. Information obtained by a computer and analyzed using spss software was used.*
Results. *The analysis results show that the most capable teachers in the principles above set of tests, High percentage of test questions are two levels of knowledge and understanding of the six levels of cognitive domain can be measured And a lower percentage of two-level analysis was devoted to Very small percentage of the surface composition and the evaluation questions were included.*
Credit average and above average teachers' groups that represent the Is that enough tests have been discredited. Correlation tests were carried out most of the questions teachers has been positive with the total score.
Implications for practice. *The results of the present study can be used as a guideline for teachers and schools principals to develop valid and acceptable tests.*

KEY WORDS:

psychometric properties, final tests, sciences teachers.

INTRODUCTION

One of the most dreaded parts of school life has to be the class test. All the way through school, children have to take tests in one form or another. From first grade onwards, there will be some point at which children have to go over everything they have learned. School tests take various forms - oral question and answer sessions, multiple choice questions, essay questions, practical demonstrations, and written short questions. These methods vary depending on the subject studied and the age of the students.

Testing is extremely important however, because without it no teacher can really know how much

Please cite this Article as :Hasan Sabetdivshali , maryam Najafi Moghadamnejhad,fatemeh Khoshnava Foomani And Rahmatollah Kharazmi Rahimabadi : The Study Of Psychometric Properties Of Final Tests Questions In Sciences Teachers : Review Of Research (Oct; 2012)

the students have learned. This is necessary, not only in terms of the students but also for the teacher so that he or she can know where the class is holding when preparing the material for the next lessons. It can also show who the weaker and stronger students are - who needs extra help and who needs more of a challenge (Bayes Rules, 2006).

For the student, testing is a good idea because this is an ideal opportunity to pause, take stock of the material studied over the recent period, and process it so that it is properly understood. In addition, there is always the satisfaction of passing the test and really feeling that you know something. And if you don't pass, there is the challenge of having to relearn the material and make sure that you do know it next time (Impara, 1996).

Standardized testing has been called the greatest single social contribution of modern psychology, and it may be the most useful evaluation method available for human resource intensive endeavors. For most of their history, however, standardized tests have been developed and administered on a large scale and large, typically politically-sensitive organizations have controlled their use (Sireci, 2005).

With powerful forces opposed to the use (or to the proper use) of a beneficial technology that is typically provided by large, politically-sensitive organizations, perhaps it is time to consider alternative methods of providing that beneficial technology (Lieberman, M, 2007). One such alternative method is the topic of today's session.

Standardized tests are not perfect evaluation tools. Used validly and reliably, however, standardized tests provide decision-makers useful information that no other evaluation method can provide (Phelps, 2005). Many research studies on educational testing dating back to the early part of the 19 century have compared different teachers' evaluations of identical student work or compared the consistency of teachers' marks to those of standardized test results over time. Not surprisingly, researchers found wide variance from teacher to teacher in grading identical student work or over time with the same teacher.

VALIDITY OF TESTS

Validity is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted. Validity isn't determined by a single statistic, but by a body of research that demonstrates the relationship between the test and the behavior it is intended to measure.

Test validity concerns the test and assessment procedures used in psychological and educational testing, and the extent to which these measure what they purport to measure. “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.” (American Educational Research Association, 1995) Although classical models divided the concept into various "validities" (such as content validity, criterion validity, and construct validity) (Guion, R. M., 1980), the currently dominant view is that validity is a single unitary construct. (Messick, S., 1995).

Validity is generally considered the most important issue in psychological and educational testing (Popham, W. J., 2008) because it concerns the meaning placed on test results. (Messick, S., 1995)Though many textbooks present validity as a static construct (Brookhart, S. M., 2004), various models of validity have evolved since the first published recommendations for constructing psychological and education tests. These models can be categorized into two primary groups: classical models, which include several types of validity, and modern models, which present validity as a single construct. The modern models reorganize classical "validities" into either "aspects" of validity (Messick, S., 1995) or types of validity-supporting evidence

Although psychologists and educators were aware of several facets of validity before World War II, their methods for establishing validity were commonly restricted to correlations of test scores with some known criterion. Under the direction of Lee Cranach, the 1954 Technical Recommendations for Psychological Tests and Diagnostic Techniques attempted to clarify and broaden the scope of validity by dividing it into four parts: (a) concurrent validity, (b) predictive validity, (c) content validity, and (d) construct validity. Cronbach and Meehl's subsequent publication grouped predictive and concurrent validity into a "criterion-orientation", which eventually became criterion validity.

Over the next four decades, many theorists, including Cronbach himself, voiced their dissatisfaction with this three-in-one model of validity. Their arguments culminated in Samuel Messick's 1995 article that described validity as a single construct composed of six "aspects"[3]. In his view, various inferences made from test scores may require different types of evidence, but not different validities.

The 1999 Standards for Educational and Psychological Testing largely codified Messick's model. They describe five types of validity-supporting evidence that incorporate each of Messick's aspects, and make no mention of the classical models' content, criterion, and construct validities.

VALIDATION PROCESS.

According to the 1999 Standards, validation is the process of gathering evidence to provide "a sound scientific basis" for interpreting the scores as proposed by the test developer and/or the test user. Validation therefore begins with a framework that defines the scope and aspects (in the case of multi-dimensional scales) of the proposed interpretation. The framework also includes a rational justification linking the interpretation to the test in question.

Validity researchers then list a series of propositions that must be met if the interpretation is to be valid. Or, conversely, they may compile a list of issues that may threaten the validity of the interpretations. In either case the researchers precede by gathering evidence – be it original empirical research, meta-analysis or review of existing literature, or logical analysis of the issues – to support or to question the interpretation's propositions (or the threats to the interpretation's validity). Emphasis is placed on quality, rather than quantity, of the evidence.

A single interpretation of any test may require several propositions to be true (or may be questioned by any one of a set of threats to its validity). Strong evidence in support of a single proposition does not lessen the requirement to support the other propositions.

Evidence to support (or question) the validity of an interpretation can be categorized into one of five categories:

1.Evidence based on test content
2.Evidence based on response processes
3.Evidence based on internal structure
4.Evidence based on relations to other variables
5.Evidence based on consequences of testing

Techniques to gather each type of evidence should only be employed when they yield information that would support or question the propositions required for the interpretation in question. Each piece of evidence is finally integrated into a validity argument. The argument may call for a revision to the test, its administration protocol, or the theoretical constructs underlying the interpretations. If the test and/or the interpretations meant to be made of the test's results are revised in any way, a new validation process must gather evidence to support the new version.

RELIABILITY OF TEST

Reliability refers to the consistency of a measure. A test is considered reliable if we get the same result repeatedly. For example, if a test is designed to measure a trait (such as introversion), then each time the test is administered to a subject, the results should be approximately the same. Unfortunately, it is impossible to calculate reliability exactly, but it can be estimated in a number of different ways.

Test-Retest Reliability. To gauge test-retest reliability, the test is administered twice at two different points in time. This kind of reliability is used to assess the consistency of a test across time. This type of reliability assumes that there will be no change in the quality or construct being measured. Test-retest reliability is best used for things that are stable over time, such as intelligence. Generally, reliability will be higher when little time has passed between tests.

Inter-rater Reliability. This type of reliability is assessed by having two or more independent judges score the test. The scores are then compared to determine the consistency of the raters estimates. One way to test inter-rater reliability is to have each rater assign each test item a score. For example, each rater might score items on a scale from 1 to 10. Next, you would calculate the correlation between the two ratings to determine the level of inter-rater reliability. Another means of testing inter-rater reliability is to have raters determine which category each observation falls into and then calculate the percentage of agreement between the raters. So, if the raters agree 8 out of 10 times, the test has an 80% inter-rater reliability rate.

Parallel-Forms Reliability. Parallel-forms reliability is gauged by comparing two different tests that were created using the same content. This is accomplished by creating a large pool of test items that measure the same quality and then randomly dividing the items into two separate tests. The two tests should then be administered to the same subjects at the same time.

Internal Consistency Reliability. This form of reliability is used to judge the consistency of results across items on the same test. Essentially, you are comparing test items that measure the same construct to determine the tests internal consistency. When you see a question that seems very similar to another test question, it may indicate that the two questions are being used to gauge reliability. Because the two questions are similar and designed to measure the same thing, the test taker should answer both questions

the same, which would indicate that the test has internal consistency.

In the 1910s, for example, researchers Starch and Elliott (1912) made copies of two actual English examination papers and sent them to teachers to grade and return. The marks ranged from 50 to 98 percent. One paper, graded by 142 teachers, received fourteen marks below 80 percent and fourteen above 94 percent. "That is, a paper which was considered too poor for a passing grade by some teachers was rated as excellent by others." Starch and Elliot repeated the procedure with duplicate Geometry tests (1913). Teachers' marks on the 116 returned papers ranged from 28 to 92 percent, with twenty grades below 60 percent and nine of 85 percent and above. According to Lincoln and Workman (1936, 7):

This type of experiment has been repeated many times by investigators and always with similar results. Therefore there is abundant evidence that teachers' marks are a very unreliable means of measurement. Without standardized tests (or standardized grading protocols) in education, we would increase our reliance on individual teacher grading and testing. Are teacher evaluations free of standardized testing's alleged failings? No. Individual teachers can narrow the curriculum to that which they prefer. Grades are susceptible to inflation with ordinary teachers, as students get to know a teacher better and learn his idiosyncrasies. A teacher's (or school's) grades and test scores are far less likely to be generalizable than any standardized tests' (See, for example, Gullickson & Ellwein, 1985; Impara & Plake, 1996; Stiggins, Frisbee, & Griswold, 1989; Woodruff & Ziomek, 2004a, 2004b). (In Phelps, 2008, Table 1 lists some common fallacies proffered by testing opponents, along with citations to responsible refutations.)

When individual teachers, or individual employers for that matter, are given the responsibility to make judgments unanchored by common standards or rules, those judgments tend to float freely in the currents of time, fitting first one context, then another, and then another. Being idiosyncratic to each particular, temporary context, each free-floating evaluation result is not generalizable to any permanent context. It is a judgment that makes sense only to a particular teacher or employer at a particular point in time and space.

According to Professor Stephen G. Sireci (2005, 113), the bad reputation of standardized tests portrayed by some critics "is an undeserved one." He continues People accuse standardized tests of being unfair, biased and discriminatory. Believe it or not, standardized tests are actually designed to promote test fairness. Standardized simply means that the test content is equivalent across administrations and that the conditions under which the test is administered are the same for all test takers.

There is more to subjectivity in decision-making than ethnic, racial, gender, or class bias, however. The fact is that true objectivity requires too much time to be practical in making everyday decisions. Double-blind controlled experiments or program evaluations with random assignment require time, money, and trained professional observation to monitor their progress.

In our daily lives, we make judgments and decisions continuously. We cannot set up a controlled experiment, and wait for the results, every time we must choose which laundry detergent to purchase, where to go on vacation or, for that matter, whom to hire for a job or whom to admit to the last available place at university. The standardized test is more than an antidote to biased judgment. We need standardized tests because each of us is a prisoner of our own limited experiences and observations.

Standardized tests provide an opportunity to make decisions about individuals that are free of subjectivity, be that subjectivity due to bias or Bayesian shortcuts. In developing standardized tests, trained professionals collect empirical data, apply statistical benchmarks, and make detached, objective evaluations.

Standardized tests have provided information for making important decisions at least since the first administration of the Chinese civil service examination many centuries ago (Zeng, 1999, 8). The "scientific" standardized test (with statistically-calibrated score scales), however, is just a century old (Phelps, 2007b, chapter 2). The innovators responsible for the development of the scientific standardized test—e.g., Binet, Simon, Rice, Thorndike—though, likely would be amazed by the improvements made in testing technology within the relatively brief period since—e.g., computer-adaptive testing or open-source, Web-based platforms, such as the Examination Assessment Management System (ExAMS). It would seem that testing technology has improved over time exponentially. Test developers have increased the complexity and technical sophistication of their product in response to market and regulatory demands. Today's standardized tests are better in most every way than their progenitors. They provide more information for the price, and they are more reliable, fair, and valid (when used as they are designed to be used).

Some of today's standardized tests might seem to the average citizen or policymaker as different in character from their 100-year-old ancestors as today's airplanes or automobiles do from their 100-year-old antecedents. Any of you who have tried in plain language to explain to policy makers the concepts of item response theory, differential item functioning, computeradaptive testing, or point-biserial correlation will

know what I mean. The combination of technical complexity and the widespread use of testing for public purposes should elicit a clear, measured, and open public discussion on testing policy. And, I hope that it does where you live. In the United States, unfortunately, the public and policymakers are generally showered with obfuscation, misinformation, and disinformation. The testing policy debate in the United States: The sound of one hand clapping Standardized testing in the United States is an enigma. Arguably, the country hosts much of the world's most advanced technical research and innovation. Yet, debates on testing policy remain primitive and one-sided.

METHODS AND MATERIAL.

Aim and Objectives

The main aim of the present study is to obtain accurate information from basic science to teachers Anzali city Suitable for writing test questions, test principles and preparation of appropriate tests to assess the validity of traits measured.

Methods and material.

Pupulation and sample:

The population of the study included the all of final test questions of the teachers in science subjects of Anzali city in Iran. Using of cluster method sampling, 4 groups of science subjects selected and in the second period 6 teachers for each subject selected. At last 20 paper of each teacher selected totaling 480 number plates were examined.

Tools:

Regarding to gendering the data, the researchers developed a questionnaire with four dimensions. The first dimension (Appearance of the questions)included 7 parts , Readability, The quality of publish, the negative points, the scores of the questions, distance between questions, and space for responding. The second dimensions included the technical appearance of the questions, the third dimensions is literature of the questions and the fourth dimensions is using of sixth level of learning.

Procedure:

Regarding to the nature of the present study, the descriptive methods is applied in the process of study. In this study, three questions about the quality tests of the fundamental principles of testing, Consistent with the objectives of education and was considered valid. Information obtained by a computer and analyzed using spss software was used.

RESULTS AND DISCUSSION

The fundamental of the study focused to obtain accurate information from basic science to teachers Anzali city Suitable for writing test questions, test principles and preparation of appropriate tests to assess the validity of traits measured.

The analysis results show below:

Question 1: How much the teacher has used the regulation of developing scales in final questions? The finding of study shows that all of the groups have appropriate ability in developing questions. the questions in this area has a qood validity. This finding is in agreed with stalings 1982 and sobhani (2001).

Question 2: the second question is that how much the teacher has used the educational aims in final questions? The finding of study shows that most of the teachers are not able to use of rul of learning in their questions and the groups have not enough skill in developing questions in this area. This findings is in agree with Taner 1995 and Esfahani (2001).

Question 3: The last question of the study, mentioned that how much the developed questions has appropriate validity in final questions? The judgment in this area shows that validity of questions in some

groups are satisfy but in the other groups there is not enough validity for developed questions. The questions in this area have a good validity. This finding is in agreed with chiko (1990) and sobhani (2001).
In summary it can be mentioned that the most capable teachers in the principles above set of tests, High percentage of test questions are two levels of knowledge and understanding of the six levels of cognitive domain can be measured And a lower percentage of two-level analysis was devoted to Very small percentage of the surface composition and the evaluation questions were included.
Credit average and above average teachers' groups that represent the Is that enough tests have been discredited. Correlation tests were carried out most of the questions teachers has been positive with the total score.

LIMITATIONS AND SUGGESTIONS

The most important limitations of the study is that in spite of more than one year's searching on planning, the researcher couldn't find all identified researches in this part but This study is a suitable foundation for further research on a similar design. Further study can be carried out on the whole population of province or other parts of the country.

IMPLICATIONS FOR PRACTICE.

This study has produced important and useful information for the official of the Ministry of Education, educators, parents and students. The results of the present study can be used as a guideline for teachers and schools principals to develop valid and acceptable tests.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999) Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
Bayes Rules. (2006). The Economist. Retrieved April 27, 2008 from http://www.economist. com/science /displaystory.cfm?story_id=E1_VPVQGJG
Brookhart, S. M. (2004). Educational assessment of students. Upper Saddle River, NJ: Merrill-Prentice Hall.
Dempster, F. N. (1997). Using tests to promote classroom learning. (pp. 332–346). In R. F.
Dillon, (Ed.). Handbook on testing. Westport, CT, USA: Greenwood Press.
Educational Measurement: Issues and Practice, 12(3), 23.
Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of
Guion, R. M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-398.
Impara, J. C. & Plake, B. S. (1996). Professional development in student assessment for educational administrators. Educational Measurement: Issues and Practice, 15(2), 14–20.
Lieberman, M. (2007). The educational morass. Lanham, MD, USA: Rowman & Littlefield.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.
Phelps, R. P., Ed. (2005). Defending standardized testing. Mahwah, NJ, USA: Lawrence Erlbaum.
Phelps, R.P. (2005b). The source of Lake Wobegon. Nonpartisan Education Review / Articles, 1(2). Retrievable at http://www.npe.ednews.org/Review /Articles/v1n2.htm
Phelps, R.P. (2007a). The dissolution of education knowledge. Educational Horizons, 85(4), 232–247.
Phelps, R.P. (2008). Educational achievement testing fallacies, Chapter 3 in R.P. Phelps (Ed.), Correcting fallacies about educational and psychological testing. Washington, DC, USA: American Psychological Association.
Popham, W. J. (2008). All About Assessment / A Misunderstood Grail. Educational Leadership, 66(1), 82-83.
secondary school teachers of academic subjects: Implications for instruction in measurement,
Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.). Defending standardized testing (111–122). Mahwah, NJ, USA: Lawrence Erlbaum.
Woodruff, D. J., & Ziomek, R. L. (2004a, March). Differential grading standards among high schools. ACT Research Report 2004-2, Iowa City, IA, USA: ACT.