Journal of Education; ISSN 2298-0172; Volume 3, Issue 1, 2014

# Practical strategies for teachers to enhance the degree of reliability and validity in assessment

**Ekaterine PIPIA***

## Abstract

In the context of measurement, reliability refers to consistency and stability of the scores, not the tests themselves, while validity refers to the accurate interpretation of test scores (Reynolds, Livingston & Willson, 2009). The concern of reliability together with validity stems from the original nature of assessment to provide precise information that helps educators make necessary changes to enhance the quality of education. The paper underlines that reliability and validity of this information is of great importance in decreasing the degree of measurement error that is inherent in all measurement. Therefore, the role of educational professionals is to identify the sources of measurement error and minimize their impact on the obtained sources. The research revealed that the teachers need to apply different approaches to eliminate the sources of measurement errors and to estimate the reliability and validity of assessment in practice.

**Keywords:** assessment, reliability, validity, measurement errors, strategies, consistency, stability

## Introduction

It's notable that a number of theories and models have been developed regarding measurement error, but the most noteworthy is **Classical Test Theory** which is very often called **True Score Theory**. According to this theory, every obtained score encompasses two key components: the **true score**, which is the score that would be obtained if there were no errors, and the **error score**: Obtained Score= True Score +Error. The theory is represented in the following equation:

**X= T + E**

**X** (Obtained Score) = represents an observed score of a test taker

**T** (True Score) = represents a test taker's true skills, knowledge and abilities, which is always free form measurement errors

**E** (Error) = represents measurement error, which limits the extent to which test results can be generalized.

It's obvious for many educators that the main interest of assessment is to obtain a true score, but due to the presence of measurement error, we can never know what the true score is. The only feasible way to tackle the problem could be to gain information about the reliability of measurement, so that we could establish intervals around the obtained score and calculate the probability that the true score falls within the intervals specified (AERA et al., 1999). Practically speaking, if we administered parallel forms of a test and had the same person take them at different times, the presence of measurement error would prevent the person from earning the same score every time. The details of time intervals will be discussed below in relation to time sampling error. Consequently, the sources of measurement error should be detected in order to reduce its impact on the final result and of course, measure the reliability and validity of the obtained score. I think it's quite beneficial to be aware of the key characteristic traits of measurement errors to identify the sources and somehow eliminate their existence in the process of test development, administration, scoring and interpretation.

## Sources of measurement error

### Content sampling error

It is quite obvious for many educators that every single test represents a sample which could be deemed a representative of the domain. The discrepancy that exists between the sample of the items (i.e., the test) and the domain of the items (i.e., all the feasible items) causes a **content sampling error** or **domain sampling error** (Aiken, 2000).

The fact that content sampling error is the largest source of measurement error makes the observation process much easier. There is always a high degree of probability to identify how well the test makers (writers) sample the total domain of items. According to Reynolds, Livingston, & Willson (2009): 'if the items on a test are a good sample of the domain, the amount of measurement error due to content sampling will be relatively small'(p. 94). Consequently, if the test items are poor sample of the domain, the amount of measurement error will be of a large scale.

We all agree that a single test may not include every possible question or evaluate every possible relevant behavior. For example, if a teacher administers a test, which is designed to assess students' knowledge in Early American

* Assoc.Prof.Dr., Faculty of Education and Humanities, International Black Sea University, Tbilisi, Georgia
E-mail: ekapipia@ibsu.edu.ge

History, and all the questions refer to American Revolution and no other aspects of American history were covered, we would conclude that these questions are simply a sample and may not be representative of the domain from which they are drawn.

### Time sampling error

Time sampling errors are provoked by the situations in which random changes over time in the test taker (e.g. illness, tiredness, anxiety) or the testing environment (e.g. temperature, noise) affect performance on the test (Reynolds, 1982). Imagine that one of your students did not have breakfast and your exam was just before lunch, s/he might not perform as well as if s/he took the test after lunch. Or take an example of a testing session, where a neighboring class was making noise; the class might have performed better in the afternoon when the neighboring class was less disruptive. According to Reynolds, Livingston & Willson (2009), 'measurement error due to time sampling reflects random fluctuation in performance from one situation to another and limits our ability to generalize test scores across different situations'(p.94).

**Clerical errors** could be detected while adding up a student's score. It is a minor source of measurement error, but still exists in our experience.

## Reliability: practical strategies for teachers

Many educators have multiple options for estimating the reliability of scores produced by their classroom tests. Due to this fact, I would like to discuss below the major approaches that are used in educational assessment to estimate reliability. Generally, there are many ways to estimate reliability, but the following ones could be very easily implemented in our everyday teaching and assessment experience.

a) Test- retest reliability- the same test is administered to the same group in two different situations and the reliability coefficient is obtained by calculating the correlation between the scores (Sheslow &Adams, 2003). If we consider the example of the student who could not perform well on the test in the morning due to sleep-deprivation, we can conclude that test-retest reliability is sensitive to measurement error due to time sampling and provides the stability of scores over time. But at the same time, the length of the interval between the two test administrations should be taken into consideration. If the test-retest interval is too short, the reliability estimate will be affected by memory and if this interval period is long, the reliability estimate may be lowered by the actual changes in the test taker during this period. Therefore, it should be noted here that the way the test is used is an important consideration in determining what an appropriate test-retest interval is in different assessment accommodations.

b) Alternate-form reliability- two forms of the test (parallel forms) are administered to the same group and the reliability coefficient is obtained by the scores of two assessments (Sheslow & Adams, 2003). In this sense, reliability is estimated through simultaneous or delayed administrations of the parallel forms. Alternate form with simultaneous administration implies two forms of the test administered on the same occasion and is sensitive to measurement error due to content sampling. The other, alternate form with delayed administration implies two forms of the test administered on two different occasions and is sensitive to measurement error due to both content sampling and time sampling.

c) Inter-rater reliability- the test is administered one time, but scored by different individuals independently and the correlation is calculated between the scores by the scorers (Sheslow, & Adams, 2003). This approach is designed to eliminate subjective judgments and evaluate the degree of agreement when different teachers score the same test. Inter-rater reliability is mostly estimated in constructed- response items, when teachers' personal biases, preferences or mood influence the score.

In order to draw a clear-cut picture of different processes estimating reliability, I have summed the main characteristic traits of each in the table below.

**Table 1.** Ways to estimate reliability

| Type of reliability to estimate | Test forms | Testing session | Error variance |
|---|---|---|---|
| Test-retest | 1 form | 2 sessions | Time sampling |
| Alternate form (simultaneous administration) | 2 forms | 1 session | Content sampling |
| Alternate form (delayed administration) | 2 forms | 2 sessions | Time sampling and Content sampling |
| Inter-rater | 1 form | 1 session | Subjectivity |

The main aim of estimating reliability of the assessment of the test results is to provide teachers with practical opportunities to make better decisions in their teaching and assessment. It is thought-provoking that there is a close relationship between reliability and validity, but reliability of the test does not guarantee validity of score interpretations. It means that after administering the tests in a reliable manner, the valid interpretations of the assessment are required.

## Validity: practical strategies for teachers

Even though many teachers may complain about the time and resources to conduct validity studies, they can use some practical procedures to evaluate the validity of the results of their classroom assessment.

### Examination of test content

Within this framework, we can discuss content based validity evidence to examine the relationship between the content of the test and the construct it is designed to measure. The focal point here is to identify whether the content of the test is relevant to the content domain. According to Standards (AERA et al., 1999): 'test content includes the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines…regarding administration and scoring' (p.11). Therefore, it is important to design a table of specifications, which is a blueprint guiding the development

Journal of Education; ISSN 2298-0172; Volume 3, Issue 1, 2014

of the test through defining the topics and objectives to be covered at the early stage of writing actual test items.
Examination of student response processes

In this sense, learning objectives and outcomes are of paramount importance, as the teachers are supposed to examine the cognitive and behavioral processes engaged in by students during the test. In other words, the testing items should reflect the same cognitive activities and behavioral abilities that are specified in the learning objectives.

## Examination of test fairness

The teachers should ensure a high degree of fairness to all students regardless of their different ethnic, cultural and political backgrounds. We all come up with one prevalent idea that, it is teachers' responsibility to make sure that the assessment activities they employ in the classroom are developed, administered, scored and interpreted in a technically, ethically, and legally sound manner.

## Examination of practical features

It is obvious that a number of factors may limit the validity of interpretations, but the two major internal threats should be mentioned here: 1) construct under-representation and 2) construct- irrelevant variance. Construct underrepresentation takes place when the items on test do not measure the needed construct- essential content in the specified domain. Construct- irrelevant variance takes place when the items on the test measure the content or skills unrelated to the construct (Feldt, 1997). Teachers should take into account these two factors while developing a classroom test in order to evaluate the correspondence between the test content and its construct. This process will guarantee the valid interpretations of assessment.

Many teachers are not aware of the fact that their positions give them an opportunity to hold considerable power. Every day they make decisions that significantly impact their students' performance inside and outside the classroom. How and when these decisions are made by the teachers in educational assessment was the primary aim of my research, which was conducted with the participation of the lecturers at International Black Sea University, Tbilisi, Georgia.

## Method

I have applied web-based online questionnaires because of its apparent advantages over paper approaches. It gave me an opportunity to reach respondents by sending email invitations to online surveys. Online survey software package www.surveymonkey.com was used for conducting the Internet based surveys. To see the questionnaire online please click the following link - https://www.surveymonkey.com/s/TDMZKQN . The respondents could access the survey questionnaire by clicking on the link emailed to them. Some responses from paper-based questionnaire were added through Manual Data Entry. Totally 25 questionnaires were completed and returned. The aim of the questionnaire was to identify teachers' approaches to eliminate the sources of measurement errors and the ways how to estimate reliability and validity of assessment in practice.

## Limitations of the Questionnaire

Several limitations to this study suggest the need for future research. Even though the questionnaires were very carefully designed, instrumented and analyzed, its validity still needs to be questioned. A number of faults can be identified with questionnaire layout: Results obtained from a horizontally presented four-point scale showed that some respondents repeatedly used one point of the scale. It suggests careless answers that can be explained by the inability of the questionnaire to maintain the respondent's interest. If the questionnaires had offered answer choices in drop-down boxes, the respondents might have chosen their responses with more consideration.

Finally, a range of organizational performance improvements were suggested such as smooth introduction of change, better decision making, more effective resolution of disagreement, increased enthusiasm motivation to get the job done at its best.

## Results

**Question 1: How often do you specify educational objectives and tables of specifications before developing a classroom test?** - As summing up the collected responses, I have received the following data for analysis, where equally 46.15 % of the teachers pointed that they always and sometimes specify educational objectives and 7.69 % responded that they do this rarely. None of them marked the option- Never; which gives a promising picture that teachers more or less define the objectives and tables of specifications in advance and create healthy assessment accommodation in the classroom.

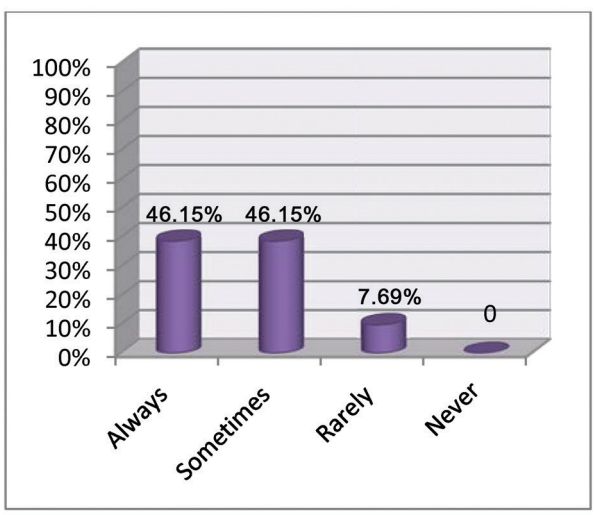


**Figure 1.** *Specification of Educational Objectives and Tables of Specifications*

**Question 2: What kind of testing items do you employ while developing a classroom test?** When asked what kind of items on the test they use while developing a classroom test, 58.33% of the subjects pointed the mixed format of all above mentioned items, 25% of them marked selected-response items and surprisingly, equal percentages (8.33%) were allocated for both constructed-response items and performance assessment. The question revealed that none of the teachers uses portfolio assessment. Only one subject skipped the question.

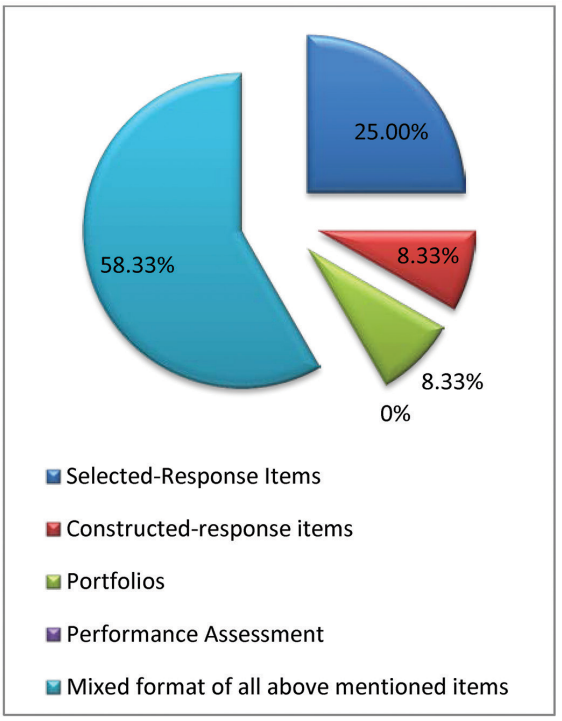


**Figure 2.** *Specification of Educational Objectives and Tables of Specifications*

**Question 3: How often do you provide information to students on the assessment before administering the test?** Related to the frequency of the provided information to students on the assessment before test administration, 92.31% of the subjects responded that they do this always, while 7.69% of them do it sometimes. None of them marked the options- rarely and never.

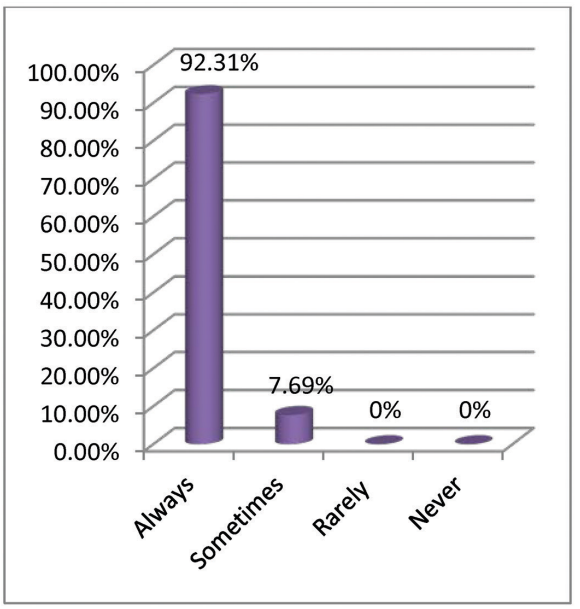


**Figure 3.** *Specification of Educational Objectives and Tables of Specifications*

**Questions 4: Do you develop guidelines for test administration?** The data highlighted that the majority of the teachers (69.23%) develops guidelines for test administration, while 15.38% of them showed their negative attitudes towards the guidelines.

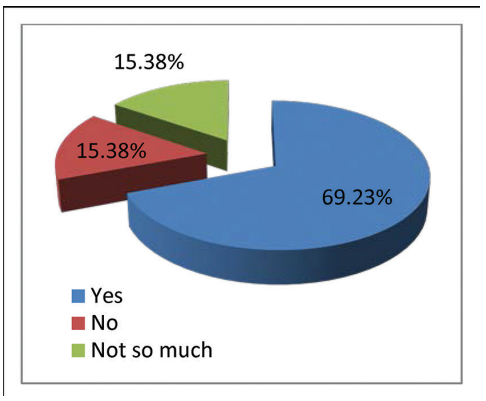


**Figure 4.** *Development of Guidelines for Test Administration*

**Question 5: Do you check if procedures are in place to ensure that assessments are scored properly and the results are reported accurately?** The data revealed that the majority of the subjects (69.23%) check the reliability and validity of the scores. 23.08% of the teachers consider these procedures less important and 7.69 % ignores the importance of properly reported assessment results.

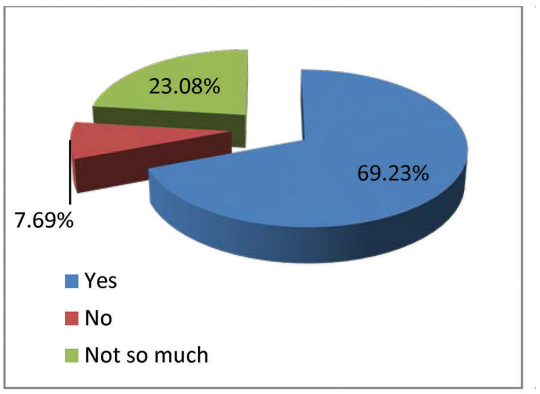


**Figure 5.** *Reliability and Validity of the scores*

**Question 6: If yes, what are your ways of quality control on scoring?** Of those who responded that the process of checking reliability and validity of scores are focal factors in educational assessment,….. pointed out some ways and techniques of quality control on scoring from their own experience:

*'I occasionally check procedures and people who are responsible for administration of exam'*

*'Before tests are held I specify scoring of assessment system in course syllabus and then try to stay on the track'*

*'I report students' background data, statistical relation between data and scores and comparing them'.*

*'Recalculating the points Accepting as correct the answers that did not come to my mind when I made up the test, but which may be correct'*

**Question 7: Do you take into consideration the limitations of the assessment result?** To the question whether they consider the limitations of the assessment result, the majority of the subjects (72.72%) responded that they do care about it and just 27% of them (still) think that the total ignorance of these limitations is not fair. 2 subjects skipped the question.

Journal of Education; ISSN 2298-0172; Volume 3, Issue 1, 2014

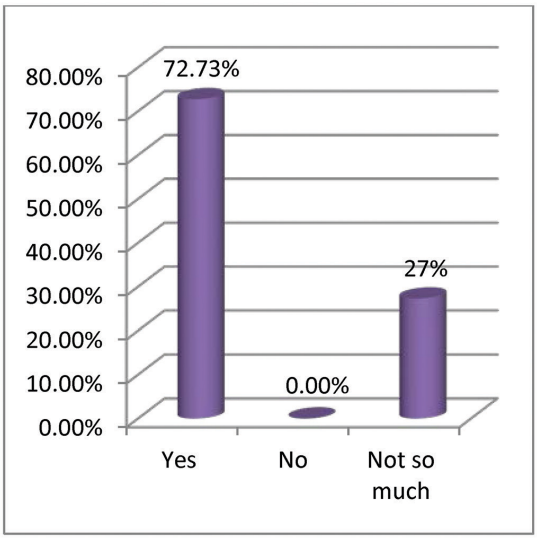


**Figure 6.** *Limitations of the Assessment Results*

**Question 8: Do you take into consideration personal factors or extraneous events that might have influenced test performance?**

**Question 9: If yes, please share your experience**

**Question 10: If no, please explain the reason**

The data revealed that 9. 9% of the teachers do not take into account personal factors and extraneous events while assessing the test results. It is not surprising that many subjects skipped the constructed-response items (questions 9 and 10) to explain their reasons. But thanks to some teachers, I have some comments for my analysis:

*'It is almost impossible to take into consideration everything about human based things'*

*'In written test I cannot take into consideration any kind of student factor. It is very objective'.*

45.45% showed their sensitivity towards these factors and almost the same number of teachers (45%) marked that they not so much pay attention to these factors in students' test performance. 2 subjects skipped the question.

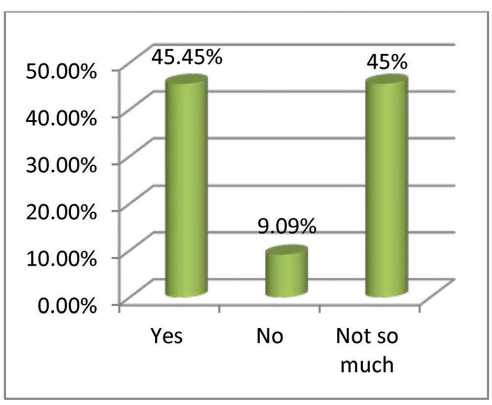


**Figure 7.** *Consideration of Personal and Extraneous Factors*

The teachers, who showed their preference towards the sensitivity of personal and extraneous factors in students' test performance, shared their experience:

*'If the problem is originated from health conditions, I change the exam day for the student'*

*'If a student had a serious reason it needs to be taken into consideration'*

*'Once the best (during the semester) student in the group wrote the paper worse than everybody else in the group. I realized it was due to anxiety. Since then I always include bonus points for a creative task which only a bright student can do'.*

## Conclusions

• The obtained data revealed that 46.15% of teachers always specify the educational objectives and tables of specifications before developing a classroom test. Surprisingly the same percentage of teachers (46.15%) noted that they sometimes determine these factors before the test is administered, while 7.69% totally ignores this process. The data once again stresses that contemporary teachers are very busy and have limited time, it may be tempting to skip these steps and simply commence writing the test. But it should be noted that, this is actually one of the important steps to produce quality tests, as the table of specifications define the content of the test which itself is linked to the educational objectives. The predetermined goals of assessment together with the learning outcomes create the content based validity evidence, which examines the relationship between the content of the test and the construct it is designed to measure. Consequently, the blueprint prepared by the teachers will eliminate the construct underrepresentation and construct-irrelevant variance (both of them refer to content sampling error) which are the main sources of measurement error.

• The majority of teachers (58.33%) pointed out that no single assessment format can effectively measure the diverse range of educational objectives and outcomes. Once the table of specifications is designed, it should be used to develop items of different types: selected-response items, constructed-response items, portfolios and performance assessment. The type of testing items is usually shaped by the specifications of the construct, but this variety really serves to check the mixed ability of the students.

• The data reflected that still there are some teachers (31.77%) who have negative attitudes towards developing guidelines for test administration. In addition to characteristics of the test itself, extraneous factors may impact the reliability and validity of assessment. Failure to give appropriate instructions, suitable testing conditions or follow time limits may lower the students' performance on the test.

• Even though many teachers (69.23%) responded that they do examine and estimate reliability and validity of assessment, most of them were not able to name the practical strategies for estimation and quality control. The discussed practical strategies for teachers to estimate reliability and validity in the points 1.1.2 and 1.1.3 could be a good guideline. It is important to note here that there is no universal approach to estimate reliability and validity and these strategies should be chosen by the teachers in accordance to their particular situation (considering teaching context, specifica-

tions of the construct, students' characteristics, administrative regulations, etc.).

• It is notable that 45.45% of the teachers showed their sensitivity towards personal factors and extraneous events that might influence test performance and almost the same number of teachers (45%) marked that they not so much pay attention to these factors in students' test performance. It is thought-provoking that many teachers are not given an opportunity to take into account the personal and extraneous factors to measure reliability and validity of assessment, as they have to comply with the administrative regulations (internal or external). I think it has to be regulated by the curriculum developers, faculty members and teachers. Consideration of these factors is of a paramount importance, as it is designed to eliminate content (domain) and time sampling errors. In sum, teachers should try to use different ways and techniques to eliminate the existence of measurement errors in order to obtain a true score.

• Reliability and validity of assessment are of paramount importance, as they enhance the quality of the product attained by the end of educational process. The reliable data help professionals to make good decisions in teaching and assessment.

## References

Aiken, L.R. (2000). Psychological Testing and Assessment. Boston: Allyn & Bacon

American Educational Research Association. (1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association

Feldt, L. (1997). *Can Validity Rise When Reliability Declines?* Applied Measurement in Education

Reynolds, C.R., Livingston, R. B. & Willson, V. (2009). Measurement and Assessment in Education, second edition, Pearson

Reynolds, C.R. (1982). The Problem of Bias in Psychological Assessment. New York: Wiley

Sheslow, D.,&Adams, W. (2003). Wide Range Assessment of Memory and Learning 2 (WRAML), Wilmington, DE: Wide Range