

# Establishing Reliability And Validity: An Ongoing Process

Rebecca Lindell<sup>1</sup> and Lin Ding<sup>2</sup>

<sup>1</sup>*Department of Physics, Purdue University, West Lafayette, IN 47907*

<sup>2</sup>*School of Teaching and Learning, The Ohio State University, Columbus, OH 43210*

**Abstract.** Establishing validity and reliability is a necessary step in any conceptual assessment instrument. But once validity and reliability are established, it is not the end of the story. Reliability and validity are not an inherent property of the assessment instrument or its individual items, but something that must be reestablished with any changes of the instrument items, order, administration techniques or population being studied. In this paper we will discuss how validity and reliability can be established or reestablished. We will also discuss common instances in instrument development and use that requires reliability and validity to be reestablished.

**Keywords:** Reliability, Validity, Assessment Development and Use

**PACS:** 01.40.Fk, 01.40.gf, 01.40.G-, 01.50.Kw

## INTRODUCTION

Since the creation of the Force Concept Inventory (FCI) [1] the development of research-based distracter driven multiple-choice instruments has surged. Now nearly every scientific discipline has multiple concept instruments available for their use [2]. The creation of these conceptual assessment instruments (CAI's), require a detailed methodology to produce a valid and reliable instrument. This research process often takes many years to complete [2]. Once created a CAI, it can only be considered to be reliable and valid under the circumstances for which it was established. Reliability and validity are not an inherent property of the CAI or its individual items, but something that must be reestablished with any changes of the instrument items, order, administration techniques or population being studied [3]. In this paper we will discuss how validity and reliability can be established or reestablished. We will also discuss common instances in instrument development and use that requires reliability and validity to be reestablished.

## ESTABLISHING RELIABILITY AND VALIDITY

Establishing validity and reliability is a necessary step in development of any CAI. The following definitions may help to clarify the difference between these two components [3, 4, 5].

**Reliability:** How consistently each CAI is at assessing the concept it purports to measure. More specifically, it is referred to the extent to which the measurement results can be replicated.

**Validity:** How well an instrument measures the construct it is attempting to measure. In other words,

validity is the extent to which the measurement results allow us to make inferences about student conceptual understanding of a concept.

There are multiple methods for establishing reliability and validity. These are discussed below.

### Establishing Reliability

There are several recognized ways for establishing the reliability of a CAI and to produce a reliability coefficient [3, 6]. These include the following:

**Alternative Form Method:** Requires two administrations of two alternative, but similar format of the CAI. Half of the population being assessed complete form 1 of the CAI first followed by form 2, while the other half complete form 2 followed by form 1. The *coefficient of equivalence* between the two forms is determined.

**Test-Retest Method:** Requires two administrations of the same CAI at different times to see if the population consistently scores the same on the CAI. The *coefficient of stability* measures the correlation coefficient between the two administrations.

**Split-Half Method:** Only requires one administration of the CAI. After the CAI has been administered, different items are assigned to a new instrument. Each new instrument is correlated to determine the correlation coefficient. This method works best with very large number of items.

**Method of Item Covariance:** By comparing the variance of each item to the total variance of the CAI, a measure of the internal consistency of the CAI is established. Often referred to as *coefficient alpha*, this can be calculated using several formulas. The most accurate coefficient alpha is *Cronbach's Alpha*.

Several factors can affect the reliability coefficients including the homogeneity of the group, the length of the CAI and the time allowed for completion of the CAI.

### Establishing Validity

Unlike reliability coefficients that require statistical computations, establishing validity can be more difficult. There are many different types of validity that can be established with the development of a CAI. We will focus only on three of the main types [3].

**Criterion Validity:** The degree to which scores on a CAI predicts another criterion. Typically established through comparison to other standardized instruments or course grades.

**Construct Validity:** The degree to which scores can be utilized to draw an inference on the content domain. Typically established through open-ended student interviews.

**Content Validity:** The degree to which CAI measure the content covered in the content domain. Typically established through expert review (a.k.a. Delphi method).

Without establishing the reliability and validity of a CAI, users of the instrument cannot be confident about whether they are measuring what they think they are measuring nor can they be confident that the scores will only change due to student ability and not another variable.

### NEED TO REESTABLISH VALIDITY AND RELIABILITY

Based on our development of CAI's [7, 8] and our experience with the methodologies used to create different CAI's [2], we have determined several instances where changes are made to the CAI that in essence create an alternative CAI, which must independently establish its reliability and validity. These changes include, but are not limited to those shown in table 1.

**Table 1.** Changes to a CAI that requires reestablishment of the validity and reliability and reasons for this requirement.

Changes that Create an Alternative CAI	Reason for Reestablishment
Change in Population:	

Different Age Group	Previous research has shown that for some concepts, students' conceptual understanding and difficulties tends to mature as they increase in age. [7] This can cause differences in the responses selected either correct or incorrect.
Different Courses	In our previous experience, we observed that different courses might attract different majors and or populations even though they may seem similar on the outset [8]. While the Hake factor can take care of differences in pre-tests [1], it will not account for differences in distribution of scores.
Different Geographical Locations	Different geographical locations can have different cultures or different locations for observation. This is more of a case with Astronomy CIA's.
Different Language	When the language of a CAI is changed the reliability and validity must be reestablished. This goes beyond just confirming that the items in the new language say what the original items. The prevalence of difficulties and interpretations may change for this population.

#### Change in administration Different delivery method

Different delivery method	With the advent of modern technologies, many CAI users wish to use the internet to administer the CAI. Differences in how the test is perceived – in questions on a page, where responses are located, time for each question, etc can greatly affect the performance on the CAI.
Different delivery location	If the CAI is no longer administered in the manner for which it is established, difficulties may be encountered. For example if a test is designed to be given in a proctored location, if the test is given as a take home instrument, the results are subject. Students can either use the internet or other resources to determine the "correct" answer without giving what they really think.

#### Change in questions

Different order of questions	Research [11] has shown that different questions can trigger different mental models and that the order affects the results obtained.
Different wording of questions	Changing the language of a question on a CAI requires the reestablishment of the reliability and validity. It is no longer the same instrument and you should not assume it is.

## CONCLUSIONS

Prior to using the CAI for assessing students' conceptual understanding, CAI developers must establish the instruments reliability and validity. However, this reliability and validity is only established for specific testing conditions and specific testing populations. It is the responsibility of developers to communicate this information, as well as the responsibility of the user to become aware of this information prior to use. In this paper, we have presented alternative ways for establishing the reliability and the validity for a CAI, as well as different situations that require users to reestablish the reliability and validity. It is our hope that this paper will serve as both a reference and a caution for anyone developing or using CAI's in their research.

## ACKNOWLEDGMENTS

We wish to acknowledge Neville "Bill" Reay for his assistance in building collaborative research opportunities between the authors. . RL would like to also acknowledge Andrew Hirsch for his continual support and discussions.

## REFERENCES

1. D. Hestenes, M. Wells and G. Swackhamer, *Phys. Teach* **30**, 141-158 (1992).
2. R. Lindell, E. Peak, and T. Foster, Are they all created equal? A comparison of different concept inventory development methodologies, 2006 PERC conf. proc., (2006).
3. L. Crocker and J. Algina, *Introduction to classical and modern test theory*, Reinhart and Winston, Inc, New York: Holt (1986).
4. R. Doran, *Basic measurement and evaluation of science instruction*, National Science Teachers Association, Washington D.C., (1980).
5. S. Messick, Validity. In R. L. Linn (Ed.), *Educational measurement*, 3rd ed., American Council on Education, New York, (1989).
6. AERA (American education research association), APA (American psychological association) and NCME (National council on measurement and education), *Standards for educational and psychological testing*. Washington DC.
7. R. Lindell and J. Olsen, *Proc. 2002 PERC*, New York: PERC Publishing, NY, (2002).
8. L. Ding, R. Chabay, B. Sherwood, and R. Beichner, *Phys. Rev. ST Phys. Educ. Res.*, **2** (1), 7 (2006).
9. R. Lindell, *Enhancing college students' understanding of lunar phases*. Unpublished Dissertation. University of Nebraska-Lincoln (2001).
10. In an unpublished study, courses designed for pre-service elementary education showed vastly different results on the LPCI as compared to non-technical

majors. Part of the difficulty was determined to be differences in the populations. The elementary education majors showed a greater preference for the many of the elementary misconceptions as uncovered in previous research. See reference [9] for references for different misconceptions among younger children that were not discovered among college students.

11. K. Gray, unpublished thesis, *The effect of question order on student responses to multiple choice physics questions*, Kansas State University, Manhattan, KS (2003).