International Journal of Research Studies in Education
2020 Volume 9 Number 5, 15-24

Key considerations in test construction, scoring and analysis: A guide to pre-service and in-service teachers

Khanal, Peshal ✉
Tribhuvan University, Nepal (Peshal.khanal@cded.tu.edu.np)

Received: 13 May 2020
Revised: 11 June 2020
Accepted: 30 June 2020
Available Online: 6 July 2020
DOI: 10.5861/ijrse.2020.5027

ISSN: 2243-7703
Online ISSN: 2243-7711

OPEN ACCESS

Abstract

This article presents key considerations while constructing, scoring and analyzing test items that serve one of the major requirements of teaching both at the school and university levels. In specific, this article provides major issues to be considered in a full cycle of testing, starting from the preparation of test objectives and specification tables to the step of analyzing the effectiveness of each item, the process commonly called item analysis. Two methodological approaches have been used to prepare this guideline – drawing summary steps of testing by reviewing the literature and contextualizing the process and examples for teacher education programs through a teachers' workshop. The analytical framework for reviewing the testing literature constitutes three components – formal curriculum, classroom testing, and content validity. Considering the testing requirements of learning in a formal educational setting, this article helps faculties and students in all teacher education programs for constructing, administrating and analyzing test items and writing a report in a scientific format.

Keywords: test; specification table; scoring; rubrics; item analysis; difficulty level; discrimination index; power of distractors

Key considerations in test construction, scoring and analysis: A guide to pre-service and in-service teachers

1. Introduction

As the accountability in education has been a growing public concern in the twenty-first century, teachers' knowledge, skills and competency for ensuring reliability and validity of classroom tests remain at the heart of teacher education programs in Nepal and beyond (Miller, Linn, & Gronlund, 2009; Reynolds, Livingston, & Willson, 2011). Reliability and validity of a test is an essential component of any classroom test, as such test not only helps teachers decide student grades but also provides a basis for improving classroom instruction and providing feedback to students for better learning (Fives & DiDonato-Barnes, 2013). Therefore, preparing prospective teachers for test construction and standardization is a pressing need for teacher education programs globally.

As part of the pre-service teacher education course in Nepal and beyond, students are required to prepare a test, administer it in a classroom and analyze the effectiveness of the test by carrying out the process of item analysis. This paper aims to enable students of teacher education programs as well as in-service teachers for carrying out the key tasks of test construction, administration and analysis. Rather than being prescriptive, this paper provides some guidelines and examples which help students to adapt them to suit the needs of their subjects and units. In other words, using the guidelines and examples, students can plan an assessment on their own and prepare and analyze test items and scores accordingly. In the sections that follow, I first describe the methodology and framework used for preparing this guideline and then present key requirements of the test development process, in some sequential order. Finally, I conclude the article with some implications for both in-service and pre-service teachers and teacher education programs.

2. Methodology

This guideline is prepared based on two study approaches. First, some key literature on test construction and analysis is reviewed and a summary of test construction, administration and analysis is developed based mainly on the requirements for assessing whether students achieve learning outcomes projected through formal curricula. The analytical framework for reviewing the testing literature constitutes three components – formal curriculum, classroom testing, and content validity (Popham, 2003; Miller, Linn, & Gronlund, 2009). This means, while reviewing the literature on testing, the focus is placed on the measuring strategies of learning achievement that is expected to achieve through the transaction of the formal curriculum in the classroom. And, rather than looking at the strategies and requirements of high-stakes standardized testing, the teacher-made classroom testing is selected as a basis for the analysis. A content validity criterion for testing is set for another component of the review framework which ensures the extent to which the elements within a measurement procedure are relevant and representative of the construct that they will be used to measure. As a second approach, the draft guideline is revised with contextual examples in a workshop organized among the faculties at the Central Department of Education, Tribhuvan University in Nepal. In this workshop, a rigorous discussion was held and a consensus was made for meeting minimum requirements of teacher education courses under the Bachelor and Master degree's programs.

3. Core steps and issues of testing and analysis

Effective testing requires a number of steps to follow, each of which has its own merits and issues. Careful consideration on each of these steps help teachers to make the test reliable, valid, practical and objective. The following sections provide major steps involved in the testing process with some practical problems and key considerations attached in each step.

3.1 Preparation of test objectives and a specification table

The first issue of preparing formal assessment is to ascertain the aims and objectives of testing preferably displayed through a specification table. The major aim of preparing the specification table is to ensure the validity of a test, which is the degree to which the evaluations or judgments we make as teachers about our students can be trusted based on the quality of evidence we gathered (Wolming & Wilkstrom, 2010). Such a table provides a two-way chart to help teachers relate their instructional objectives, the cognitive level of instruction, and the amount of the test that should assess each objective (Notar, Zuelke, Wilson, & Yunker, 2004; Wolming & Wilkstrom, 2010). A major point to note is that teaching is a purposeful activity and assessment is the process of assuring that the students achieve the level of learning specified by the curriculum. Therefore, every test should have a purpose to assess the extent to which students acquire a certain level of learning in terms of knowledge, skill and behavior. Students' learning outcomes are categorized at different levels, from a low recall stage to higher abilities such as analysis and evaluation. Benjamin Bloom's taxonomy of educational objectives suggests that students' learning achievement is categorized at six levels, from simple to complex. They are knowledge, understanding, application, analysis, synthesis and evaluation (Krathwohl, 2002). While devising a test, students should review the objectives and learning outcomes of the content and make a specification table, specifying the levels of learning they want to assess. This table should also include the types and number of test items to be used in a test paper. This is one of the best ways of ensuring the content validity of the test.

An example of specification tables showing levels of learning outcomes and the types and number of questions to assess them are given below:

Table 1

An example of specification table

Unit	Specific Objectives	Type of the questions	Levels of learning			
			Knowledge	Understanding	Application	Higher Abilities
	.....	Short answer		1	1	
		Long answer				2
		Multiple choice	10	3	2	

Note. The numbers in the cells are illustrative only.

3.2 Writing test items, organizing them onto a test paper, typing and proofreading and printing

After preparing the specification table specifying the levels of learning and types of test items, the next step is to write the items in a simple and understandable language. The test items are to be proofread by a language expert to make sure that there are no syntax and grammatical errors. The finalized items are then organized into a test paper with necessary information and instructions. The test items can be organized into a few groups according to the types of questions or the areas and units of learning. The major information to be included in the final test is grade or level, subject, full marks and the breakdown of marks into each item, pass marks, time, and guidelines for answering the items. The finalized test should be printed in good quality in adequate numbers.

3.3 Administering the items

The next step is an arrangement and management of all tasks associated with testing that allows students to respond to each item in a conducive environment and ensures that there is not any hindrance and restriction for responding to the items. While conducting the test, the physical ambiance of the test center is very important. Equally important is providing support to the students to help students facilitate responding to the questions. To facilitate the item analysis process, it would be better if the number of students taking the test is more than 30. If the number of students in one class is very low, the test can be taken in more than one class or school. Before administering the test, students can make a guideline for examinees including the basic rules and consideration for writing answers in the test center.

3.4 Preparing answer key and rubrics

Along with the administration of the test, there is a need for preparing an answer key for multiple-choice items and rubrics for a subjective answer to make scoring more accurate and reliable. The answer key provides the correct answer of the multiple-choice items whereas rubrics provide the basis for scoring the essay-type test. The rubric breaks down the answer of subjective items into different components and provides guidelines to provide marks for each component. It is a plan or scheme for providing certain marks to the answers against specific criteria prepared before the test is administered (Kubiszyn & Borich, 2003). Scoring rubrics are prepared based on the expected response from the students. If the response requires components that are distinct and separate, rubrics specify these components and allocate marks for each of these components. This is called an analytical rubric (Miller, Linn, & Gronlund, 2008). An example of a rubric for writing assignment presented by Miller, Linn, and Gronlund (2008) includes the following seven components – ideas and content, organization, voice, word choice, sentence fluency, conventions and citing sources. The scoring points are allocated accordingly. On the other hand, students' responses can also be rated holistically, in which "rubrics yield a single overall score taking into account the entire response" (Miller, Linn, & Gronlund, 2008, p. 251). In scoring student's work at the school level, the analytical rubric is used more often than a holistic rubric. An example of a rubric is given below:

Table 2

An example of a rubric

Multiple choice questions	Correct answer	Subjective question rubrics	
1	a	Short answer question	Introduction - 1
2	c	1 (full mark 5)	First reason - 2
3	b		Second reason - 2
4	a		
5	d		
6	a	Short answer question	……. 2
7	c	2 (full mark 5)	……. 2
8	c		……. 1
9	d		
10	b	Long answer question	…… 2
11	a	1 (full mark 10)	…… 4
12	d		…… 4
13	c	Long answer question	…… 2
14	b	2 (full mark 10)	…… 4
15	b		…… 4

3.5 Scoring the test

Following the answer key and rubric, the next step is to score each item against the rubric criteria and calculate the total marks obtained by each student. In this step, the examiner should be careful in marking each item, making sure that each item is scored accurately taking account of each of the scoring criteria. Equally important is to double-check the total marks and make sure no answers are left unscored.

3.6 Item analysis of multiple-choice items

Item analysis is the process of analyzing the effectiveness of each item on a test, rather than the test as a whole, for its difficulty, appropriateness, relationship to the rest of the test (Miller, Linn, & Gronlund, 2008; Wright, 2008; Reynolds, Livingstone, & Willson, 2011). Item analysis is useful in helping test designers determine which items to keep, modify or discard in a given test, and how to finalize the score for a student (Reynolds, Livingstone, & Willson, 2011). When we improve the individual test item, it helps to improve the overall quality of the test – hence improve both reliability and validity. It is worth noting that each test item

intends to measure certain knowledge, fact, concept, or understanding and item analysis aims to look at the extent to which item is effective enough to measure what the test intended to measure. When analyzing together, item analysis also examines the class-wide performance of individual test items. Therefore, through analyzing student responses to the individual test item, one can evaluate the overall quality of the test and ensure test effectiveness and fairness (Reynolds, Livingstone, & Willson, 2011).

An item analysis includes the analysis of three important measures of individual test items: difficulty level, discrimination index and power of distractors.

Difficulty level (p) - The difficulty level of an item is the proportion of examinees who answer the item correctly. It is the relative frequency with which examinees choose the correct response (Thorndike, Cunningham, Thorndike, & Hagen, 1991). Therefore, the value of difficulty level ranges from 0 to 1. An important point to note is that higher difficulty indexes indicate easier items. For example, if an item has item difficulty 0.95, this is a very easy item since, by definition, 95% answer this item correctly. In general, Items with difficulties less than 30 percent or more than 90 percent need attention. Such items should either be revised or replaced. An exception might be at the beginning of a test where easier items (90 percent or higher) may be desirable.

What should be the difficulty level of an item in a test? This should depend on the types and use of the test (Kaplan & Saccuzzo, 2009). The first thing is to consider the probability of a correct answer by chance. For example, in a true-false item, 50% of students answer the item correctly by guessing (or chance), therefore, difficulty level .5 could be a low value. But, for a 4-alternative multiple-choice item, there is only a 25% chance to choose the correct answer by guessing. Therefore, the difficulty level of 0.5 could be higher.

The optimal difficulty level for an item is "about halfway between 100% of the respondents getting the item correct and the level of success expected by chance alone" (Kaplan & Saccuzzo, 2009, p. 171). This is calculated by taking an average of 100% (1.00) success and chance performance. For multiple-choice items, the optimum difficulty level is therefore 0.625 and for a true-false item, this is 0.75.

Interpretation of difficulty level of an item: According to Backhoff, Larrazolo, and Rosas (2000), the median difficulty level of the examination should range between 0.5 and 0.6, the values of p being distributed in the following manner: easy items (0.8 and above) 5%; items of medium-low difficulty (0.6 to 0.8) 20%; items of medium difficulty (0.5 – 0.6) 50%; medium-hard items (0.2 – 0.5) 20%; and difficult items (0.2 and less) 5%. The following table summarizes the interpretation criteria of the difficulty level of test items.

Table 3

Criteria for interpreting p-value

p-value	Meaning	Expected number of items
0.8 – 1.00	Easy items	5 %
0.6 – 0.8	Medium-low difficulty	20%
0.5 – 0.6	Medium difficulty	50%
0.2 – 0.5	Medium-hard items	20%
0.0 – 0.2	Hard items	5%

Source: Backhoff, Larrazolo, and Rosas (2000)

Discriminating index (D) - Discrimination index is the value that shows the ability of an item to differentiate between the high achieving and low achieving students. In other words, D-value measures 'the extent to which a test item discriminates or differentiates between students who do well on the overall test and those who do not do well on the overall test" (Kubiszyn & Borich, 2003, p. 198). The value of D ranges from -1 to +1 with an ideal score of +1.00. Positive coefficients indicate that high-scoring examinees tended to have higher scores on the item, while a negative coefficient indicates that low-scoring students tended to have lower scores. According to Kubiszyn and Borich (2003, p. 198) there are three types of discrimination indexes:

- Positive discrimination index – those who did well the overall test chose the correct answer for a particular item more often than those who did poorly on the overall test
- Negative discrimination index – those who did poorly on the overall test chose the correct answer for a particular item more often than those who did well on the overall test
- Zero discrimination index – those who did well the overall test and those who did poorly on the overall test chose the correct answer for a particular item with equal frequency.

Interpretation of D-value: According to Ebel and Frisbie (1986), a highly discriminating item has a value greater than 0.4, and a lower discriminating item has a value closer to 0 (see table below). If a D-value of an item is negative, the item is discriminating against the students negatively, which is undesirable at all. Negative D values show that high achievers making the items wrong and low achievers making the items correct.

Table 4

Criteria for interpreting D-value

D =	Quality	Recommendations
0.40 and up	Excellent	Very good items
0.30 – 0.39	Good	Reasonably good but possibly subject to improvement
0.20 – 0.29	Mediocre	Marginal items, usually needing and being subject to improvement
Below 0.19	Poor	Poor items, to be rejected or improved for revision

Source: Ebel and Frisbie (1986)

Power of distractors - In this process, we analyze the responses in each of the alternatives or options in multiple-choice items to assess the extent to which the distractors are plausible. In other words, the distractor analysis aims to assess how the distracters are able to function effectively by drawing the test takers away from the correct answer. Distractor analysis can be a useful tool in evaluating the effectiveness of distractors in multiple-choice items. In a multiple-choice format, distractors are to be effective, so that each distractor has the potential to attract students while they are looking for the correct alternatives. Otherwise, there is a greater possibility that students will be able to select the correct answer by guessing as the options have been reduced.

Calculation and analysis of difficulty level, discrimination index and power of distractors - The general procedures for analyzing an item based on the above three methods are as follows:

- Arrange the answer books from the highest total scores to lowest scores.
- Keep the upper and lower 27% answer books separately and use these two groups for analysis (middle 46% copies are not considered, keep them away). The percentage of 27 percent is used because "this value will maximize differences in normal distributions while providing enough cases for analysis" (Wiersma & Jurs, 1990, p. 145). For example, if you have 30 answer books arranged from the higher scores to lower scores, 27 % of 30 = 8.1 (= 8). Therefore, select 8 copies from the upper group and 8 copies from the lower group. Now we can calculate the difficulty level and discriminating index of each item.
- Difficulty level. Let's start from item 1. First, we will count the number of students who answer item 1 correctly from the upper group. Let's say this number is 7. Similarly, we will count the number of students who answer item 1 correctly from the lower group? Let's say this number is 2. We know that the total number of students in the upper and lower group is 8 each.

- Then, difficulty level of item 1 $(p_1) = \frac{7+2}{8+8} = \frac{9}{16} = 0.56$

  Formula,

  If the number of students answers the item correctly in the upper group = U_r

  If the number of students answers the item correctly in the lower group = L_r

  If the total number of students in the upper group = U_n

  If the number of students in the lower group = L_n

  The difficulty level of the item $(p) = \frac{Ur + Lr}{Un + Ln}$

  The easiest item has a p-value closer to 1 and a difficult item has a p-value closer to 0.

- Discriminating value. The discrimination value of item 1 $(D_1) = \frac{7-2}{8} = \frac{5}{8} = 0.62$

  Formula,

  If the number of students who answers the item correctly in the upper group = U_r

  If the number of students who answers the item correctly in the lower group = L_r

  If the total number of students in the upper group = U_r

  If the number of students in the lower group = L_r

  Discrimination index of item $(D) = \frac{Ur - Lr}{Un \; or \; Ln}$

- Power of distractors. To analyze the effectiveness or power of distractors, we compare responses on key and distractors between upper and lower groups. We count responses on the correct answer and distractors and use logical analysis to assess the extent to which distractors are plausible. For a good item, distractors and key (correct option) should meet the following two criteria (DiBattista & Kurzawa, 2011): the correct option or key should be chosen by a majority of students and the number of students selecting the key in the high scoring group should always be greater than the number of students selecting it in the lower group, and, all distractors should be plausible, which should reasonably lure students away from the key. Therefore, some students who are not fully knowledgeable about the content area are likely to select them. Haladyna and Downing (1993) have suggested that at least 5% of examinees should select each of an item's distractors, and this value is a common benchmark for distractor functionality. In the tables below, there are examples of various types of multiple-choice questions based on the power of distractors.

Table 5

An example of 18 students' response on a good multiple-choice item

Item No. 1	a	b*	c	d	* correct answer
Upper	1	7	1	2	
Lower	1	4	1	1	
Total	2	11	2	3	

Analysis - Correct option 'b' is chosen by 11 students (out of 18). Among them, 7 students in the upper group answer the item whereas only 4 students from the lower group answer it. All three distractors attract the students reasonably (at least 5%). Therefore, Item no. 1 is a good item. While analyzing the responses in this way, there are three different cases in which items to be considered revising. They are whether an item is miskeyed, whether responses to the item are characterized by guessing, or whether an item is ambiguous (Kubiszyn &

Borich, 2003). Let's see the examples provided.

Miskeying – When an item is miskeyed, most students from the upper group will select an option that is a distractor, rather than the option that is keyed.

Table 6

An example of 15 students' responses on a miskeyed multiple-choice item.

Item No. 2	a	b	c*	d
Upper	1	5	1	1
Lower	3	1	2	1
Total	4	6	3	2

Note: In this example, the correct option is 'b', this is miskeyed to 'c'.

Ambiguity – There could be an ambiguous distractor when students from the upper group choose one of the distractors with about the same frequency.

Table 7

An example of 15 students' response an ambiguous multiple-choice item

Item No. 3	a*	b	c	d
Upper	4	4	1	1
Lower	2	1	2	0
Total	6	5	3	1

Note: In this example, distractor 'b' is ambiguous.

Guessing – In case of guessing, students from the upper group respond in a more or less random fashion. According to Kubiszyn and Borich (2003, p. 203), this case is likely to exist when "the item measures content that is i) not covered in class or text, ii) so difficult that even the upper-half students have no idea what the correct answer is, or iii) so trivial that students are unable to choose from among the options provided.

Table 8

An example of 20 students' response on a multiple-choice item mostly by guessing

Item No. 4	a*	b	c	d
Upper	4	4	3	4
Lower	2	1	2	0
Total	6	5	5	4

The above examples suggest that while distractors in multiple-choice items should be clearly incorrect and key must be definitely correct, the distractors should seem likely or reasonable to students who are not sufficiently knowledgeable in the content area. If a distractor is so obvious to be incorrect that almost no examinee will select it, such items do not contribute to the discriminatory power and performance of the item. Therefore, the plausibility of distractors is a major requisite for any functional multiple-choice items.

4. Conclusion and implications

Test construction, scoring and analysis and use of test results are key areas for teachers in order to make assessments reliable and accurate as well as to help students improve learning and administrators to plan for school improvement as a whole. A good teacher education program, therefore, should incorporate courses that prepare students for designing and analyzing the test. Test analysis is particularly useful in multiple-choice items which are used increasingly both as classroom and standardized tests. An effective multiple-choice item should have reasonable levels of difficulty level and discriminatory power, and more importantly, its distractors should be equally plausible so that test items could differentiate students based on their cognitive ability and mastery over the content. In this context, this article could be useful for those responsible for constructing, administrating and analyzing test items and thereby they could contribute to standardizing the test and enhance the credibility of

the test results.

A significant aspect of this article is the reflection from the literature about what requires a good test and what procedures are essential for constructing test items and assessing whether each item of the test is effective in multiple-choice items in particular. As a useful resource, this article provides a framework in which educational objectives are set and pupils' progress charted and expressed. Certainly, no specific instrument for measuring is perfect. However, considering the use of tests as a pivotal means of assessing and reporting students' progress, this article contributes to the field of educational testing by offering practical steps involved in assessing children's learning. More importantly, this article provides useful steps and examples for analyzing the effectiveness of each item (item analysis) of a test containing multiple-choice items. As the plausibility of distractors in multiple-choice items is always a concern and many novice teachers fail to select appropriate keys and distractors while devising multiple-choice items (Kubiszyn & Borich, 2003; Miller, Linn, & Gronlund, 2009), the examples and suggestions in this article could provide greater insight into the key areas of difficulties. While distractors are prone to ambiguity, and miskeying and guessing are most likely due to the poorly constructed multiple-choice items, educators should consider these three possible areas of error as a parameter while reviewing and finalizing the items.

This article has a practical implication for both students pursuing a teacher education degree and teachers working in schools and colleges. The students in the teacher education program could prepare themselves for their teaching career by broadening their knowledge and understanding of testing and by developing skills for constructing the test, scoring students' responses and analyzing the test items to ensure their effectiveness and suitability. In-service teachers could employ the above strategies to inform, improve and evaluate classroom instruction. Effective teaching creates learning opportunities for students, but without testing, teachers fail to ensure that all students are learning effectively. As Popham (2003) rightly contends, "*how* a teacher tests— the way a teacher designs tests and applies test data—can profoundly affect *how* well that teacher teaches" (p. 1), the success of teaching is measured through the ability of teachers for providing an equitable learning opportunity for learners as well as for ensuring whether students grasp the intended learning through effective testing. One of the major responsibilities of teachers is to decide students' learning, and such a decision should be based on the evidence of learning. In most cases, teachers make inferences using test scores. As the accuracy of such inferences is always crucial, testing should ensure that each test item is effective enough to measure the learning outcomes of the students and the procedure for transforming response to score is free from error and bias. Any professional teacher training program, therefore, should consider the systematic and scientific way of designing, administrating, scoring and reporting the test result. Otherwise, assessment in the forms of test and examination may lead to some destructive consequences, which Stobart (2008) indicated clearly in his seminal book 'uses and abuses of assessment'. The main point is that making claims and decision about students' learning through ill-prepared testing could make a debilitating impact not only on students' learning and performance, but largely on their psychological and emotional wellbeing and, in some circumstances, on their life as a whole.

5. References

Backhoff, E., Larrazolo, N., & Rosas, M. (2000). The level of difficulty and discrimination power of the basic knowledge and skills examination. *Revista Electrónica de Investigación Educativa*, *2*(1), 1–16.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 1-23. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of education measurement*. Englewood Cliffs, NJ: Prentice Hall.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research, and Evaluation*, *18*(3), 1-7.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*, 999-1010. https://doi.org/10.1177/0013164493053004013

Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing: principles, applications and issues* (7th ed.). Belmont, CA: Wadsworth.

Kline, T. J. B. (2008). *Psychological testing*. Thousands Oak, CA: Sage.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice*. River Street, Hoboken, NJ: Wiley & Sons.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Educational test and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson Education.

Notar, C. E., Zuelke, D. C., Wilson, J. D., & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. *Journal of Instructional Psychology*, *31*, 115-129.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2011). *Measurement and assessment in education*. New Delhi: Prentice Hall.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. New York, NY: Routledge. https://doi.org/10.4324/9780203930502

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York, NY: MacMillan.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Boston, MA: Allyn and Bacon.

Wolming, S., & Wikstrom, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, *17*, 117-132. https://doi.org/10.1080/09695941003693856

Wright, R. J. (2008). *Educational assessment: Test and measurements in the age of accountability*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781483329673