PJERE, June 2017, Vol. 2, No. 1, pp. 101-104

Book Review

Title of Book:	**Developing and Validating Test Items**
Authors:	**Thomas M. Haladyna** **Michael C. Odriguez**
Year of Publication:	**2013**
Publishing Agency:	**Routledge, Taylor & Francis Group**
City:	**New York**
Reviewer:	**Dr. Muhammad Saeed**

It is a fact that instruction is guided by assessment and evaluation. No doubt, literature cites ample evidence on the significance of varying techniques of assessment which most commonly, referred to as alternative assessment techniques (AATs) and plays a significant role in assessment of students' learning at all levels of education. But perhaps all these techniques do not appropriate when we have to: a) accurately measure or quantify the students' achievement, b) make international comparisons of students' achievement in various disciplines and/or c) award scholarships to selected individuals. In all such cases, we have to depend upon tests. Hence from this notion, need and importance of testing cannot be denied.

I think this book (Haladyna & Rodiguez, 2013) is one of the good master piece in the field of development and validation of different kinds and categories of test items. On one hand, the book describes the science and philosophy of item development while on the other hand; it addresses the issues, warnings and offers suggestions about future efforts.

The general outlay of the book is attractive. This comprehensive book of 446 pages is divided into six parts. Each part is split into reasonable number of chapters or unit ranging from three to five, except the last part which explores the future prospects of item development and validation. The book presents reasonable number of headings and subheadings in bold and/or Italic that are mostly recommended in every good academic work. Each section/part starts with an overview of chapters included in it and each chapter ends with a summary of about 100-200 words. A long list of references is given by alphabetical order at the end that allows the readers to consult further books/other readings on the subject.

Part 1 of the book titled as "A Foundation for Developing and Validating Test Items". It describes role of validity in item development, process of test item development, pros and limitation of cognitive taxonomy,

and procedure of selecting items, but the philosophy of testing is somehow not well described in this section. The authors have explained the concept of test item referring to other worthwhile work (e.g. Kane, 2006b; Downing & Haladyna, 2006; Brennan, 2006; Cronbach, 1971; Messick 1989, 1994). The authors have explained the planning and development process of both selected response (SR) items and constructed response (CR) items. This section also entails the properties of items bank procedure of recruiting item writers and their key assignments, guidelines and training on item writing, and reviewing test items with regard to mainly fairness, language complexity and editorial. With regard to content and cognitive demand of test items, it describes the limitation of the cognitive taxonomy for classifying cognitive demand and also adds recommendations in the context of knowledge, skills and abilities. The section analyses the notion that cognitive ability is represented by a model of cognition of learning which includes: a) Specification of declarative and procedural knowledge, b) a measurement plan, c) hypotheses and evidence that accepts or rejects hypotheses, d) description of threads that more learners from novice toward expert, e) consideration of factors affecting learning, and f) consideration of construct-irrelevant variance that may diminish validity. Further the discussion on procedure of item format generates four fundamental types of item formats the three i.e. objective versus subjective scoring, selection versus production, fixed-response versus free-response, and product versus performance. Each format has advantages and limitations for the users in view of the nature of subject and objectives to be measured. The chapter discusses the concept of Differential Format Functioning (DFF) quoting the merits of Beller and Gafui (2000), Demars (2000), Garner and Engel Hard (1999), Hamilton (1998), and Wightman (1998) who conclude that performance is not directly related to the format of the item. The discussion ends with the key recommendations about choosing an item format that the first priority should be given to what is measured rather than how it is measured (Beller & Gafni, 2000) where CR and SR items are viable options, the SR item is generally the optional choice (Rodriguez, 2002).

Part-II "Developing Selected Response Items" discusses the different selected response formats in case of MCQs, Alternative Choice/Response (True/False) Format etc. with adequate examples along with relevant examples to explain each format. These guidelines have evolved since the publication of an initial taxonomy by Haladyna and Downing (1989a, 1989b); and Ellsworth, Dunnel, and Duell (1990); and Haladyna, Downing and Rodriguez (2002). The authors clarify each format with examples how to write the stem/statement of an item and what are the basic principles/rules of writing last options/alternatives. The guidelines are also provided for different types of matching formats—normal test matching, extended matching and triple matching. There seems continuity and sequence which

highlights rules of each format. The authors have nicely discussed the Automatic Item Generation (AIG). Referring to this, they quote definition of "item shell" (closing items) as a hollow item containing a syntactic structure that is useful for writing sets of similar items. They also explain guidelines for survey items, especially avoiding double—barreled items, clarity, avoiding slang and negative words etc.

Part-III is titled as "Developing Constructed-Response Items" starts with the characteristics and procedure of constructed –Response item formats. The fundamental design of all CR items has the universal four elements: a) content and cognitive demand, b) instruction to test criteria. It further highlights that CR item taxonomy is based on four level competency-predictive, analytical, interpretive and factual recalls (Oterlind & Merz, 1994). The typology of CR item formats generally includes anecdotal record, close (a technique for measuring reading comprehension and fluency), discussion, essay, exhibition, experiment, interview, oral exam/viva, performance tasks, portfolio, research papers and writing samples. The guidelines on CR items are given referring to the previous work (Downing, 2006, Farrara & DeMauro, 2006; Lane & Stone, 2006; Schmeiser & Welch, 2006; Surecu & Zebusjt, 2006; Nulch, 2006). This section discusses the scoring procedure of CR items. This procedure is usually employed when we have to measure a complex cognitive behavior. The authors discuss three main scoring formats: a) objectively scored formats; b) subjectively scored formats; and c) automated scoring (scoring procedure used for large scale assessment/testing).

Part-IV is titled as "Unique Applications of Selected Response and Constructed Response Formats" starts with discussion on developing and validating items to measure writing ability of students. The measurement of writing ability has many challenges to overcome to achieve a high degree of validity. For most is the definition of the construct of writing. Then there may be an issue of scoring the response to the item. The authors' idea of credentialing carries its uniqueness, which they mean to earn a certificate indicating high achievement in a profession or get a license to practice in profession usually in the respective state/province/country. The discussion leads to the general theory for test accessibility, review the characteristics of exception ability with special attention to English language learners and individuals with disabilities, and review accessibility research supporting the existing SR and CR item—writing guidelines. The authors discuss guidelines for writing SR items to support item accessibility (e.g. context fort and style concerns, and writing the stem and options) and guidelines for writing CR items (e.g. content and context concerns and writing the direction/stimulus). Further the guidelines explain the item characteristics with reference to analysis in the terms of item difficulty, item discrimination, role of item response theory (IRT) in item discrimination, the relationship

between item difficulty and discrimination, and the standards for evaluating difficulty and discrimination. The last chapter entails the brief history of testing, the areas of inquiry that may affect item development future-role of theory, development of new formats, research, and the influence of technology. All these are interrelated, but each represents a specific scholarly focus that constantly needs attention if the science of item development is to flourish.

Summing up the above discussion, I would say that the authors have covered most of the important aspects of test item development and validation in excellent way that has contributed to the existing knowledge in educational assessment. Perhaps it would be nice if the authors may consider the following aspects in the future editions.

1. No doubt as per title of the book, the thrust of the book should be on test development and its validation, but I think it may be useful if the authors may add some main advantages and limitations of the tests in the start of the book as tests do not truly assess students' learning that is why other alternative assessment techniques such as observations, oral questioning, quiz, portfolio etc. are used in classroom context.
2. More discussion on Theories of Test Development, especially Item Response Theory (IRT), and Bloom's taxonomy of educational objectives because these form the theoretical bases or philosophy of test development.
3. More description of item analysis may further improve the scope of this book, as this is one of the key measures of validation of selected response and constructed response test items.
4. Somewhere examples from developing countries context should also be given as a significant majority of the readers belong to such countries and such examples or interpretations would be perhaps useful for the readers of developed nations.
5. The authors may further probe into ways of validating test items of different categories while developing a classroom test and a standardized test, as there exist certain commonalities and differences in the validation of test items in both of these streams.