Available online at www.sciencedirect.com

ScienceDirect

Currents in Pharmacy Teaching and Learning 8 (2016) 31–38

Currents in Pharmacy Teaching & Learning

http://www.pharmacyteaching.com

Research

Measurement characteristics of a concept classification exam using multiple case examples: A Rasch analysis

Nicholas B. Jennings, MPH[a], Marion K. Slack, PhD[b,*], Lea E. Mollon, PharmD[b], Terri L. Warholak, PhD[b]

[a] Division of Community, Environment, and Policy, College of Public Health, University of Arizona, Tucson, AZ
[b] Department of Pharmacy Practice and Science, College of Pharmacy, University of Arizona, Tucson, AZ

Abstract

Objective: To determine if an exam using multiple cases to test research design concepts measured only one cognitive skill, concept classification, and to determine if item difficulty varied according to the research design used for the case.
Methods: The exam consisted of 50 multiple choice items associated with five example abstracts: a randomized controlled trial, pretest-posttest, crossover, retrospective cohort, and descriptive designs. A Rasch analysis was conducted to determine dimensionality (i.e., measured a single skill). Items were stratified by design to explore the relationship between item difficulty and study design. Overall difficulty was assessed using an item–person map.
Results: The exam was administered to 101 students; the mean was 88.4% (mean score = 44.2; SD = 3.5). The Rasch analysis indicated the exam primarily measured one cognitive skill, presumably concept classification. The stratified analysis indicated that overall no single research design was more difficult than other designs; however, the type of research design and item topic interacted so that an easy item for one design could be difficult when associated with a different study design.
Conclusions: The exam appeared to function more like a mastery exam documenting that most students performed well rather than as an exam for ranking students by ability. That item topic interacted with study design to affect item difficulty, indicates that items on the same topic are needed to test basic design concepts across study designs.

Keywords: Concepts; Multiple choice exam; Rasch analysis; Research design

Introduction

Cases are often used for testing students on clinical or scientific skills in the health sciences. The student is presented with a clinical case then asked, for example, to identify an appropriate diagnosis for the patient or an appropriate drug to treat the disease. If the case is an abstract of a research report, the student is asked to identify the type of research design or identify an appropriate statistical test for the reported findings. The cases represent examples from categories related to the topic under consideration. For instance, the case could be a patient with type 2 diabetes who is being treated with metformin or the case could be an abstract from a randomized controlled trial (RCT) of metformin in the treatment of type 2 diabetes. In both instances, students need classification skills to correctly categorize (diagnose) the patient as having type 2 diabetes, categorize metformin as an appropriate therapy, or categorize the design of the example research as that of an RCT.

The ability to categorize items or cases is described by concept classification theory also known as schema theory. Concept classification theory posits that humans organize

* Corresponding author: Marion K. Slack, PhD, College of Pharmacy, University of Arizona, 1295 N. Martin, PO Box 210202, Tucson, AZ 85721-0202.
E-mail: slack@pharmacy.arizona.edu

http://dx.doi.org/10.1016/j.cptl.2015.09.010
1877-1297/

their knowledge about objects and ideas into groups of similar items, that is, into categories.[1] Skilled performance is thought to depend on information being stored as categories or schemas in long-term memory that are instantly available and that contain large amounts of information in expert performers. Instruction should facilitate the acquisition of the categories relevant to particular domains such as research design.[2,3] In medicine, the concept classification skill required to recognize studies as representing specific research design categories has been identified as a skill students should have at graduation.[4]

The approach of Merrill and Tennyson[5] was used to the develop instruction and exams for research design concepts. The objectives for this module on research design were for students to be able to recognize the type of research design used in a specific study and to be able to identify the study characteristics and the validity issues associated with each type of study design. The basic instructional strategy was to present an example study representing a specific design (e.g., a RCT), identify the pertinent characteristics (e.g., uses random assignment to groups), identify the level of measurement of the dependent variable and the statistical test used (e.g., *t*-test or Chi square), and identify the primary threat to internal validity for that research design (e.g., for a RCT, the primary threat is attrition). Exams consisted of abstracts for five different study designs with 50 accompanying multiple choice questions asking students to identify the specific research design, identify characteristics of the study, including an appropriate statistical test to use with the reported data, and identify the primary threat to internal validity. Practice homeworks similar to the exam were provided throughout the module; however, all cases on the exam were new to the student; there were no study designs on the exam that were not discussed in class. The module is part of a four-course sequence; basic statistics are taught first, then research design, followed by the drug information course, and the student research project course.

Students generally do well on the exam; the class average has ranged from 82% to 88% correct for individual classes of 93–101 students over the past five years on the 50-item exam. However, the measurement characteristics of the exam were unknown. For a similar exam on research design, Jackson et al.[6] used a Rasch analysis to identify the characteristics of the exam. Rasch analysis statistically compares the target instrument to a specific response model based on the premise that scores represent a unidimensional scale (a scale that measures one construct; it is important for a scale to measure only one construct because if it measures more than one, the user does not know which one the student used to respond) that progresses from easy to difficult.[7]

Person ability is calculated from the odds of a specific person (total number of items correct divided by the total number of items incorrect) being able to correctly respond to a random question. For instance, the odds of a person with average ability being able to correctly answer easy questions is high but becomes progressively lower as the items become more difficult. Item difficulty is calculated from the odds of a specific question being answered correctly by a random individual. Person ability and item difficulty are then transformed onto a single logit scale (i.e., the log of the odds ratio) so that person ability and item difficulty can be compared directly, for example, on a graph that displays both person ability and item difficulty. Based on the measure of item difficulty, the target instrument can be statistically compared to the Rasch model to determine if the target instrument fits the model, that is, whether or not the target instrument is unidimensional and the items independent (i.e., a response to one item does not depend on the response to another item). Individual items that do not fit the model can indicate poor wording or poor response options. The responses of individual persons can be compared to the model to identify problems such as cheating, carelessness, miscoding, and lucky guessing.[7]

Classical test reliability can be calculated for persons to indicate how likely a specific person will be similarly ranked relative to other persons on a second administration of the instrument. A measure of item reliability is also provided, which indicates how likely a specific item would be similarly ranked as easy or difficult on a second administration of the exam.[7,8] A Rasch analysis can be used with either a mastery exam when the objective of testing is to document that individuals can perform a specific skill or with a classical exam when the objective of testing is to rank individuals as having more or less ability.

In contrast, classical test statistics are based on the assumption that the purpose of the exam is to rank persons according to ability and provides no information on mastery. Classical test reliability measures such as test-retest reliability provide a measure of consistency in the ranking of individual abilities.[9] If individuals rank high compared to other individuals on the first administration then, if the test is reliable, they should rank high compared to other individuals on the second administration. The ranking does not indicate if the person has mastered the material; the person with the highest score on the exam may know very little. A Cronbach's α or K-R 20 indicates whether or not all the items rank individuals consistently,[9] that is, whether the items are measuring one versus multiple skills. Classical test statistics also include a point-biserial statistic for individual items, which indicates the degree to which the responses on a specific item correlate to the total test score.[10] Point-biserials are highest for items in which all high-scoring individuals respond correctly and all low-scoring individuals respond incorrectly so that the item reliably discriminates between high-ability individuals and low-ability individuals. Point-biserials are low if everyone responds correctly—which can happen if the purpose of the item is to document that everyone can perform the skill.

When Jackson et al.[6] used Rasch analysis for their exam using a single research design case with 25 questions and an