Determinants of Difficulty and Discriminating Power of Image-based Test Items in Postgraduate Radiological Examinations

Dirk R. Rutgers, MD, PhD, Fleur van Raamt, MD, PhD, Anouk van der Gijp, MD, PhD, Christian Mol, MSc, Olle ten Cate, PhD

Rationale and Objectives: The psychometric characteristics of image-based test items in radiological written examinations are not well known. In this study, we explored difficulty and discriminating power of these test items in postgraduate radiological digital examinations.

Materials and Methods: We reviewed test items of seven Dutch Radiology Progress Tests (DRPTs) that were taken from October 2013 to April 2017. The DRPT is a semiannual formative examination, required for all Dutch radiology residents. We assessed several stimulus and response characteristics of test items. The response format of test items included true or false, single right multiple choice with 2, 3, 4, or ≥5 answer options, pick-N multiple-choice, drag-and-drop, and long-list-menu formats. We calculated item *P* values and item-rest-correlation (Rir) values to assess difficulty and discriminating power. We performed linear regression analysis in image-based test items to investigate whether *P* and Rir values were significantly related to stimulus and response characteristics. Also, we compared psychometric indices between image-based test items and text-alone items.

Results: *P* and Rir values of image-based items ($n = 369$) were significantly related to the type of response format ($P < .001$), and not to which of the seven DRPTs the item was obtained from, radiological subspecialty domain, nonvolumetric or volumetric character of images, or context-rich or context-free character of the stimulus. When accounted for type of response format, difficulty and discriminating power of image-based items did not differ significantly from text-alone items ($n = 881$). Test items with a relatively large number of answer options were generally more difficult, and discriminated better among high- and low-performing candidates.

Conclusion: In postgraduate radiological written examinations, difficulty and discriminating power of image-based test items are related to the type of response format and are comparable to those of text-alone items. We recommend a response format with a relatively large number of answer options to optimize psychometric indices of radiological image-based test items.

Key Words: Educational measurement; internship and residency; learning; radiology.

INTRODUCTION

Radiological knowledge and visual skills are imperative in the profession of a radiologist and are therefore trained during residency. To assess whether residents have reached sufficient competence in these domains, both workplace assessments and written examinations are essential. In many countries, written examinations are included in radiology residency programs (1–4). These examinations usually contain image-based test items in addition to text-alone items. Although radiological images were traditionally x-ray photos, they are nowadays, to a large extent, obtained through cross-sectional techniques, such as computed tomography scanning and magnetic resonance imaging.

Test Item Characteristics

Viewed from the perspective of test theory, test items in examinations can be characterized by their stimulus and their response format (5). The stimulus refers to what is being asked by the test item and is considered the main determinant of what type of competence is tested (5). The stimulus may be context-rich or context-free. Context-rich items consist of a (case) scenario and ask for decisions that are related to the scenario, whereas context-free items are mainly aimed at testing factual knowledge (5). The response format refers to how the

Acad Radiol 2017; ■:■■–■■

From the Department of Radiology, University Medical Center, Heidelberglaan 100, Utrecht 3584 CX, The Netherlands (D.R.R., A.v.d.G.); Radiological Society of the Netherlands, Vught 5260 CB, The Netherlands (D.R.R., F.v.R.); Department of Radiology, Gelre Hospitals, Apeldoorn 7334 DZ, The Netherlands (F.v.R.); Center for Image Sciences, University Medical Center, Utrecht 3584 CX, The Netherlands (C.M.); Center for Research and Development of Education, University Medical Center, Utrecht 3508 GA, The Netherlands (O.t.C.). Received August 10, 2017; revised October 8, 2017; accepted October 9, 2017. Funding Sources: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. **Address correspondence to:** D.R.R. e-mail: d.rutgers@umcutrecht.nl

https://doi.org/10.1016/j.acra.2017.10.014

candidate's answer is captured. Examples of response formats are true or false items, multiple-choice questions with a single best answer option, and long-list-menu items with a large number of predefined answer options (5–7). All answers of the candidate lead to a test score for the examination. According to classical test theory, the observed test score is the summation of the true score (the candidate's actual knowledge) and measurement errors (8). Sources of measurement errors are the test itself, the testee, and the tester (8). The amount of measurement error is estimated by calculating the reliability of the test. If measurement errors are reduced, the observed score of the candidate will approach the true score (9). Postexamination item analysis is recommended to investigate how test items perform in light of the objectives of the examination (10). In classical test theory, this analysis typically involves computing item difficulty and discrimination indices (10,11). Obviously, such psychometric characteristics have to be sound if the objective of the examination is to identify candidates who have not attained expected learning outcomes (12).

To quantify item difficulty, the *P* value is often used. This difficulty index reflects the proportion of candidates that has answered a given test item correctly. A value between .3 and .8 is generally considered as an acceptable level of difficulty (8,10). Values >.9 and <.2 may indicate that items are too easy or too difficult, but a range of *P* values in a long test is quite acceptable. Items that no-one or everyone answers correctly are not helpful in a test. The item-rest correlation (Rir) is an index that reflects the discriminating power of a test item, that is, the ability to discriminate between good and poor candidates. Rir values can vary from −1 to 1. High Rir values are found in test items that have been answered correctly by candidates who have performed well on the test as a whole, whereas they have been answered incorrectly by candidates performing weakly on the entire test. The opposite is found in test items with negative Rir values. These items have been answered incorrectly by high performers and correctly by weak performers. Generally, this should, of course, be avoided in an examination. A Rir value of zero reflects that the test item does not discriminate at all. Values >.2 indicate sufficient discriminating power (10), with values >.4 indicating excellent discriminatory performance.

Quality of Items with and without Images

Image-based test items can be used in examinations in various medical domains, such as anatomy, histology, radiology and pathology. The literature on psychometric properties of image-based test items in written examinations is limited and shows conflicting results. In anatomic and histologic examinations, some found no effect on item difficulty and discrimination (13), whereas others did find differences between image-based and text-alone items, although not in a consistent pattern (14,15). Studies on radiological written examinations are scarce. Radiological image-based test items comprise either volumetric images, that is, a series of cross-sectional slices of a body part through which the candidate can scroll, or two-dimensional (2D) nonvolumetric images, such as x-ray photos or a single cross-sectional slice. In a study of medical students, test items with volumetric images appeared to be more difficult than nonvolumetric items (16), but the discriminating power of individual test items was not explored and a comparison with text-alone items was not made. Previous study in radiology residents, using data from the same national examination as the current study, showed a steeper learning curve over time for image-based test items than for text-alone items (4). It has been suggested that this variation in learning curves may be related to differences in psychometric properties between image-based and text-alone items (17). However, this remains to be investigated as in these previous studies of postgraduate radiological examinations no psychometric characteristics of individual test items were reported.

Aim of the Study

The overall purpose of the present study was to explore the role of images in test item difficulty and discriminating power in postgraduate radiological written examinations. Our first aim was to investigate in image-based test items whether these psychometric indices were related to stimulus and response format characteristics. Our second aim was to study whether image-based test items differed from text-alone items with respect to difficulty and discriminating power.

METHODS

Dutch Radiology Progress Test

We obtained radiological test items from the Dutch Radiology Progress Test (DRPT). The DRPT is a semiannual formative knowledge test for radiology residents in the Netherlands. It is a required test for all residents during all training years. Radiology residency in the Netherlands does not include a separate certifying examination. Throughout their training program, radiology residents are formatively and summatively assessed in numerous workplace observations and written examinations such as the DRPT. The results of these assessments help the local program director to determine during residency whether a resident can continue in the training program, and to decide at the end of residency whether the resident has reached sufficient competence to graduate from the training program. This latter decision has to be reinforced by the national registration committee for medical specialists in order for the resident to register as a radiologist.

The content of the DRPT is drafted by the Examination Committee of the Radiological Society of the Netherlands. This committee is composed of at least nine radiologists from teaching and nonteaching hospitals throughout the Netherlands, who collectively cover all subspecialty domains of the DRPT: cardiothoracic radiology, neuro- and head-and-neck radiology, abdominal radiology, musculoskeletal radiology, pediatric radiology, breast radiology, interventional radiology,