Constructing Multiple-Choice Items to Measure Higher-Order Thinking

Darina Scully

Follow this and additional works at: https://scholarworks.umass.edu/pare

Recommended Citation
DOI: https://doi.org/10.7275/swgt-rj52
Available at: https://scholarworks.umass.edu/pare/vol22/iss1/4

This Article is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Practical Assessment, Research, and Evaluation by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Constructing Multiple-Choice Items to Measure Higher-Order Thinking

Darina Scully, Dublin City University

Across education, certification and licensure, there are repeated calls for the development of assessments that target higher-order thinking, as opposed to mere recall of facts. A common assumption is that this necessitates the use of constructed response or essay-style test questions; however, empirical evidence suggests that this may not be the case. In this paper, it is argued that multiple-choice items have the capacity to assess certain higher-order skills. In addition, a series of practical recommendations for test developers seeking to purposefully construct such items is provided.

The concept of 'higher-order thinking' is often linked to Bloom’s Taxonomy of Educational Objectives (Bloom, Englehart, Furst, Hill & Krathwohl, 1956); which set out six increasingly sophisticated cognitive processes in which a learner may engage. Revisions of and alternatives to Bloom’s taxonomy have been proposed over the years, but the underlying framework has remained a stable and important influence in education. There is now widespread recognition of the importance of invoking higher-order processes in both curriculum and assessment design (e.g., Lord & Baviskar, 2007; Momsen, Long, Wyse & Ebert-May, 2010).

This article provides an overview of what is meant by 'higher-order thinking', and outlines why it is valued in assessment, not only in K-12 and higher education contexts, but also in the field of professional certification and licensure. It discusses research that has investigated the capacity of multiple-choice (MC) items to assess higher-order thinking, and argues that this item format is not as restricted as once thought. Intended for both researchers and practitioners in the field of assessment and test development, this paper highlights the potential of MC items to assess higher-order thinking skills and compiles some practical suggestions for how such items may be constructed.

Higher-Order Thinking: Bloom’s Taxonomy & Related Frameworks

Higher-order thinking has typically been defined with specific reference to the cognitive domain of Bloom’s Taxonomy (outlined in Table 1), a trend that is still evident in contemporary research and discourse (e.g. Barnett & Francis, 2012; Jensen, McDaniel, Woodard & Kummer, 2014). The persistent influence of Bloom’s framework most likely stems from its intuitively appealing nature, and the fact that each level of cognitive sophistication, although designed to transcend specific subject matters and educational stages, can be interpreted and operationalized to suit individual contexts. The most basic level of the taxonomy is knowledge. Within any subject area, a learner can possess mere knowledge, and may demonstrate the ability to recall this learned knowledge in an assessment. They may not, however, understand the meaning of this knowledge. Furthermore, they may not possess the ability to apply it in situations other than that in which it was learnt, or to combine it with additional knowledge.
to create new insights. Such abilities are represented by subsequent levels of the taxonomy.

### Table 1. Bloom’s Taxonomy - Cognitive Domain

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>Recall or recognition of learned knowledge – without necessarily having the ability to apply this knowledge</td>
</tr>
<tr>
<td>Comprehension</td>
<td>Describing and explaining learned knowledge</td>
</tr>
<tr>
<td>Application</td>
<td>Using learned knowledge to solve problems in novel (but structurally similar) contexts</td>
</tr>
<tr>
<td>Analysis</td>
<td>Using learned knowledge to decompose situations into components, recognize unstated assumptions &amp; identify motives</td>
</tr>
<tr>
<td>Synthesis</td>
<td>Combining elements of learned knowledge into new integrated wholes</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Critiquing or judging the value or worth of learned knowledge</td>
</tr>
</tbody>
</table>

Although the taxonomy appears to assume a hierarchical structure; with the implication that processes such as knowledge and comprehension are prerequisites for processes such as application and analysis, Bloom et al. (1956) made no specific references to lower-order and higher-order thinking. Consequently, others’ interpretations of where lower-order thinking ‘ends’ and higher-order thinking ‘begins’ have been inconsistent. Hopson, Simms and Knezek (2001), for example, included only analysis, synthesis and evaluation in their definition of higher-order thinking, whilst Fives and DiDonato-Barnes (2013) made the cut-point between comprehension and application. Wiggins (2015), on the other hand, posited that only knowledge constitutes lower-order thinking. In the present paper, the taxonomy is conceived of as a continuum, with each level identifying a higher level of thinking than that which preceded it. As such, the term ‘higher-order thinking’ is understood to include all levels from comprehension onwards.

Since its inception, Bloom’s taxonomy has been formally revised (Anderson et al., 2001, see Table 2). The revised taxonomy substitutes the noun forms used to name the levels with equivalent verb forms, with the aim of drawing greater attention to the actions in which learners engage. Furthermore, it provides a range of additional verbs associated with each level, facilitating greater precision in classifying particular learning objectives at different levels. Anderson et al.’s revision also reverses the top two levels of the taxonomy, with create categorized as a higher level of thinking than evaluate. No empirical evidence was provided for this decision, however, and it has been argued (e.g. Huitt, 2011) that these two levels are best thought of as being equal in terms of complexity.

### Table 2. Anderson et al.’s (2001) Revision of Bloom’s Taxonomy

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Verbs Associated with Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remember</td>
<td>Retrieving relevant knowledge from long-term memory</td>
<td>Recognize, Recall</td>
</tr>
<tr>
<td>Understand</td>
<td>Determining the meaning of instructional messages, including oral, written &amp; graphic communication</td>
<td>Interpret, Exemplify, Classify, Summarize, Infer, Compare, Explain</td>
</tr>
<tr>
<td>Apply</td>
<td>Carrying out or using a procedure in a given situation</td>
<td>Execute, Implement</td>
</tr>
<tr>
<td>Analyse</td>
<td>Breaking material into its constituent parts, detecting how the parts relate to one another and to an overall structure or purpose</td>
<td>Differentiate, Organize,</td>
</tr>
<tr>
<td>Evaluate</td>
<td>Making judgements based on criteria and standards</td>
<td>Check, Critique</td>
</tr>
<tr>
<td>Create</td>
<td>Putting elements together to form a novel whole</td>
<td>Generate, Plan, Produce</td>
</tr>
</tbody>
</table>

An additional and important element of Anderson et al.’s (2001) taxonomy is its recognition of the fact that ‘knowledge’ itself is not a unitary concept. Drawing on concepts from the field of cognitive psychology that emerged in the latter half of the 20th century, the revised taxonomy differentiates between four types of knowledge: factual (knowledge of the basic elements of a
discipline such as terminology and specific details), conceptual (knowledge of the inter-relationships between these elements within larger structures), procedural (‘how-to’ knowledge) and metacognitive (awareness and knowledge of one’s own cognition). It follows that any cognitive process can interact with any type of knowledge. Krathwohl (2002) provided a helpful template to illustrate this concept, by plotting these different types of knowledge and the various cognitive levels on opposing axes of a two-dimensional table (Table 3). The cells formed by the intersections of these two dimensions give rise to a wide range of potential cognitive activities.

### Table 3. Krathwohl’s ‘Taxonomy Table’

<table>
<thead>
<tr>
<th>Cognitive Process Dimension</th>
<th>Knowledge Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Remember</td>
</tr>
<tr>
<td>Factual Knowledge</td>
<td></td>
</tr>
<tr>
<td>Conceptual Knowledge</td>
<td></td>
</tr>
<tr>
<td>Procedural Knowledge</td>
<td></td>
</tr>
<tr>
<td>Metacognitive Knowledge</td>
<td></td>
</tr>
</tbody>
</table>

Indeed, it has been shown that students who experience assessments demanding higher-order thinking are subsequently more likely to adopt meaningful, holistic approaches to their study, as opposed to engaging in mere surface-level or rote learning techniques (Jensen et al., 2014; Leung, Mok & Wong, 2008). In addition, such assessments allow instructors to provide more detailed and specific feedback (Momsen et al., 2010), which in turn can promote and guide future learning.

In higher education, there is a particularly strong interest in the assessment of higher-order skills, as universities and third-level institutions face growing demands to bridge the perceived gap between what students learn, and what is valued by employers. The need for ‘T-shaped professionals’ – i.e. university graduates equipped not only with disciplinary specialization (represented by the vertical stroke of the T), but also with ‘soft skills’ that allow them to operate effectively across a broad range of contexts (represented by the horizontal bar of the T) – is increasingly emphasized in both the academic literature and the mainstream media (e.g. Bitner & Brown, 2008; MacCraith, 2016; Oskam, 2009; Selingo, 2015). Examples of these ‘soft skills’ include creativity, collaborative problem-solving and critical thinking, all of which can be aligned with the upper levels of the various cognitive taxonomies.

In the field of certification and licensure, the primary objective of assessment is to reliably distinguish between candidates who do and do not possess the necessary knowledge, skills and abilities to practise a particular profession. Indeed, as LaDuca, Downing and Henzel (1995, p.138) asserted, the purpose of these types of assessment can be defined as the ‘protection of the public and the profession from unqualified practitioners’. With this in mind, issues such as the potential impact of these assessments on subsequent learning behaviours, or the value of generic, ‘transferrable’ skills may seem less relevant. This does not imply that the importance of
higher order thinking in this context is diminished, however. As Webb, Cizek and Kaloh (1993) pointed out, test items requiring higher-order thinking can improve the breadth and depth of content coverage within a licensure test. More importantly, the abilities to think strategically, to reflect, and to apply learned knowledge in a range of situations have been identified as key indicators of competency for a wide range of professions, including, but not limited to: nursing (Morrison & Free, 2001), medicine and the allied health professions (Choudhury, Gouldsborough & Shaw, 2015; Mann, 2008), accountancy (Hansen, 2006), and teaching (Struyven, Blieck & De Roeck, 2014). As such, it is vital that items measuring these skills are included to ensure the validity of pass/fail decisions arising from these tests.

**MC Items as an Assessment Format**

MC items (typically consisting of a stem and a choice of 3-5 response options) are an attractive assessment option for both educators and professional test developers for several reasons. Unlike constructed-response (CR) items such as short-answer or essay-style questions, they can be quickly administered and machine-scored, rendering them suitable for use with large groups of students or test candidates (Morrison & Free, 2001). In addition, they facilitate higher sampling of content per unit time (Schuwirth & Van der Vleuten, 2003), are associated with greater objectivity and reliability (Newstead & Dennis 1994, kniveton, 1996) and have even been shown to demonstrate superior concurrent validity with other measures of achievement (Bleks-Rechek, Zeug, & Webb, 2007). The use of multiple-choice items also provides the opportunity for test developers to quickly analyse the performance of each test item, and use this information to improve future assessments.

Despite these advantages, MC items have also received a great deal of criticism. Veloski, Rabinowitz, Robeson and Young (1999, p. 539), for example, condemned their use in the context of medical education, arguing that professional competence requires being able ‘to perform in a real-life setting that does not offer short lists of five choices.’ This reflects a general perception that MC items are incapable of assessing cognitive process beyond recall or recognition of knowledge, given that the correct answer is provided amongst the response options.

A number of empirical studies have attempted to investigate this issue, by comparing student performance across MC items and CR/performance items. Many have revealed that these item types measure two distinct constructs (e.g. Becker & Johnston, 1999; Frederiksen, 1984; Hickson & Reed, 2011; Krieg & Uyar, 2001), which is usually interpreted as a demonstration of the inferiority of MC items in assessing higher-order thinking. This interpretation, however, rests on the assumption that all CR items are a valid measure of higher-order thinking, which is not necessarily the case. Furthermore, other studies have failed to find differences in student performance across the two item types (e.g. Hickson, Reed & Sander, 2012; Thissen, Wainer & Wang, 1994; Walstad & Becker, 1994).

As Martinez (1999, p.207) argued, it is likely that any differences emerging between MC and CR items are ‘less a reflection of the limitations of these formats than they are of typical use’. Indeed, when subject-matter experts are given the task of classifying MC items according to the different levels of Bloom’s Taxonomy, the overwhelming majority are typically deemed to be recall or recognition items (e.g. Masters, Hulsmeyer, Pike, Leichy, Miller & Verst, 2001; Momsen et al., 2010; Tarrant, Knierim, Hayes & Ware, 2006; Webb et al., 1993). A small number of comprehension, application and analysis items have also been identified in these studies, however, which suggests that MC items do indeed have the potential to assess these skills, but that lower level MC items are simply over-represented. It follows that comparisons of different item types have been constrained by this fact; and that the potential of MC items to assess higher-order thinking may be chronically underestimated. Strong support for this assertion arises from studies that use Bloom’s Taxonomy to classify each item from both their MC and CR item banks at the offset. Indeed, Hancock (1994), Simkin and Kuechler (2005) and Traub (1993) followed this method, and observed that when subsequent performance comparisons were restricted to MC vs. CR items written at the same taxonomic level, moderate to strong correlations emerged.

At this point, it should be noted that there is currently insufficient evidence to suggest that MC items can measure a fully exhaustve range of thinking skills. Educational research distinguishes between convergent thinking, which refers to working with knowledge, concepts and processes that already exist; and divergent thinking, which is required in situations wherein there is no pre-determined solution. Given these definitions, it could be argued that the nature of the MC item format necessarily precludes its ability to assess the two highest
levels of the taxonomy. Indeed, the research discussed to this point supports the contention that the capabilities of MC items extend as far as the analyse level, but there have been few instances of evaluate or create MC items identified to date. A glance at some of the verbs suggested by Anderson et al. to describe actions reflecting these levels (refer back to Table 2) highlights the potential difficulty to this end. Despite this, some strategies may be employed to support ‘pseudo-assessment’ of these higher levels using MC items. These will be discussed in the following section.

MC items, like all forms of assessment, are associated with certain limitations. It is fully acknowledged that the authenticity of an assessment can – and where possible, should - be maximized by encompassing various methods of measurement. The aim of this paper is not to advocate for the exclusive use of one particular method; rather, it is to draw attention to the possibility that the capacity of MC items to assess higher order thinking has been masked by studies that treat this item format as a homogenous entity.

Constructing MC items assessing higher-order thinking is undoubtedly a challenging and time-consuming task. Yet, it is possible; moreover, it is worthwhile investing the time and resources in doing so. Cognitively challenging MC items offer an attractive balance, as they have the potential to simultaneously meet the needs of (i) students seeking to improve their learning through the medium of assessment, (ii) educators who wish to obtain meaningful information about their students’ abilities, (iii) universities seeking to measure skills valued by employers, and (iv) certification/licensure test developers aiming to improve the fidelity of their test items and the validity of their decision-making in a cost-effective manner.

### Strategies for Constructing MC items Assessing Higher-Order Thinking

Targeting an item at a particular cognitive level requires, above all, explicit reference to a well-established taxonomic framework such as that of Anderson et al. (2001). For example, items can be written specifically to assess a student’s ability to ‘remember procedural knowledge’, to ‘apply factual knowledge’, etc. (refer back to Table 3). Whilst this is undoubtedly a helpful starting point, it is not enough to refer solely to process descriptions, as they are often quite broad in nature and open to subjective interpretations (Hancock, 1994). Rather, careful consideration should be given to the precise definitional criteria set out both for the cognitive level and type of knowledge in question, and these criteria mapped closely to the structure of the item.

Furthermore, although the taxonomy was initially designed to transcend specific subject matters, measurement experts argue that the more sophisticated cognitive processes are inherently domain-specific. Indeed, Anderson et al. (2001) acknowledged that, ideally, each major field should have its own taxonomy, ‘closer to the special language and thinking of its experts, and reflecting its own appropriate sub-divisions’. All subject-matter experts are faced with this challenge of operationalizing general taxonomic levels for their specific area (Morrison & Freie, 2001). As such, the extent to which generic rules for item construction at various cognitive levels can be generated is somewhat limited. Nevertheless, some strategies have been identified that may help guide the production of items that reach beyond mere recall.

1. **Manipulation of Target Verbs** Specific verbs have been linked to the various cognitive processes (Morrison & Freie, 2001; Table 4). These verbs, when

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Comprehension</th>
<th>Application</th>
<th>Analysis</th>
<th>Synthesis</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify</td>
<td>Describe</td>
<td>Apply</td>
<td>Analyse</td>
<td>Compose</td>
<td>Appraise</td>
</tr>
<tr>
<td>Define</td>
<td>Differentiate</td>
<td>Calculate</td>
<td>Categorize</td>
<td>Construct</td>
<td>Assess</td>
</tr>
<tr>
<td>Know</td>
<td>Discuss</td>
<td>Classify</td>
<td>Compare</td>
<td>Create</td>
<td>Evaluate</td>
</tr>
<tr>
<td>List</td>
<td>Explain</td>
<td>Develop</td>
<td>Contrast</td>
<td>Design</td>
<td>Judge</td>
</tr>
<tr>
<td>Name</td>
<td>Rephrase</td>
<td>Examine</td>
<td>Distinguish</td>
<td>Formulate</td>
<td></td>
</tr>
<tr>
<td>Recognize</td>
<td>Restate</td>
<td>Solve</td>
<td>Determine</td>
<td>Modify</td>
<td></td>
</tr>
<tr>
<td>State</td>
<td>Reword</td>
<td>Use</td>
<td>Investigate</td>
<td>Plan</td>
<td></td>
</tr>
</tbody>
</table>

Published by ScholarWorks@UMass Amherst, 2017
placed in an item, may serve as rudimental indicators of the cognitive level it is likely to assess. This strategy should be used with caution, as some verbs could arguably be placed in multiple categories, and much depends on the context of the item in which they are placed. They may, however, provide an objective, transparent basis for item-writers.

At a first glance, many of these verbs may appear to be incompatible with MC items - even those associated with relatively low cognitive levels such as ‘describe’ and ‘explain’. This may explain why there is usually an abundance of knowledge level items. However, as Dickinson (2011) pointed out, MC items assessing higher levels can be constructed by replacing the desired ‘unconstrained’ verb with its noun derivative; and preceding it with a knowledge level verb, resulting in stems such as ‘select the best description’ or ‘identify the most accurate explanation’. This strategy could theoretically be extended to the synthesis and evaluation levels (e.g. ‘select the best plan’/ ‘identify the best modification’). However, as students are not required to independently generate the solutions in such scenarios, this is best thought of as ‘pseudo-assessment’ of the highest cognitive levels.

(ii) Item Flipping  Items that present an overarching concept or category, and require examinees to recognize a specific instance of this concept, can usually be classified at the lowest taxonomic level. Examinees can successfully answer these items without necessarily having an understanding of the concept, or what it means to be an exemplar of that concept, by simply drawing on a memorized list of terms. Dickinson (2011) suggests that such items can be ‘flipped’, by presenting the specific instance in the item stem, and asking the examinee to identify the underlying rule or concept instead. To correctly answer flipped items, test-takers must also have a complete understanding of the alternative distractor concepts, and consider whether or not the characteristic presented in the stem could fit in with any of these; thus the item is moved from the knowledge to the comprehension level. Examples from the fields of education and psychology are provided in Table 5.

<table>
<thead>
<tr>
<th>Original Items:</th>
<th>Flipped Items:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. is based on the student’s attitudes, interests and values</td>
<td>A. affective assessment</td>
</tr>
<tr>
<td>B. is designed primarily to evaluate learning</td>
<td>B. formative assessment*</td>
</tr>
<tr>
<td>C. is usually high-stakes</td>
<td>C. diagnostic assessment</td>
</tr>
<tr>
<td>D. provides information to modify teaching and learning*</td>
<td>D. summative assessment</td>
</tr>
</tbody>
</table>

(Source: O’Leary, personal communication, May 1, 2017)

According to Piaget’s theory of cognitive development, what is ‘accommodation’?

A. the ability to think logically
B. the diminishing of a response to a frequently repeated stimulus
C. altering one’s existing schemas as a result of new information* | A. Accommodation* |
D. an inability to understand perspectives besides one’s own | B. Conservation |
C. Habitation |
D. Egocentrism | (Adapted from: ProProfs, n.d.)

(iii) Use of High Quality Distractors  Regardless of how an item is constructed, if one or more of its distractors are implausible to even the weakest students, it will not assess higher level thinking (Hancock, 1994). Distractors that are superficially similar to the item key, on the other hand, demand a high level of discriminating judgement. It has thus been recommended that, where possible, all of the given options are theoretically

https://scholarworks.umass.edu/pare/vol22/iss1/4
DOI: https://doi.org/10.7275/swgt-rj52
plausible, with the key being the ‘best’ answer, as opposed to the only correct option. The item stem should also be worded appropriately to reflect this. Great care must be taken to ensure that this strategy does not introduce subjectivity - the item key must remain indisputably and objectively ‘more correct’ than any of the distractors.

Table 6 provides an example of the effectiveness of high quality distractors, using two items from the field of nursing (Morrison & Free, 2001). Although both of these items assess similar content, the item on the right demands a higher level of cognitive processing. To correctly answer this item, an examinee must know that a shuffling gait is characteristic of Parkinson’s disease, and use this knowledge to understand that the presence of throw rugs throughout the home pose a significant safety hazard to the client, and thus has the ‘greatest’ implication for his care. All three of the distractors are plausible, as they may also be construed as having implications for the client’s care. For example, options C and D are indicative of additional symptoms of Parkinson’s disease that may require attention, whilst option A could potentially, but not necessarily, have implications for care, depending on how the client feels about visits from his grandchildren.

(iv) Tapping ‘Multiple Neurons’ Burns (2010, p.332) distinguished between ‘one-neuron’ items, whereby, ‘figuratively, the student only has to fire one neuron to obtain the memorized, tidbit answer’, and ‘multiple-neuron’ items, which require an understanding of ‘interconnections between knowledge’. Practically speaking, multiple-neuron items assess higher-level processes such as knowledge application, because they require examinees to have knowledge of more than one fact or concept, and to successfully combine these to arrive at the correct answer. Table 7 tracks the transformation of an item stem from a ‘one-neuron’ to a ‘five-neuron’ classification.

As is evident from Table 7, the process of literally transforming an item from one- to five-neuron status in this way may result in rather cumbersome, poorly-worded items. It is important to strike a balance between achieving the desired level of cognitive complexity, whilst simultaneously maintaining basic principles for good item-writing, such as clarity of wording (see Haladyna, Downing, & Rodriguez, 2002 for a comprehensive overview of these). Furthermore, some content areas simply may not lend themselves to the use of five-neuron questions. An appropriate rule-of-thumb may be to strive for items that could be classified at least at the ‘two-neuron’ level or, as Morrison and Free (2001)
suggested; items for which the answer could not theoretically be located on one page of a textbook.

Table 8 provides some examples of single- and multiple-neuron items from the fields of pharmacy, statistics, and education respectively. It should be noted that multiple-neuron items are often (but not always) context-dependent. That is, they may present a stimulus or scenario, requiring the examinee to draw on various elements of their subject knowledge to interpret the scenario and subsequently select the most appropriate response. In some cases, several items may accompany a given scenario. These are known as context-dependent item sets.

One potential disadvantage of context-dependent items is their inherently greater reading load (Airasian et al., 1994). Theoretically, this can introduce construct irrelevant variance and disadvantage test-takers with low verbal ability or poor English proficiency. Consideration of test-taker characteristics and differential item functioning analyses may help monitor whether or not this is an issue in particular testing contexts. Where concerns are raised, the use of video (Chan & Schmitt, 1997) or animation (Dancy & Beichner, 2006) to present the information contained in the stem may be an effective solution.

Table 8. Examples of ‘multiple-neuron’ items

Which of the following is a contraindication for spironolactone?
- A. Serum creatinine = 3.0mg/dL*
- B. Serum potassium = 3.5mEq/L
- C. Resting heart rate = 68bpm
- D. Blood pressure = 130/85mmHg

(Source: Tiemeier, Stacy & Burke, 2011)

You have carried out a 3 x 2 ANOVA for independent groups. There were 60 participants with 10 participants randomly assigned to each cell. You have now analysed the data and are checking your work. Which of the following would immediately let you know that you have made an error?
- A. You found the total degrees of freedom to be 60.*
- B. You found the mean square for the error term to be 6.25
- C. You found the F-statistic for the interaction effect to be 2.34
- D. You found degrees of freedom for the interaction effect to be 2.

(Source: DiBattista, 2011)

James is a fourth class student. His results from a standardized reading assessment are below:

<table>
<thead>
<tr>
<th>Test: Reading</th>
<th>Standard Score: 81</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level: 4</td>
<td>STEN Score: 2</td>
</tr>
<tr>
<td>Form: A</td>
<td>Percentile Rank: 12</td>
</tr>
</tbody>
</table>

James’ teacher, Mrs Smith is preparing to explain the test results to James’ parents. Which of the following represents a correct interpretation of the results?
- A) James did as well or better than 12% of students in his class
- B) James did as well or better than 12% of 4th class students nationally*
- C) James did better than 81% of 4th class students nationally
- D) James did better than 81% of students in his class

What additional information would be most important for Mrs. Smith to communicate to James’ parents to help them fully understand the meaning of these results?
- A) James’ raw score on the standardized test
- B) The mean standard score for the class
- C) James’ performance in everyday reading activities*
- D) James’ standardized scores from last year
Each of the strategies described to this point may aid item-writers to construct MC items at higher cognitive levels. However, it should be appreciated that none will automatically produce items aligned to one particular level. Constructing analysis level items, for example, is especially challenging, and requires an additional layer of abstraction. Indeed, as Oermann and Gaberson (2009) pointed out, differences between application and analysis items are not always readily apparent. Analysis has been described in various terms, such as ‘breaking material into its constituent parts’ (Anderson et al., 2001) and ‘recognizing unstated assumptions’ (Bloom et al., 1956). Items that assess analysis are often accompanied by complex stimuli that must be interpreted; alternatively, they may require examinees to digest and make sense of multiple pieces of information with respect to each response option in order to determine which is the most appropriate. Generally speaking, these items require high-level problem-solving skills such as interpreting abstract information, recognizing discrepancies, scrutinizing decisions, or inferring causality in complex situations (Lord & Baviskar, 2007; Simkin & Kuechler, 2005; Vacc, Loesch, & Lubik, 2001). Of course, it is necessary to tailor these criteria to the specific subject matter, and also to the level of the exam. Two examples of analysis items from a nursing certification exam and a primary school science test, respectively, are provided in Table 9.

### Validity Considerations

Following attempts to construct MC items that measure higher-order thinking, the next logical step is to investigate whether or not these items succeed in doing so. This is essential to establish support for the validity of any subsequent judgements based on the assessment. However, it can present some challenges. Ideally, the

---

**Table 9. Examples of MC items that could be classified at the ‘analysis’ level**

You receive a report on the following patients at the beginning of your evening shift. Which patient should you assess first?

A. An 82-year-old with pneumonia who seems confused at time*
B. A 76-year-old patient with cancer with 300 mL remaining of an intravenous infusion
C. A 40-year-old who had an emergency appendectomy 8 hours ago
D. An 18-year-old with chest tubes for treatment of an pneumothorax following an accident

*(Source: Oermann & Gaberson, 2009)*

Linda and Steve did a survey of the fruit that children in their class liked best. Look at the chart and answer the two questions.

1. Oranges are more popular than grapes
   - A. True
   - B. False*
   - C. Can’t say

2. Children eat bananas most often
   - A. True
   - B. False
   - C. Can’t say*

*(Adapted from Kilfeather, O’Leary & Varley, 2011)*
The cognitive complexity of an item should be rated by subject-matter experts, trained explicitly in the use of one or more of the relevant taxonomies (e.g. Tarrant et al., 2006). A drawback to this method is that there is necessarily an element of subjectivity associated with these ratings. Accordingly, it is advisable that items are classified by diverse groups of experts, who have been instructed to focus closely on the definitional criteria outlined in the frameworks, and to consider potential discrepancies that could arise from these as they are classifying the items. Measures of inter-rater reliability should then be considered in determining whether an item can be eventually classified at any given level, or whether it should be revised/removed.

For those favouring more objective criteria, it may seem appealing to refer to the items’ difficulty indices. However, this is not advisable. Difficulty and complexity are two distinct attributes, with the former simply referring to the proportion of test-takers who answer an item correctly. As Hancock (1994) pointed out, a test item may be difficult on the basis that it requires a test-taker to recall a relatively obscure fact. This fact, however, may be trivial with regard to a learner’s overall level of understanding of a complex concept, or a licensure candidate’s competence in the given profession. Empirical evidence has supported this view, with weak – and occasionally inverse correlations emerging between complexity ratings and difficulty indices (e.g. Hancock, 1994; Schneider, Huff, Egan, Gaines, & Ferrara, 2013).

One particularly promising method that may be suitable for identifying the cognitive complexity of a test item is the think aloud protocol (TAP; Ericsson 2006), whereby individuals are asked to verbalize their thoughts whilst engaged in a learning activity. Evidence suggests that TAP may represent an accurate measure of both cognitive and metacognitive processes (Azevedo, Moos, Johnson & Chauncey, 2010); as such, if TAP was employed to measure the cognitive processes of a test-taker whilst completing a given item, this could potentially give valuable insights into the level of thinking assessed by this item.

Concluding Comments

The importance of measuring higher-order thinking, for a variety of reasons, is well recognized in both educational and professional assessment circles. Many have argued that MC items – although valued for their objective and cost-efficient nature – are incapable of measuring complex cognitive processes. A more accurate assertion, however, may be that MC items measuring complex cognitive processes are simply rarely constructed. That is, the format itself is not necessarily restricted to the assessment of superficial recall and recognition. By adhering to certain strategies, it is possible to construct MC items measuring processes such as knowledge application and analysis. This can benefit both learners and test developers in a variety of contexts.

References


Bloom, B., Englehart, M., Furst, E., Hill, W. & Krathwohl, D. (1956). A taxonomy of educational objectives,


Scully, Constructing Multiple-Choice Items to Measure Higher-Order Thinking


Tarrant, M., Knierim, A., Hayes, S., Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Education in Practice, 6, 354-363


Webb, N.L. (2005). Depth of Knowledge levels for four content areas. Presentation to the Florida Education
Author Notes

The work of the Centre for Assessment Research, Policy and Practice in Education at Dublin City University is supported by Prometric. The author would like to thank the test development team at Prometric for stimulating her interest in this topic. Acknowledgements are also due to Anastasios Karakolidis, Michael O’Leary and Linda Waters for their helpful comments on drafts of the manuscript. The views expressed in this article are those of the author and do not necessarily represent the views of Prometric.

Citation:


Corresponding Author

Dr. Darina Scully
Centre for Assessment Research, Policy & Practice in Education (CARPE)
Institute of Education, St. Patrick’s Campus
Dublin City University
Dublin 9
Republic of Ireland

email: darina.scully [at] dcu.ie