On varying the difficulty of test items

LEONG See Cheng
Singapore Examinations and Assessment Board, Singapore

A paper presented at the 32nd Annual Conference of the
International Association for Educational Assessment, Singapore, May 2006.

Abstract

This paper discusses the practice of varying the difficulty of test items in educational measurement. An item difficulty framework comprising concepts such as content difficulty, stimulus difficulty, task difficulty and expected response difficulty is introduced. The paper concludes with some suggestions on how to vary the difficulty of test items.

The views and opinions expressed in this paper are those of the author and are not to be taken as official policy and practice of the Singapore Examinations and Assessment Board.

Contact details:
Address: 298, Jalan Bukit Ho Swee, Singapore 169565

Email address: Leong_See_Cheng@seab.gov.sg

Tel: 65-63777723
Fax: 65-62729439

On varying the difficulty of test items

Someone by the name of Stenner once said, “If you don’t know why this question is harder than that one, then you don’t know what you are measuring.” (cited in Fisher-Hoch & Hughes, 1996). This statement puts into focus the role of item difficulty in educational measurement. While it is very often in testing agencies worldwide that item writers are reminded to write test items to measure the construct that they are measuring, it is less often that item writers are advised to think about the difficulty of items in relation to the construct that they are measuring.

There is a host of construct validation procedures (see Sireci, 1998) to aid item writers in ensuring that test items measure the construct they are intended to measure; but there are only a few documents (e.g., Pollitt, Hutchinson, Entwistle, & De Luca, 1985; Fisher-Hoch, Hughes, & Bramley, 1997; Ahmed & Pollitt, 1999) on how to vary the difficulty of test items that item writers may refer to. This paper aims to add to the literature on how the difficulty of test items may be varied and to generate discussion among practitioners on the appropriate practices in controlling the difficulty of test items.

The need to control difficulty in an item

Besides contributing to the measurement of the construct that item writers want to measure, there are other rationales for controlling the difficulty of items. First, in some achievement testing circumstances, there is a need to spread candidates over a wide range of marks. Test items of a wide range of difficulty levels are needed to test the entire range of candidates’ achievement levels. Tests that contain too many easy or too many difficult test items of would result in skewed mark distributions. Second, in situations where there is a need to construct parallel tests (e.g., to maintain the rigour and standards of assessment from year to year), the ability to vary the difficulty of test items is crucial. The distribution of item difficulty levels in one year should be comparable to the distribution of item difficulty levels in another, among other considerations. Third, in test development, the pilot-testing of test items of unsuitable difficulty levels is a waste of time and effort. Test items must be set at suitable difficulty levels so that the results of pilot-tests can be used to confirm their difficulty level. Fourth, in assessments where choices from optional items are offered to candidates, there is a responsibility for item writers to ensure that the items are of comparable difficulty. It is only when the optional items are of comparable difficulty that the test results may be reliable.

Locations of difficulty in a test item

Ahmed and Pollitt (1999) have suggested that the difficulty of a test item is in the question-answering process. In their paper, they list “sources of difficulty” in the five stages of the question-answering process (namely, learning, reading the question, searching the subject knowledge, matching the question and subject models, generating the answer, and writing the answer). Is there another way of thinking about the locations of difficulty in a test item? In other words, is there a way of thinking about difficulty that does not require a psychological understanding of the question-answering process? We can begin with the definition of a test item in Osterlind (1990).

> “A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form of answering; and is intended to yield a response from an examinee from which performance in some psychological construct (such as knowledge, ability, predisposition, or trait) may be inferred.”

An analysis of Osterlind’s definition of a test item suggests there are four locations in an item where difficulty may reside. These are: (1) content assessed; (2) stimulus; (3) task to be performed; (4) expected response. I shall refer to the difficulty in the four locations as content difficulty, stimulus difficulty, task difficulty and expected response difficulty.

Content difficulty refers to the difficulty in the subject matter assessed. In the assessment of knowledge, the difficulty of a test item resides in the various elements of knowledge such as facts, concepts, principles and procedures. These knowledge elements may be basic, appropriate or advanced. Basic knowledge elements are those in which candidates have learnt at lower levels. They are very likely to be familiar to candidates because they would have the opportunity to learn them well, and they are not likely to pose difficulty to many candidates. Advanced knowledge elements are usually those that will be covered more adequately at advanced levels and hence are peripheral to the core curriculum, and candidates may not have sufficient opportunity to learn. These knowledge elements are likely to be difficult for most of the candidates. Knowledge elements at the appropriate level are those that are central to the core curriculum. Depending on the level of preparedness of the candidates, these knowledge elements may be easy or difficult to candidates; overall, items that test knowledge elements at the appropriate level may be moderately difficult to candidates. Content difficulty may also be varied by changing the number of knowledge elements assessed. Generally, the difficulty of an item increases with the number of knowledge elements assessed. Test items that assess candidates on two or more knowledge elements are generally more difficult than test items that assess candidates on a single knowledge element. The difficulty of a test item may be further increased by assessing candidates on a combination of knowledge elements that are seldom combined (Ahmed, Pollitt, Crisp, & Sweiry, 2003).

Stimulus difficulty refers to the difficulty that candidates face when they attempt to comprehend the words and phrases in a test item and the information that accompanies the item (e.g., diagrams, tables and graphs). Test items that contain words and phrases that require only simple and straightforward comprehension are usually easier than those that require careful or technical comprehension. The manner in which information is packed in a test item also affects the difficulty level of the test item. Test items that contain information that is tailored to an expected response (i.e., no irrelevant information in the test item) are generally easier than test items that require candidates to select relevant information or unpack a large amount of information.

Task difficulty refers to the difficulty that candidates face when they generate a response or formulate an answer. In most test items, to generate a response, candidates have to work through the steps of a solution. Generally, test items that require more steps in a solution are more difficult to than test items that require fewer steps. In addition, the task difficulty of a test item may be mediated by the amount of guidance present. Test items that contain guided steps are generally easier than those that require candidates to devise the steps. The task difficulty of a test item may also be affected by the order of thinking or cognitive processing required. Taxonomies of cognitive processes, in particular the Bloom's Taxonomy, have suggested that cognitive processes exist in a cumulative hierarchy (i.e., the more complex processes include the simpler processes). Thus test items that assess candidates on higher order processes (e.g., analysis and synthesis) may generally be more difficult than test items that assess candidates on lower order processes (e.g., recall and comprehension). Similarly, in the assessment of skills, test items that assess candidates in higher order skills such as application and improvisation are generally more difficult than test items that assess candidates in lower order skills such as imitation and patterning.

Expected response difficulty refers to the difficulty imposed by examiners in a mark scheme or scoring rubrics. This location of difficulty in a test item is applicable only to constructed-response items; it is not applicable to selected-response (e.g., multiple-choice, true-false and matching). When examiners expect few or no details in a response to a test item, the test item is generally easier than a test item in which examiners expect a lot of details. Another aspect of expected response difficulty is the complexity in structure of an expected response. When simple connections among ideas are expected in a response, the test item is generally easier than a test item in which the significance of the relations between the parts and the whole is expected to be discussed in a response. In other words, a test item in which a unistructural response is expected is generally easier than a test item in which relational response is expected. A third aspect of expected response difficulty is in the clarity of marks allocation. Test items in which the allocation of marks is straight-forward or logical (e.g., 3 marks for listing 3 points) are generally easier than test items in which the mark allocation is unclear (e.g., 20 marks for a discussion of a concept, without any hint of how much and what to write in a response). This aspect of expected response difficulty affects the difficulty of an item because candidates who are unclear about the demand in a response may not produce sufficient amount of answers in a response that will earn the marks that befit their ability.

A similar item difficulty framework has been proposed in 1985. Pollit, et al. (1985) suggested three general categories of difficulty: subject or concept difficulty; process difficulty and question (stimulus) difficulty.

Valid and invalid moderators of difficulty

In the foregoing discussion of the four locations of difficulty in a test item, the various aspects of difficulty may be termed as moderators of difficulty (after Ahmed, et al., 2003). There are valid and invalid moderators of difficulty. Valid moderators of difficulty are those that contribute to the measurement of the construct under consideration. Conversely, invalid moderators of difficulty are those that impede or confound the measurement of the construct. Invalid moderators of difficulty prevent examiners from achieving the goal of assessing what s/he wants to assess and candidates doing what examiners wants them to do. Invalid moderators of difficulty also hinder candidates from showing their true ability or competence. Table 1 presents a list of probable invalid moderators of difficulty.

Table 1
<u>Probable invalid moderators of difficulty</u>

Content
Testing of obscure concepts or facts (e.g., facts that are hardly mentioned in major textbooks)
Testing of unimportant facts that are not central to learning outcomes and objectives
Testing of advanced concepts which candidates have little opportunity to learn
Stimulus
Inaccuracy or inconsistency in data or information given
Insufficient information
Meaning of words unknown or unclear
Question asked is not the one that examiners want candidates to answer
Grammatical errors in the question that can cause misunderstanding
Unclear resources (e.g., badly drawn / printed diagram, inappropriate graph, unconventional table)
Dense presentation (too many important points packed in a certain part of the stimulus)
Demand on reading comprehension when reading comprehension is irrelevant to the construct measured
Task
Illogical order of parts of the items
Mark allocation is unclear or illogical
Level of detail required in an answer is unclear
Context is unrelated / unnatural to the task that candidates have to do
Details of a context distract candidates from recalling the right bits of knowledge
Interference from a previous question
Insufficient space or insufficient time allocated for responding
Expected Response
Mark scheme and questions are incongruent – mark scheme spells out expectations to a slightly different question, not the actual question
Answer is indeterminable
Large number of plausible alternative answers
Rigid mark scheme
Demand on producing a written answer when producing a written answer is not part of the construct measured

Suggestions to moderate difficulty

How then should item writers increase or decrease the difficulty of items in ways that do not impede the measurement of the construct under consideration? Are there ways in which the difficulty of a test item can be varied without hindering candidates from showing what examiners expect to see? Table 2 presents some suggestions.

Table 2
Suggestions on how to increase or decrease the difficulty in test items

Location	Demands that increase difficulty	Supports that decrease difficulty
Content	Test knowledge in the curriculum that requires deep learning and understanding Test two concepts / topics that are rarely combined	Connect knowledge tested with basic level knowledge or knowledge learnt at lower level Reduce the number of concepts / topics tested
Stimulus	Use relevant technical terms, without elaboration or clarification, in the item; Remove references to the concept tested in the item; Use novel or foreign contexts that are appropriate Pack more information than needed (if it is appropriate to test selection of information); Present information in such a way that requires candidate to do some re-organisation	Highlight or emphasize terms that require careful comprehension; State the topic or concept tested in the form of a heading to help candidates recall or focus at the right bits of knowledge; Use contexts that are closely related to the task that candidates have to do Improve the physical layout of the item; Replace words that may mislead some candidates; Provide a glossary of command words or replace command words with a simply and clearly stated demand; Remove irrelevant or redundant information / words in the item; Tailor the resources to the task that candidates have to do.
Task	Increase the number of steps in executing task; Without cues and leaders, present task that requires candidates to devise steps to execute the task Require candidates to use process skills in uncommon ways	Decrease the number of steps in executing task; Break up the task into a few steps (sub-questions); Order the steps such that they provide the scaffolding for subsequent steps Test lower level process skill as a precursor to the testing of higher level process skill

Conclusion

The control of item difficulty in test items is currently not an exact science. The item difficulty framework that is presented in this paper is merely an attempt at creating a conceptual framework to think about item difficulty. It is not an explanatory theory yet because it still cannot explain why certain low-order test items that assess candidates on specific knowledge (e.g., “How many bones are there in the inner ear of the human body?”), among others. can be more difficult than a multi-step, multiple-concept science question. Further, the item difficulty framework in this paper does not state the relationships and interactions among the concepts of the framework. More work has to be done to understand item difficulty at a deeper level.

Although we are still far away from understanding item difficulty completely, it is important to keep in mind that the skill in varying the item difficulty of a test items contributes to the construct validity of an educational assessment. The skilfulness of an item writer is not only in assuring that an item measures what it is supposed to measure, it is also in varying the difficulty at will to measure the entire range of what s/he is supposed to measure.

References

Ahmed, A., & Pollitt, A. (1999). *Curriculum demands and question difficulty.* Paper presented at IAEA Conference, Slovenia.

Ahmed, A, Pollitt, A., Crisp, V., & Sweiry, E. (2003). *Writing examinations questions*. A course created by the Research & Evaluation Division, University of Cambridge Local Examinations Syndicate.

Fisher-Hoch, H., & Hughes, S. (1996). What *makes mathematics exam questions difficult?* Paper presented at the British Educational Research Association Annual Conference, Lancaster University.

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997). *What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions*. Paper presented at the British Educational Research Annual Conference, University of York.

Pollitt, A., Huchinson, C., Entwistle, N., & De Luca, C. (1985). *What makes exam questions difficult?* Edinburgh, UK: Scottish Academic Press.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117.