IAEA conference, 2010
Bangkok, Thailand

A three-way classification of sources of item difficulty in tests and examinations

Gabrielle Matters
Australian Council for Educational Research

A three-way classification of sources of item difficulty

What do test t*akers* mean when they say 'this item is difficult'? What do test *analysts* mean when they say 'this item is difficult'? The answer to the first question comes out of experience. The answer to the second question comes out of empirics. The notion of difficulty covers a considerable diversity of sources, materials and methods. Test analysts seem obliged to collapse all senses of difficulty under one heading and so it might be useful to attempt a classification or typology of some of the possible sources of difficulty in test items.

This presentation describes such a system and applies it to test items in multiple-choice and short-response items, in tests that are discipline-specific and in tests of generic skills.

The interest is for policy makers who are inclined to fixate on test data without ever querying the properties of the assessment instruments from which the data were generated, and for educators who have always been interested in the questions of the item–person interaction.

The classification system gives difficulty some meaning beyond measurement.

Proposition

The following three sources, alone or in combination, may explain item difficulty:

1. Nature of the cognitive task (*intrinsic difficulty* – a function of the particular cognitive skill(s) required)
2. Candidate perception of the difficulty of the task (*self-imposed difficulty* – a function of the particular candidate's mindset on viewing the stimulus material)
3. Aspects of test design (*design-imposed difficulty* – originates in features of the item that occur as a result of the test developers fulfilment of one or more of the design criteria, such as the required mode of response at the item level or, at the test level, "turbulence" as students move from one task to another).

Focus of this conference session

In this conference session we focus on intrinsic difficulty and self-imposed difficulty.

Intrinsic Difficulty

The first type of difficulty we examine is intrinsic difficulty. Items that test cognitive skills are different in three fundamental ways, which can be regarded as dimensions:

i. They involve different kinds of thinking: concrete, conceptual or personal ─ the *interactive dimension.*

ii. They involve different mental abilities: verbal, numerical or spatial ─ the *reasoning*[1] *dimension.*

[1] It is acknowledged that the word *reasoning* lends itself to a purer definition, but for the purposes of this typology it remains a convenient label possessing the restricted meaning given here.

iii. They place different emphases on the treatment of the stimulus material: it may need to be absorbed, operated on, or transformed into something new ─ the *treatment dimension*.

Each of the three dimensions of intrinsic difficulty can in turn be viewed as having factors that affect a candidate's success on the test item; these factors can be viewed as *positions* on a *dimension*. An 'item characteristic indicator' (ICI) can conveniently summarise item characteristics on the three dimensions. As there are three dimensions in the model, each having three categories or positions, there are 27 combinations of item characteristics.

Self-imposed difficulty

Self-imposed difficulty is a function of a particular student's mindset on viewing the stimulus material, and may be influenced by features such as content and context. The notion of self-efficacy is important here (Bandura, 1982). Matters & Burnett's (2003) investigation of self-efficacy in relation to the propensity to omit items in tests has relevance to wider considerations of difficulty in tests:

> [Self-efficacy] is a cognitive mechanism consisting of beliefs concerning one's capacity to perform tasks successfully. These expectations are hypothesised to affect the initiation of coping behaviour, the expenditure of effort, performance accomplishment, and persistence in overcoming obstacles. Because self-efficacy relates to mastery and persistence, which, in turn, are precursors to academic success, self-efficacy could provide a possible answer to the question of who omits test items.
>
> (Matters & Burnett, 2003: 241–242)

For the QCS Test, which tests student performance on 49 generic skills that are the threads of the senior curriculum, and for PISA, with its varied clientele and dedication to real-life contexts, self-imposed difficulty might be significant. A context may be novel to the students – not only unfamiliar in real life but also unrehearsed in an assessment situation.

Application of typology

This typology was successfully applied to all 100 multiple-choice items and 29 short-response items on the 1993 Queensland Core Skills Test (QCS Test) (Queensland Board of Studies, 1993) and to 75 non-secure PISA items designed to test scientific literacy (OECD, 2006), and then matching narratives were composed for the items that proved difficult for students who sat the test.

Attachments

Items that will be used to stimulate discussion are attached. The correct response is not given in this paper but in the conference session.

A list of words that students use as synonyms for difficulty is also attached.

References

Bandura, A. (1992). Self-efficacy mechanism in human agency. American Psychologist, 37(2), 122–147.

Matters, G.N. (1999). The QCS Test Companion: Discover your test-taking type and take control of your learning. Sydney: McGraw-Hill.

Matters, G. N., & Kelly, D. J. (2009). "Context as a source of item difficulty". Paper presented at the 35th annual conference of the International Association for Educational Assessment, Brisbane, September 2009.

Matters, G. N., & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement*, *63*(2), 239–256.

OECD (2006). Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006. http://www.pisa.oecd.org/dataoecd/63/35/37464175.pdf

Queensland Board of Studies. (1993). *The Queensland Core Skills Test.*

Queensland Board of Studies. (1994). *The Queensland Core Skills Test.*

Queensland Board of Studies. (1995). *The Queensland Core Skills Test.*

Rochex, J.-Y (2006).Social, methodological, and theoretical issues regarding assessment: Lessons from a secondary analysis of PISA 2000 literacy tests. In J. Green and A. Luke (Eds.) Rethinking Learning: What counts as learning and what learning counts (Review of Research in Education, 30, pp 163-212). Washington, DC: American Educational Research Association.

HORSE RACE

If a horse runs a distance of s metres in a time of t seconds, its average speed, in metres per second, over the distance is defined by the formula:

$$\text{average speed} = s/t$$

A race took place over a distance of 1200 metres. A marker post located alongside the race-track, 600 metres from the starting post, enabled the average speed of each horse to be calculated over each half (600 metres) of the race. The average speed of horse X over the first half of race is 12 metres per second and its average speed over the second half of race is 16 metres per second.

The average speed of horse X for the complete race was closest to

A. 13.3 metres per second.

B. 13.7 metres per second.

C. 14.0 metres per second.

D. 14.3 metres per second.

ACID RAIN

Below is a photo of statues called Caryatids that were built on the Acropolis in Athens more than 2500 years ago. The statues are made of a type of rock called marble. Marble is composed of calcium carbonate.

In 1980, the original statues were transferred inside the museum of the Acropolis and were replaced by replicas. The original statues were being eaten away by acid rain.

ACID RAIN 1

Normal rain is slightly acidic because it has absorbed some carbon dioxide from the air. Acid rain is more acidic than normal rain because it has absorbed gases like sulfur oxides and nitrogen oxides as well.

Where do these sulfur oxides and nitrogen oxides in the air come from?

.....

.....

The effect of acid on marble can be modelled by placing chips of marble in vinegar overnight. Vinegar and acid rain have about the same acidity level. When a marble chip is placed in vinegar, bubbles of gas form. The mass of the dry marble chip can be found before and after the experiment.

ACID RAIN 2

A marble chip has a mass of 2.0 grams before being immersed in vinegar overnight. The chip is removed and dried the next day. What will the mass of the dried marble chip be?

A. Less than 2.0 grams

B. Exactly 2.0 grams

C. Between 2.0 and 2.4 grams

D. More than 2.4 grams

ACID RAIN 3

Students who did this experiment also placed marble chips in pure (distilled) water overnight.

Explain why the students included this step in their experiment.

.....

.....

GREENHOUSE

Read the texts and answer the questions that follow.

THE GREENHOUSE EFFECT: FACT OR FICTION?

Living things need energy to survive. The energy that sustains life on the Earth comes from the Sun, which radiates energy into space because it is so hot. A tiny proportion of this energy reaches the Earth.

The Earth's atmosphere acts like a protective blanket over the surface of our planet, preventing the variations in temperature that would exist in an airless world.

Most of the radiated energy coming from the Sun passes through the Earth's atmosphere. The Earth absorbs some of this energy, and some is reflected back from the Earth's surface. Part of this reflected energy is absorbed by the atmosphere.

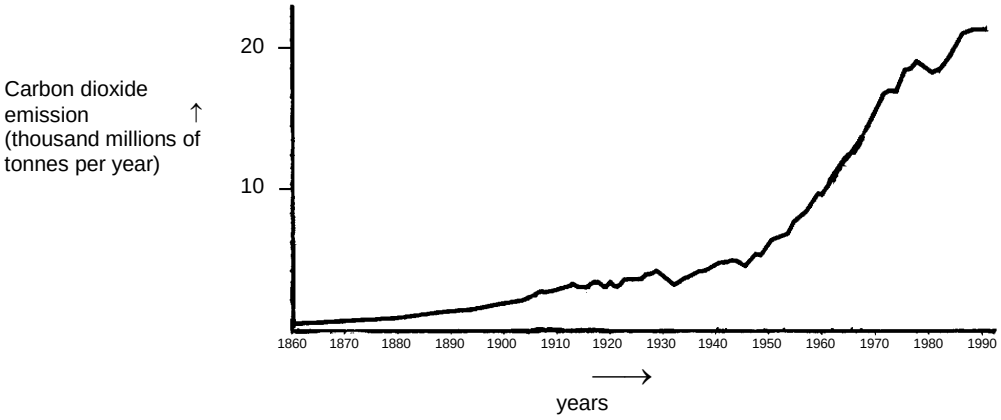
As a result of this the average temperature above the Earth's surface is higher than it would be if there were no atmosphere. The Earth's atmosphere has the same effect as a greenhouse, hence the term *greenhouse effect.*

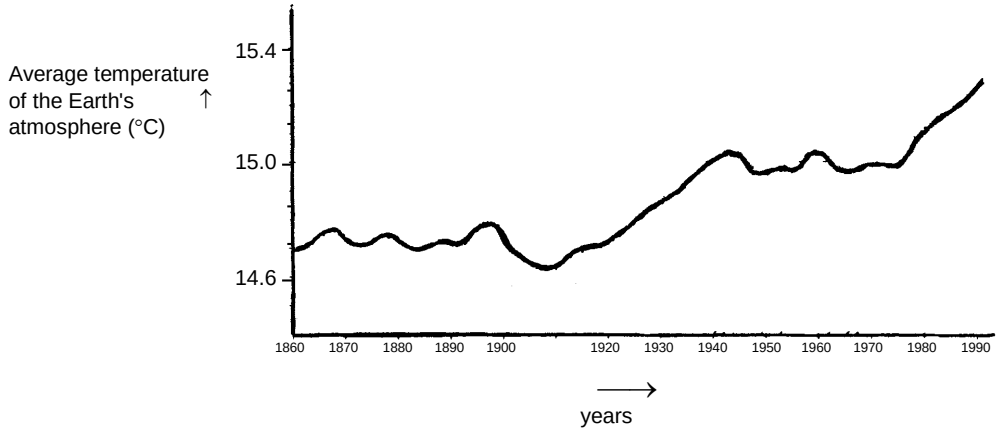
The greenhouse effect is said to have become more pronounced during the twentieth century.

It is a fact that the average temperature of the Earth's atmosphere has increased. In newspapers and periodicals the increased carbon dioxide emission is often stated as the main source of the temperature rise in the twentieth century.

A student named André becomes interested in the possible relationship between the average temperature of the Earth's atmosphere and the carbon dioxide emission on the Earth.

In a library he comes across the following two graphs.





André concludes from these two graphs that it is certain the increase in the average temperature of the Earth's atmosphere is due to the increase in the carbon dioxide emission.

GREENHOUSE 1

What is it about the graphs that supports André's conclusion?

.....

.....

GREENHOUSE 2

Another student, Jeanne, disagrees with André's conclusion. She compares the two graphs and says that some parts of the graphs do not support his conclusion.

Give an example of a part of the graphs that does not support André's conclusion. Explain your answer.

.....

.....

.....

GREENHOUSE 3

André persists in his conclusion that the average temperature rise of the Earth's atmosphere is caused by the increase in the carbon dioxide emission. But Jeanne thinks that his conclusion is premature. She says: "Before accepting this conclusion you must be sure that other factors that could influence the greenhouse effect are constant".

Name one of the factors that Jeanne means.

.....

.....

ANTS

On the planet Archid, in a galaxy far away, lives a species of intelligent ants that can be trained by Archidians to crawl in straight lines at constant speeds. The constant crawling speed of one ant may, however, differ from that of another.

An Archidian selects four trained ants, labelled I, II, III and IV, and places them on the comers of a square board of side exactly 100 centimetres.

Figure 1 shows the board, its diagonals, and the positions of the four ants.

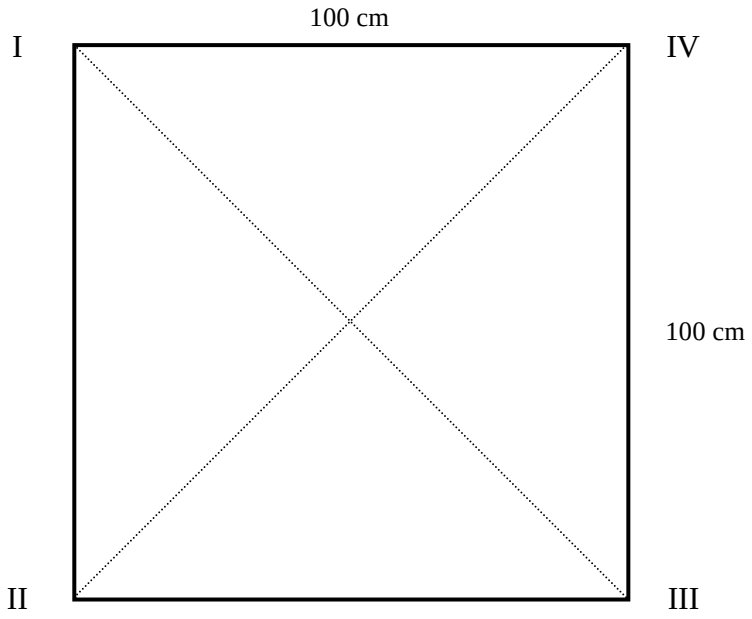


Figure 1

ITEM 9 [***]

Ants I and III are released simultaneously from their initial positions (in Figure 1) and crawl towards each other along the diagonal. Ant I crawls twice as fast as Ant III.
What distance does Ant I crawl before the two ants meet?

Show your working.

.....

.....

.....

.....

ITEM 10 [**]

Point P, not shown on the diagram, is reached by Ant IV 20 seconds after its release.

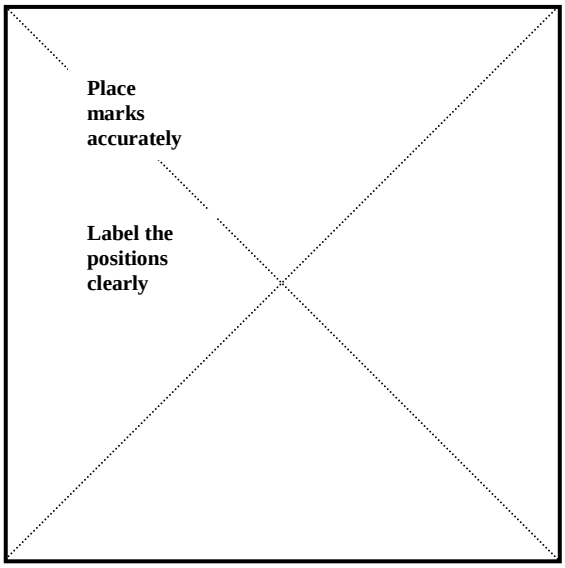
How many times does Ant IV pass point P in the 350 seconds after its release?

Give a number only. ……

ITEM 11 [***]

The diagram below is drawn to scale. On this diagram, mark the positions of Ants II and IV 50 seconds after their release. Label these positions II and IV respectively.

IV



II

OPERA

UNIT SEVEN

Opera enthusiasts wishing to familiarise themselves with an opera before attending a performance often consult an opera guide.

This unit relates to the entries in two different guides, $\mathcal{V}$ and $\mathfrak{M}$, for Verdi's opera *Il Trovatore* (The Troubadour).

ITEM 14 [*]

The cast of the opera, according to each guide, is listed below.

$\mathcal{V}$

Cast: Leonora; Inez, her attendant; Azucena, a gypsy; Manrico, her son; Ruiz, his lieutenant; Count di Luna; Ferrando, captain of the guard; an old gypsy; a messenger.

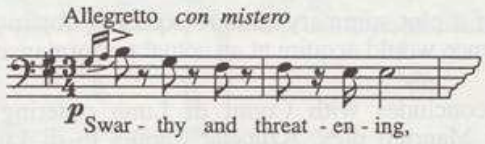
$\mathfrak{M}$

CAST: Count di Luna. Countess Leonora. Azucena, a gipsy. Manrico. Ferrando, Luna's vassal. Inez, Leonora's confidante. Ruiz, friend of Manrico.

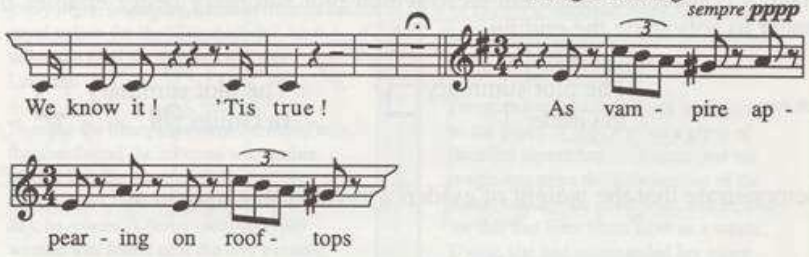
Which characters appear in the cast list in Guide $\mathcal{V}$ **only**?

Respond in this space. ☞

Excerpt 1



Excerpt 2



The words given in the excerpts are English translations whereas the required musical expression is described by conventional Italian terms and symbols. The meanings of these are given alongside.

Allegretto	moderately fast
Allegro assai agitato	lively and with much agitation
con mistero	with mystery
p	soft
sempre ***pppp***	always extremely soft

By assembling the clues, identify the specific point in the scene at which each excerpt would be performed and by whom it would be sung.

You may refer to the information in either guide.

Excerpt 1 would be sung by .....

when .....

.....

You may refer to the information in either guide.

Excerpt 2 would be sung by .....

when .....

.....

ITEM 16 [****]

The reader of a plot summary cannot expect to obtain all the information that the audience would acquire at an actual performance of a scene.

This opera concludes with Count di Luna ordering the execution of Manrico. As Manrico dies, Azucena[2] shouts to di Luna, "He was your brother! Mother, you are avenged!"

Note: [2]Azucena has by then been revealed to be the gypsy daughter mentioned in both guides.

Make a considered judgment as to which plot summary better enables the reader to anticipate the ending.

Tick one.

The plot summary in Guide $\mathcal{V}$ ❑

The plot summary in Guide $\mathfrak{M}$ ❑

Demonstrate that the weight of evidence favours this judgment.

Focus on significant differences in the plot summaries.

Fresh working copies of the plot summaries are provided on the next page.

PAINTING

Items 41-44

Within the rectangle in Figure 1 two diamond (parallelogram) shapes and a triangle have been drawn. The diamonds and triangle enclose three regions — one within each diamond and one within the triangle. Outside the diamonds and triangle but within the rectangle is a fourth region.

To paint the entire figure so that no two regions within the rectangle share the same colour on opposite sides of a boundary, three different colours will be needed, as indicated. Note that regions touching at a point only, as do the diamonds, may use the same colour.

The same colouring rules are to be used in the items that follow.

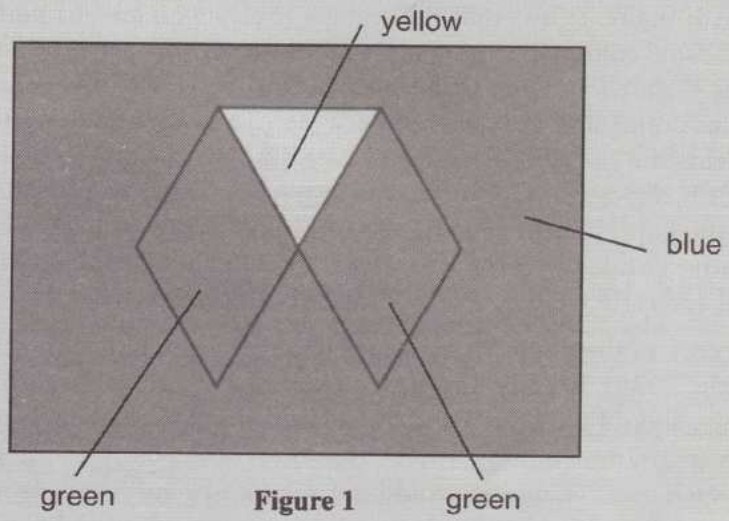


Figure 1

Item 41

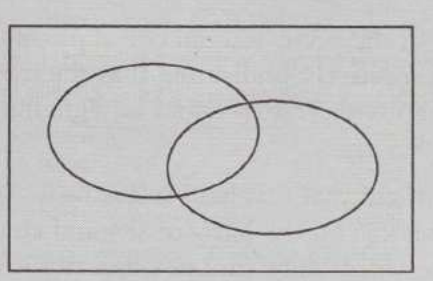


Figure 2

The smallest number of colours needed to paint Figure 2 is

A 2.

B 3.

C 4.

D 5.

THE UNBEARABLE LIGHTNESS OF BEING

Items 52-62

The two passages in this unit are from the English translation of the Czech novel, *The Unbearable Lightness of Being*.

PASSAGE I

All languages that derive from Latin form the word 'compassion' by combining the prefix meaning 'with' (*com-*) and the root meaning 'suffering' (Late Latin, *passio*). In other languages — Czech, Polish, German, and Swedish, for instance — this word is translated by a noun formed of an equivalent prefix combined with the word that means 'feeling' (Czech, *sou-cit*; Polish, *wspól-czucie*; German, *Mit-gefühl*; Swedish, *med-känsla*).

In languages that derive from Latin, 'compassion' means: we cannot look on coolly as others suffer; or, we sympathise with those who suffer. Another word with approximately the same meaning, 'pity' (French, *pitié*; Italian, *pietà* etc.), connotes a certain condescension towards the sufferer. 'To take pity on a woman' means that we are better off than she, that we stoop to her level, lower ourselves.

That is why the word 'compassion' generally inspires suspicion; it designates what is considered an inferior, second-rate sentiment that has little to do with love. To love someone out of compassion means not really to love.

In languages that form the word 'compassion' not from the root 'suffering' but from the root 'feeling', the word is used in approximately the same way, but to contend that it designates a bad or inferior sentiment is difficult. The secret strength of its etymology[1] floods the word with another light and gives it a broader meaning: to have compassion (co-feeling) means not only to be able to live with the other's misfortune but also to feel with that person any emotion — joy, anxiety, happiness, pain. This kind of compassion (in the sense of *soucit*, *wspólczucie*, *Mitgefühl*, *medkänsla*) therefore signifies the maximal capacity of affective imagination, the art of emotional telepathy. In the hierarchy of sentiments, then, it is supreme.

Note: [1]the origin or history of words

NB: The lines in this passage are numbered for ease of reference.

tem 56

The passage makes the point that a word's apparent counterpart in another language

A may incorporate the same root word yet have a different meaning.
B may have a very similar meaning even though the derivation is different.
C never has exactly the same meaning even when the derivation is the same.
D may have a significantly different meaning when the derivations are different.

List of synonyms for difficulty according to test takers

- long
- vague
- complex
- inaccessible
- complicated
- tough
- boring
- teasing
- obscurre
- intricate
- against the grain
- puzzling
- tedious
- perplexing
- ambiguous
- impossible
- plain hard
- containing several decision
- points (multistage)
- baffling
- tricky
- hohum
- confusing
- profound
- misleading
- deep
- trivial
- subtle
- novel
- verbose
- familiar
- turgid
- dense
- other