

Exploring tacit assumptions about comparability



Paul E. Newton
Date: 11 July 2008

Paper presented at the 34th Annual Conference of the International Association for Educational Assessment. 7-12 September 2008. Cambridge, United Kingdom.

Exploring tacit assumptions about comparability

From time-to-time, in England, influential stakeholder groups become concerned about a perceived lack of comparability between standards in examinations for different subjects. They typically argue that inappropriately severe grading standards are deterring students from studying the subject in question.

As the regulator for public examinations in England, Ofqual has a role in leading the national debate on standards between subjects. This is particularly challenging, since the underlying concept is contestable, let alone whether or not it is achieved in practice. So part of leading the national debate involves clarifying the terms in which it is constructed.

This presentation explores the idea of comparability between subjects. It identifies a range of different perspectives from which comparability can be understood, and a number of quite distinct definitions of comparability within these perspectives. It explains why there is no necessary relationship between methods for investigating comparability and definitions of comparability and, hence, why the same findings can be explained in different ways.

This analytical work is intended to help different stakeholder groups to appreciate the fundamental ways in which their understandings of comparability may differ. If successful, it raises questions like whether such incommensurable views might be reconciled through national debate and how to proceed if not.

Introduction

Last year, a major review of UK school leaving qualifications in modern foreign languages (MFL) concluded that:

the demands of languages in the GCSE are greater than for the great majority of subjects (Dearing & King, 2007, 3.20)

In saying this, Dearing & King were not proposing that the syllabus content for subjects like German, French and Spanish was too complicated; they were simply proposing that grading standards – the examination cut-scores – were too high.

There is a substantial amount of evidence which – when taken at face value – does indeed seem to support this conclusion (see Myers, 2006, collated on behalf of the Association for Language Learning). This can be illustrated using results from Coe (2008), which are re-presented in Figure 1. Coe draws some quite radical conclusions from his analyses:

Overall, the differences in the level of ability associated with a particular grade across different subjects are substantial. At grade C, for example, Latin is the best part of a grade higher than the next hardest subject, but even the next few subjects (Spanish, French, German) are about a grade higher than those at the other end of the scale (drama, textiles, office technology and English).

A similar pattern can be seen for grades other than C. For every pair of adjacent grades, there is substantial overlap; the higher grade in some subjects indicates less ability than the lower grade in others. For example, grade B in German, Spanish or French is about equivalent to an A in child development, textiles or PE. For the lower grades, the overlap seems bigger still, sometimes approaching two grades; a grade F in Spanish, IT or history is almost the same as a D in textiles, PE or drama. (Coe, 2008, pp.20-21)

In the light of evidence such as this, Dearing & King (2007) recommended action by the regulator for examinations in England:

This needs to be resolved one way or the other by a definitive study, followed by publication of the conclusions, because the present widely held perception in schools, whether right or wrong, is adversely affecting the continued study of languages through to the GCSE. (Dearing & King, 2007, 3.20)

The regulator was not entirely persuaded by the evidence, however; believing there to be strong counter-arguments, and arguing that levelling grading standards on the

basis of statistical evidence (like that provided by Coe) would have harmful effects on students in a range of subjects (QCA, 2008).¹

The QCA response claimed that the kind of statistical evidence provided by Coe was only indicative of problems with grading standards under one particular definition of comparability. Moreover, it claimed that the UK qualifications system ought really to be understood on the basis of an entirely different definition.

If this is true – that a range of possible definitions of comparability exist – then this challenges the very idea of definitive empirical study (contra Dearing & King); since the validity of any particular technique for monitoring comparability will depend entirely upon the definition of comparability adopted.

The idea that there might be a range of possible definitions of comparability – each with different implications for grading standards across subjects – is not universally recognised, even within the measurement profession. So this paper is an attempt to develop the argument provided in QCA (2008); defending the idea of alternative definitions, and illustrating their different implications for grading standards across subjects. However, instead of arguing the case for or against the particular definition adopted by QCA, or attributed to Coe, it simply aims to establish a commonsense conceptual framework, within which to locate such competing perspectives.

A taxonomy of definitions of comparability

In proposing a taxonomy of definitions of comparability, this paper is far from original. Similar frameworks can be found in reports as diverse as: Schools Council Forum on Comparability (1979); Christie & Forrest (1981); Orr & Nuttall (1983); Mislevy (1992); Linn (1993); Cresswell (1996); Baird, et al. (2000); Baird (2007); and Coe (2007). The aims of developing a new taxonomy were:

1. to provide a commonsense framework with which to highlight fundamental differences between a range of definitions
2. to foreground a central distinction between definitions of comparability and methods for linking standards or for monitoring comparability
3. to recognise each definition as legitimate in its own right.

¹ In early 2008, the division of the Qualifications and Curriculum Authority with the largest share of responsibility for regulatory work separated to become the Office of the Qualifications and Examinations Regulator (Ofqual); it is now an independent body which reports to Parliament.

A grounding principle of the new taxonomy is that comparability is not equating-done-less-well but something quite different instead. Exactly what it is, is given by the particular definition of comparability adopted.

The implication here is that different definitions may need to be adopted within different contexts and situations to respond to any of a range of potential comparability dilemmas. Unless these definitions are made explicit, it will not be possible to make correct decisions on methods for linking standards, or for monitoring comparability, and it will not be possible for stakeholders to draw appropriate inferences from comparability claims.

Three perspectives on comparability theory

The following taxonomy outlines three quite distinct perspectives upon comparability theory, within which are clustered sets of distinct definitions. Each of the perspectives is similar, in the sense that comparability is always defined in terms of a profile of student attainment (that is associated with a particular grade standard). But each perspective is different in terms of how that basis in student attainment is conceptualised. Each provides a different kind of answer to the question: what is it about the attainment of students – who have been awarded the same grade from different examinations – that could possibly be the same?

First, from a **phenomenal** perspective, attainments are the same in the sense of their discernible characteristics.² Comparability concerns the features, properties or dispositions that comprise attainment: the outcomes from learning that similarly graded students have in common.

Second, from a **causal** perspective, attainments are the same in the sense of their antecedents. Comparability concerns the factors that result in attainment: the inputs to learning that similarly graded students have in common.

Third, from a **predictive** perspective, attainments are the same in the sense of the prospects that they offer. Comparability concerns the potential that is implicit in attainment: the likelihood of future success that similarly graded students have in common.

1 Phenomenal definitions (the ‘present tense’)

For two examinations to have comparable grading standards, students who score at equivalent grade boundary marks must be the same in terms of the character of their attainments. They must be equally good at knowing, understanding and being able

² My thanks to Robert Coe for coining this term.

to do X, where X is the set of knowledge, skill and understanding (KSU) that is common to both examined constructs.

1a The ‘full construct’ phenomenal definition

In the first case, the set of KSU that is common to the syllabuses of both examinations also exhausts the KSU of both syllabuses. As such, this definition can only apply to situations in which alternative forms of an examination are developed to assess a single syllabus. For example, when a single examining board develops a new form of an examination (e.g. GCSE biology) for each successive session.

It is the only definition of comparability that is compatible with the traditional concept of score equating: where ‘the same’ implies precise identity of score meaning (at linked scores across test forms) expressed in terms of the construct assessed by both test forms.

1b The ‘part construct’ phenomenal definition

In the second case, the set of KSU that is common to the syllabuses of both examinations does not exhaust the KSU of those syllabuses, but represents a common subset of the distinct examined constructs. This definition can apply to situations in which examinations are designed to assess similar syllabuses, rather than a single common syllabus. For example, when examining boards develop examinations in cognate subject areas (e.g. GCSE psychology and sociology); as long as the subjects in question support the development of a common core proficiency (e.g. social research skill).

Here, the point is to define comparability purely in terms of the common core proficiency that students acquire through studying the subjects in question. Importantly, this forms the basis for defining comparability even when the distinct subjects develop the common core proficiency at different rates or to different extents. If that strikes you as unfair, then this definition is probably not for you.

2 Causal definitions (the ‘past tense’)

The weakness of phenomenal definitions is that they require there to be at least something common to the attainments, or proficiencies, arising from the study of two examination syllabuses (for their standards to be linked). For subjects as disparate as art, French and chemistry this is hard to imagine. Easier to imagine, though, is a degree of commonality in the causes of attainment across different subject areas.

From this perspective, for two examinations to have comparable grading standards, students who score at equivalent grade boundary marks must be the same in terms of the causes of their attainments. That is, they must have experienced equally the set of causes that are common to attainment in both examinations. Even though the characteristics of attainment in those examinations might differ radically – as for

examinations in different subject areas – it might still be possible to define comparability in terms of the factors that enabled those attainments to be achieved.

There are very many possible causal determinants of attainment for any examined subject area; each, in its own way, making a difference to how high or low a student attains by the end of a period of study. Figure 2 presents a reasonably large set of causes, all of which strike me as both direct (i.e. not mediated by other factors) and independent (i.e. distinct from other factors).

2a The ‘all causes’ causal definition

Under the all causes definition, comparability is understood in terms of the sum total impact of all of the possible factors that causally affect the level of attainment achieved by a student by the end of a course of study. This definition can apply to any situation in which it is possible to conceive of attainment being attributable to a common set of causes, such as those illustrated in Figure 2: from examinations designed to assess a single common syllabus, to examinations designed to assess entirely different domains.

2b The ‘specific causes’ causal definition

Any single cause, or combination of causes, of attainment could, in principle, form the basis of a definition of a specific causes definition. The aggregate of capability, flair, strategy, dedication and time would be an obvious choice, since it represents the sum total of what students contribute to their learning during their course of study. We might call this the ‘student-level’ causal definition.

Or we might want to invoke the aggregate of every input factor which operates during a course of study, i.e. all of the causes illustrated in Figure 2 with the exception of prior attainment. We might call this the ‘system-level’ causal definition.

Again, any number of specific causes definitions could be constructed, although there ought to be some good social or educational reason for the particular choice of input(s). For example, it is hard to imagine why anyone would want to create a definition based on the aggregate of strategy and peer dialogue alone.

3 Predictive definitions (the ‘future tense’)

While phenomenal definitions represent attainment in the present tense (comparable characteristics) and causal definitions represent attainment in the past tense (comparable causes) predictive definitions represent attainment in the future tense (comparable prospects).

For two examinations to have comparable grading standards, students who score at equivalent grade boundary marks must be the same in terms of the extent to which their attainments predict their future success. That is, they must have the same potential for success in the future.

3a The ‘common prospect’ predictive definition

When different examinations are assumed to prepare students for the same opportunities, this constitutes a special case in which comparability is tantamount to the predictive validity of grading standards. This could apply, for instance, when different examining boards develop different forms of an examination (e.g. A level biology) all intended as adequate and appropriate preparation for a common university course (in biology). It could equally apply more generally, were success in any of a range of examinations at a certain level (e.g. Advanced level) to be taken as evidence of a broad potential for success in any of a range of examinations at a higher level (e.g. degree course).

3b The ‘respective prospects’ predictive definition

There was once a time when the examining boards in England agreed upon the following definition of the A level grade B standard:

Good. (This is the standard considered to give a reasonable assurance that the candidate is likely successfully to pursue a University Honours course in the subject.) (SSEC, 1960, Appendix E, p.26)

So, for students at grade B boundary marks in, say, A level French and chemistry:

- even though their attainments would be different in character (from a phenomenal perspective)
- and even though their attainments might be different in cause (from a causal perspective)
- their attainments might still be considered comparably graded if they indicated the same potential for success in French and chemistry university honours courses, respectively.

Under this definition, the level of attainment in A level French that best predicted (a specified level of) success in honours French would be linked to the level of attainment in A level chemistry that best predicted (a specified level of) success in honours chemistry, both of which would represent the A level grade B standard.

Comparability dilemmas

Within England’s public examination system, there has always been an expectation that a grade z (A, B, C etc.) in one subject ought somehow to represent the same standard as a grade z in all others within a particular type of qualification. However, there has never been a clear explanation of what this might mean or how it might

translate into impacts upon grade distributions. This can be illustrated through the following comparability dilemmas.

First, consider the case of a curriculum with only optional subjects. In England, students studying for A level examinations choose around three or four subjects from a wide selection on offer. Some subjects, such as mathematics and the sciences, are associated with high status careers (e.g. medicine, rocket science); while others, such as sociology, have somewhat less kudos. Furthermore, on average, students who opt for (say) A level chemistry achieve a better GCSE grade profile than students who opt for (say) A level sociology. Some would argue that this indicates their higher general capability or aptitude for learning (e.g. Kelly, 1976). If we were to assume that both subjects were equally motivating – such that students put in the same amount of effort – would it be fair if the A level chemistry students ended up with higher grades?

Dilemma 1. If students in different subjects differ only in terms of their general capability or aptitude for learning – having applied the same amount of effort to their respective courses – should those in the lower general capability subject really be awarded lower grades?

Second, what if courses in certain subjects begin from a higher baseline of attainment than others? A very extreme example comes from GCSE Welsh, where there are two forms of the examination: one for students who grow up in Welsh-speaking communities, are educated in Welsh-medium schools, and who speak Welsh as a first language (Welsh 1L); and one for students who only speak Welsh as a second language (Welsh 2L). There is no question that the students who take Welsh 1L are vastly more proficient than those who take Welsh 2L. On the other hand, the percentages of students who are awarded grade C or better at GCSE are almost equal, at around 70% in 2006. Is this fair?

Dilemma 2. If students in different subjects differ only in terms of their baseline level of attainment upon starting their respective courses – having applied the same amount of effort and capability to their respective courses – should those in the lower prior attainment subject really be awarded lower grades?

Finally, what if only the quality of teaching differed across subjects? At GCSE it is not uncommon for certain subjects – such as biology and history – to be taught by subject experts; while other subjects – such as physical education and information technology – tend not to be. In fact, in a survey of maintained schools in England, undertaken by Smithers & Tracey (2003): 66% of secondary school history teachers had a first degree in history, while only 20% of PE teachers did; and 85% of secondary school biology teachers had a degree in their subject while only 7% of IT teachers did. If we were to assume that all other things were equal – such that

students who studied biology applied no more effort nor capability than students who studied IT – would it be fair if biology students ended up with higher grades?

Dilemma 3. If students in different subjects differ only in terms of the quality of teaching experienced – all other things being equal – should those in the lower teaching quality subject really be awarded lower grades?

Dilemmas such as these present the challenge to which comparability theory must rise. Tables 1a to 1c illustrate plausible consequences of adopting each of the different definitions of comparability, in terms of their likely impact upon grade distributions in the subjects under comparison.³ The first column identifies what is assumed to differ between subjects, assuming that all other things are equal.

The first point to notice is that neither of the phenomenal definitions can be used as a rational basis for defining comparability between subjects, except with respect to Welsh 1L and 2L (which assess a similar using/reading/writing construct at different levels). This is because the different subjects develop substantially different competencies and there is therefore little in the way of learning outcomes that might be considered similar. As far as Welsh is concerned, if comparability were to be defined from a phenomenal perspective, then there would be a clear expectation of a far higher grade distribution for Welsh 1L; in stark contrast to the situation that actually pertains.

All of the causal definitions suggest the same impact upon grade distributions for the first dilemma, where students differ only in terms of their (subject-general) capability. The implication here is that students ought to receive grades that reflect how much they have put into a course of study. If students in one cohort exercise higher levels of capability than students in another, all other things being equal, then they ought to be rewarded with higher grades.

As far as the second and third dilemmas are concerned, implications from the causal perspective differ. For Welsh 1L and 2L the difference turns on whether the definition allows that higher baseline attainment should be recognised in higher grades. For the subjects that differ only in terms of teaching quality – biology and IT – the difference turns on whether grades ought to reflect causes of attainment beyond the student.

Anticipated impacts from predictive definitions are somewhat harder to disentangle. For the respective prospects definition this is because the relationship between predictor grades (e.g. A level sociology and chemistry) and criterion grades (e.g. degree level sociology and chemistry) will depend upon grading decisions made

³ A higher grade distribution means that, on average, students would be awarded higher grades.

within the criterion subjects. But, for the sake of argument, we might assume that grades in the criterion subjects will be awarded so as to discriminate effectively between the students who opt to study them. Thus grade distributions across criterion subjects would be fairly similar; which would probably translate into similar grade distributions across predictor subjects too.

Impacts for the common prospect definition are easier to anticipate since we are talking about the prediction of a single criterion (with a single grading standard). Again, the implications will not always be obvious. However, assuming that the criterion subject is not unduly differentially related to either of the predictor subjects, the impacts will probably depend upon the extent to which the examination cohorts differ in terms of a general capability or aptitude for learning: the higher the general capability of students within a cohort the more likely they are to be successful in any given higher level course. Given the all-other-things-being-equal clause, this would imply the same grade distributions for Welsh 1L versus 2L, and IT versus biology; but a higher grade distribution for chemistry than for sociology.

The obvious but central point is that different definitions of comparability will have quite radically different implications for grade distributions across subjects. Unless the underlying definitions of comparability have been made explicit for public scrutiny these differential impacts will be impossible to defend rationally.

Implications

This paper began by setting out three aims that prompted the development of the new taxonomy. It will end by considering them in reverse order.

The legitimacy of alternative definitions

The third aim was to recognise each of a range of alternative definitions as legitimate in its own right. Not all taxonomies share this aim. Sometimes (as in Cresswell, 1996) the intention is to knock down each of a series of candidate definitions, one-by-one, before reaching a final definition which is able to withstand all criticism.

The point emphasised by the present paper is that different definitions may need to be adopted within different contexts and situations, to respond to any of a range of potential comparability dilemmas. None of the definitions is conceptually superior to any of the others in any absolute, de-contextualised sense. As with all aspects of educational assessment, the guiding principle in deciding between alternative definitions ought to be the uses to which assessment results are put.

Equally importantly, the present paper is intended to challenge a growing trend of theorising linking (comparability) as though it were a special case of equating; more specifically, equating-done-less-well. The reverse happens to be true: equating theory is simply a special case, or subset, of comparability theory. Equating is based

upon one legitimate definition of comparability amongst many (the full construct phenomenal definition).

Since equating theory is a subset of comparability theory it makes no sense to use equating criteria – such as group invariance – to judge the defensibility of linking relationships.⁴ All that violation of the group invariance criterion really indicates is that the to-be-linked examinations assess somewhat different constructs. Yet, when linking standards between different subjects, you know that already! The decision whether to link standards is not primarily technical, but conceptual, pragmatic or political.

The difference between definitions and methods

This leads on to the second aim: to foreground a central distinction between definitions of comparability and methods for linking standards or for monitoring comparability. Historically, much of the debate in the UK has concerned whether or not the assumptions of a particular method are valid. Unfortunately, the longstanding debate over adequacy of method has tended to obscure a deeper and far more important debate over appropriateness of definition.

Critics of the work by Robert Coe, for example, have tended to claim that the assumptions underlying his statistical modelling are invalid (e.g. Murphy, 2007). The particular form of statistical modelling that Coe employs requires the assumption that individual students ought, on average, to be awarded similar grades across different subjects. His critics have claimed that this assumption fails to hold in practice, since:

- even on average, students are likely to put more effort into studying an interesting and motivating subject than a boring one, and
- even on average, the quality of teaching within certain subjects is likely to be better than the quality of teaching within others.⁵

However, as Coe (e.g., Coe, 2007) has rightly responded, the assumption that individual students ought, on average, to be awarded similar grades across different subjects is not an assumption of the *method* that he uses, but is part of a *definition* of comparability that might legitimately be adopted. Under this definition, standards ought to be linked *purely* on the basis of general capability, or aptitude for learning. If

⁴ Under the group invariance criterion, the linking relationship between the to-be-linked tests must be equivalent for all identifiable subgroups.

⁵ Factors such as these can translate into differential linking relationships across subgroups, i.e. violation of the group invariance criterion; for example, if girls are more motivated than boys in some subjects but not others.

so, then the arguments traditionally voiced against the use of statistical methods simply do not follow, since it is irrelevant to this definition how much effort students apply across different subjects, or how high the quality of teaching experienced. The confusion occurs because Coe defends the legitimacy of a very narrow 'specific causes' definition of comparability (based upon general capability alone); while his critics seem to assume a position far closer to the 'all causes' definition. His method and his conclusions are quite defensible in relation his definition; just not in relation to theirs.

In summary: methods are distinct from definitions; and results from the application of methods can only be interpreted in relation to underlying definitions. This means that decisions on methods ought not to be made in advance of decisions on definitions. If you wanted to apply an 'all causes' definition of comparability then your ideal method might be a complex multiple regression. However, if you wanted to apply a very narrow 'specific causes' definition, based upon general capability alone, then your ideal method might be a simple linear regression.

The importance of a commonsense framework

Finally, to the first aim: to provide a commonsense framework with which to highlight fundamental differences between a range of definitions. Ultimately, the rationale for this work is to support rational debate on the most appropriate definition of comparability to apply in any given context or situation. To this end, the definitions need to be capable of making sense to a range of stakeholders.

To the extent that the taxonomy makes sense and adds structure to the debate on comparability, it should help users with different preferences and needs to appreciate that their tacit assumptions are not necessarily shared by others, and that these differences may have some legitimacy. This is the starting point for debate. Clearly, though, this is only the starting point. The more profound challenge is to work towards consensus on the most appropriate definitions to adopt in each of the range of contexts and situations.

References

- Baird, J. (2007). Alternative conceptions of comparability. In P.E. Newton, et al. (Eds.). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
- Baird, J., Cresswell, M.J., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, 15, 213–229.
- Christie, T., & Forrest, G.M. (1981). *Defining public examination standards*. *Schools Council Research Studies*. London: Macmillan Education.

- Coe, R. (2007). Common examinee methods. In P.E. Newton, et al. (Eds.). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
- Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education*, pre-print version.
- Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: JohnWiley and Sons.
- Dearing, R. & King, L. (2007). *Languages review*. Nottingham: Department for Education and Skills Publications.
- Kelly, A. (1976). *The comparability of examining standards in Scottish Certificate of Education Ordinary and Higher grade examinations*. Dalkeith: Scottish Certificate of Education Examination Board.
- Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83–102.
- Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, prospects*. Princeton: Educational Testing Services.
- Murphy, R. (2007). Common test methods. In P.E. Newton, et al. (Eds.). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
- Myers, H. (2006). *The “severe grading” of MFL grades at GCSE and A level*. London: Association for Language Learning.
- Orr, L. & Nuttall, D.L. (1983). *Determining standards in the proposed single system of examining at 16+*. *Comparability in Examinations Occasional Paper 2*. London: Schools Council.
- Qualifications and Curriculum Authority. (2008). *The grading of GCSE Modern Foreign Languages. Letter to Minister of State, Jim Knight, 4 February*. London: Qualifications and Curriculum Authority.
- Schools Council Forum on Comparability. (1979). *Standards in public examinations: problems and possibilities. Comparability in Examinations Occasional Paper 1*. London: Schools Council.
- Smithers, A. & Tracey, L. (2003). *Teacher qualifications*. London: The Sutton Trust.

Figure 1. Relative 'difficulties' of achieving each grade in 34 GCSE subjects, ordered by weighted average difficulty. From Coe (2008).

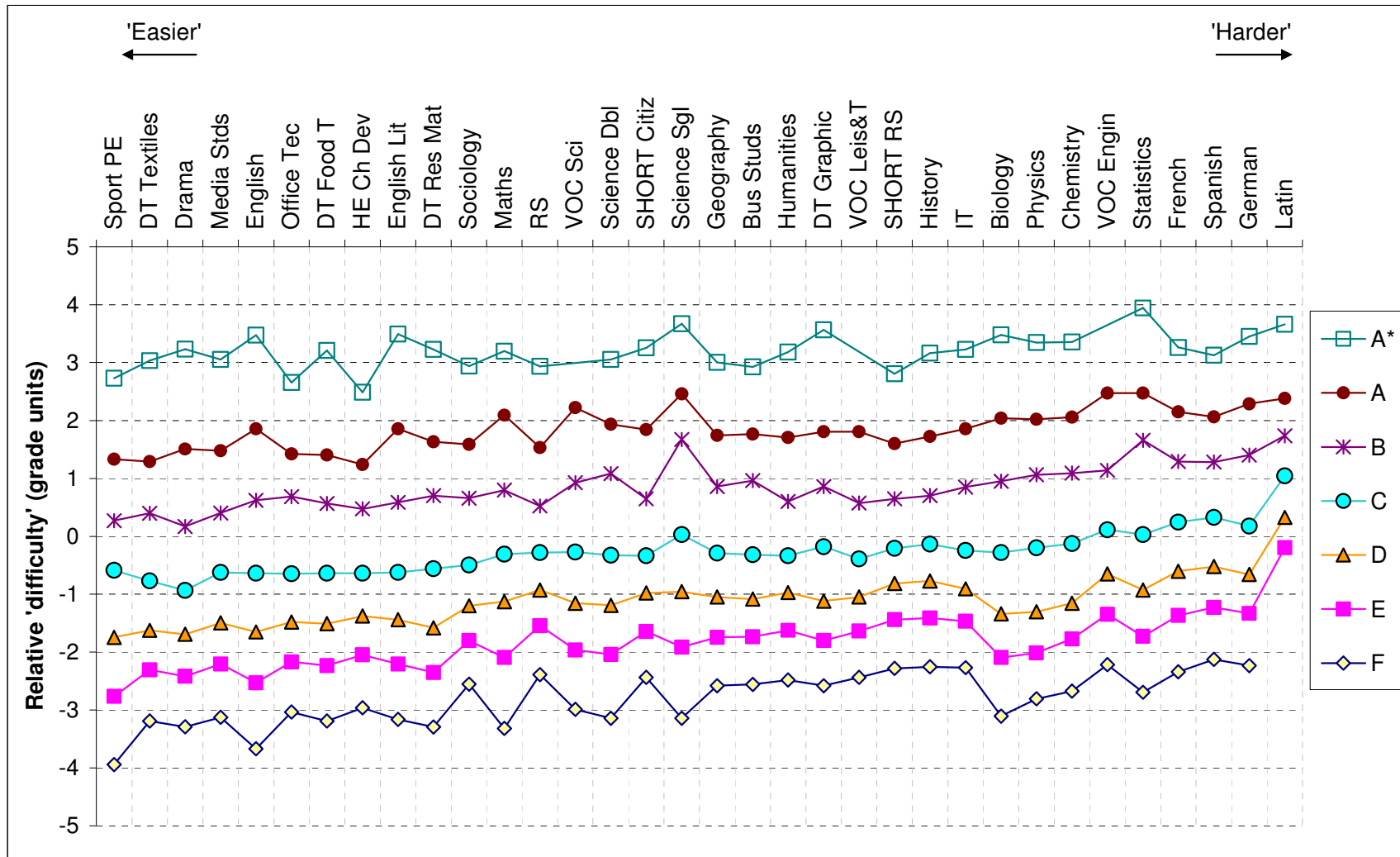


Figure 2 An illustration of direct and independent causes of attainment

First, how the student contributes to their own learning experience:

- a. **Prior attainment** – baseline level of attainment before embarking upon the course of study
- b. **Capability** – degree of general aptitude for learning (implicating constructs like working memory capacity, intelligence, problem-solving ability)
- c. **Flair** – degree of specific aptitude for learning within the domain of study
- d. **Strategy** – quality of exercise of learning skills (implicating constructs like meta-cognitive skill, memorisation skill)
- e. **Dedication** – quality of engagement with the task of learning (implicating constructs like effort, diligence, attention, concentration, focus)
- f. **Time** – amount of time spent:
 - undertaking self-initiated learning activities (revision, personal study)
 - undertaking teacher-initiated learning activities (worksheets, projects, tests, debates, homework)
 - in learning dialogue with teachers
 - in learning dialogue with peers
 - in learning dialogue with family members.

Second, how others contribute to the student's learning experience:

- a. **Teacher pedagogy** – quality of instruction (clarity, structure and pacing of delivery of syllabus content; quality of tasks and exercises; match of learning outcomes to teaching methods; choice of contexts for delivering syllabus content)
- b. **Teacher dialogue** – quality of dialogue with teachers (during lessons and through written feedback on performance)
- c. **Peer dialogue** – quality of dialogue with peers (during lessons, homework or revision)
- d. **Family dialogue** – quality of dialogue with family members (during homework or revision).

Third, how learning resources contribute to the student's learning experience:

- a. **Learning materials** – quality of learning software (text-books, revision guides, dedicated websites)
- b. **Learning tools** – quality of learning hardware (library, computers, whiteboards)
- c. **Syllabus design** – quality of design of syllabus (suitability of syllabus content and structure for the acquisition of intended learning outcomes).

Table 1a Likely grade distribution impacts according to phenomenal definitions

All other things being equal, what if...	Phenomenal definitions	
	Full construct	Part construct
Chemistry students have higher general capability than sociology students	n.a.	Probably n.a.
Welsh 1L students have higher baseline attainment than Welsh 2L students	Probably n.a.	Welsh 1L receives (far) higher grade distribution
IT students have lower quality teachers than biology students	n.a.	Probably n.a.

n.a. = not applicable, i.e. definition will not extend to this context.

Table 1b Likely grade distribution impacts according to causal definitions

All other things being equal, what if...	Causal definitions		
	All causes	Specific causes (student-level)	Specific causes (system-level)
Chemistry students have higher general capability than sociology students	Chemistry receives higher grade distribution	Chemistry receives higher grade distribution	Chemistry receives higher grade distribution
Welsh 1L students have higher baseline attainment than Welsh 2L students	Welsh 1L receives (far) higher grade distribution	Welsh 1L & 2L receive same grade distribution	Welsh 1L & 2L receive same grade distribution
IT students have lower quality teachers than biology students	IT receives lower grade distribution	IT & biology receive same grade distribution	IT receives lower grade distribution

Table 1c Likely grade distribution impacts according to predictive definitions

All other things being equal, what if...	Predictive definitions	
	Common prospect	Respective prospects
Chemistry students have higher general capability than sociology students	Depends, but chemistry probably receives higher grade distribution unless criterion is particularly related to sociology	Chemistry & sociology probably receive same grade distribution (although depends on grading decisions at higher level)
Welsh 1L students have higher baseline attainment than Welsh 2L students	Welsh 1L & 2L probably receive same grade distribution if criterion is not Welsh	Welsh 1L & 2L probably receive same grade distribution if criterion is Welsh at different levels
IT students have lower quality teachers than biology students	Depends, but IT & biology probably receive same grade distribution unless criterion is particularly related to either	IT & biology probably receive same grade distribution (although depends on grading decisions at higher level)

Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by The Office of the Qualifications and Examinations Regulator in 2008.

© Qualifications and Curriculum Authority 2008

Ofqual is part of the Qualifications and Curriculum Authority (QCA). QCA is an exempt charity under Schedule 2 of the Charities Act 1993.

Reproduction, storage or translation, in any form or by any means, of this publication is prohibited without prior written permission of the publisher, unless within the terms of the Copyright Licensing Agency. Excerpts may be reproduced for the purpose of research, private study, criticism or review, or by educational institutions solely for education purposes, without permission, provided full acknowledgement is given.

Office of the Qualifications and Examinations Regulator
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk