Can qualitative phenomena be quantified?

A study of the validity of quantitative and qualitative mathematical approaches to measuring educational attainment

Alex Scharaschkin

St Anne's College

Michaelmas Term 2024

Acknowledgements

This study was undertaken while I was working at AQA Education. I am very grateful to AQA for meeting the cost of fees, and for access to anonymised examination data. Some parts of the dissertation were presented at meetings of AQA's Research Committee, and I thank members of the Committee for their feedback.

I would like to thank my supervisors Jo-Anne Baird (for the first period of the project), Jenni Ingram, and Josh McGrane, for helpful and constructive feedback and suggestions throughout, for stimulating discussions, and support as I tried to balance the demands of work, family, and study. I would also like to thank Dougal Hutchinson, Anne Pinot de Moira, and Michelle Meadows for examining my transfer and confirmation of status submissions, and for their considered feedback which helped determine the final shape of this project.

Finally to Eleanor, Dora, and Matilda, who supported me in every respect *ab ovo usque ad mala*, my heartfelt thanks, gratitude, and love.

'In fact the idea of an essential link between measurement and number-assigning comes from taking a myopic view of mathematics and an equally myopic view of measurement.

Mathematics is not the study of numbers. Numbers are merely a particular, if useful and user-friendly instance of mathematical structure. Granted, the numbers' salience is not the only reason to think of measurement mainly in terms of numerical outcomes. Some paradigmatic examples of measurement are indeed numerical. But ... we will have to look beyond those.'

—van Fraassen (2008, p. 159)

Abstract

In educational assessment students are given tasks – such as examination questions – and information obtained from their responses is used to summarise the quality of their performances. This summary information is often said to *measure* some feature of substantive interest, such as 'attainment' or 'proficiency', and is often expressed numerically, for example, as a score.

Usually, information is collected about different attributes of each performance, in the form of sub-scores for different parts or aspects of the performance. In many educational assessment contexts, these partial valuations encode assessors' judgements of quality, against criteria that are deemed to be appropriate for the assessment in question. For example, 16-year-old students taking a GCSE English language assessment in England are asked to produce a piece of creative writing on a given theme. Students' performances are then assessed, and awarded marks, with respect to the separate criteria of 'content and organisation' and 'technical accuracy'. The encoding (mapping of performances to scores) takes place in accordance with an inter-subjectively defined notion of what constitutes *better* or *worse* performance.

In the sense that these judgements determine whether performances are 'better' or 'worse' than each other, with respect to given criteria, assessors' judgements provide only ordinally-structured information about particular attributes of performances. In

most applications of educational assessment, however, it is assumed that it is also meaningful to combine or aggregate sub-scores as if they derive from underlying proficiency factors that have *quantitative* structure (i.e. whose value-structures are isomorphic to the field of real numbers). Overall measurements of attainment or proficiency are generally derived from sums, linear combinations, or averages, of the numbers representing the sub-scores, or from statistical models that use these numbers to generate estimates of values of hypothesised quantitative, continuous, latent variables.

This dissertation explores, conceptually and empirically, the extent to which the assumption of quantitative structure actually holds in some educational assessment contexts of considerable substantive interest, namely the public examinations taken by pupils in England around the ages of 16 and 18, on the basis of which high-stakes qualifications (GCSEs and A levels) are awarded.

It considers some of the consequences that follow if the phenomena of interest in educational assessment cannot validly be regarded as quantities. It suggests that some of the problems that arise from a measurement-theoretic perspective can be overcome by adopting a structuralist account that widens the traditional notion of measurement beyond the location of the value of a quantity on an equal-interval scale. And it proposes a framework for the analysis of data arising from educational assessment procedures that models assessment outcomes as fuzzy relational systems between learners and construct-relevant attributes.

This study therefore makes a contribution both to the theory and practice of educational assessment. From a theoretical perspective, it is hoped that the development and application of 'qualitative mathematical' approaches to the measurement of assessment constructs may facilitate improving the construct-validity of assessment procedures, in particular large-scale curriculum-embedded assessments such as public examinations in England. From a practical viewpoint, a key question is the extent to which there may

be significant divergences in conclusions about students' grades, awarded on the basis of such assessments, that can ultimately be traced to divergent assumptions about how educational attainment or proficiency is theorised as a measurand – as a quality, or as a quantity. The results of this study suggest that such divergences are likely to be more significant in some subjects than in others. Consequently there may be no need substantially to change existing quantitatively-based approaches to educational measurement in some cases, where assessments seem to be relatively robust to a departure from the quantity assumption. In other cases, cleaving to unwarranted assumptions about the structure of the assessment construct may jeopardise the validity of inferences drawn from the results of the assessment.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context and motivation for this study

This thesis brings together a set of papers that collectively explore the question of when it is possible to quantify qualitative phenomena. This question underpins much of the practice of educational measurement. For example, L.L. Thurstone, one of the pioneers of educational measurement as a discipline in its own right, said: 'When the idea of measurement is applied to scholastic achievement, for example, it is necessary to *force* the qualitative variations into a scholastic linear scale of some kind' (Thurstone, 1928, p. 534, emphasis added). More recently, a leading textbook on Rasch measurement theory (Andrich and Marais, 2019, p.4), states that 'the observations from an assessment, often referred to as *responses*, are qualitative. However, as a step towards measurement, these responses have an order which immediately implies more or less of the property to be assessed'. The authors go on to treat this postulated property as a *numerical quantity* – rather than simply regarding it as an *ordered structure*, for instance.

The 'quantitative imperative' with respect to the study of educational phenomena is clearly open to question. Michell (2012a), for instance, notes the implausibility, on the grounds of cognitive theory, of supposing mathematics ability to be a quantity, suggesting rather that 'abilities are attributes composed of ordered hierarchies of cognitive resources,

the differences between which are heterogeneous' (p. 265). Baird et al. (2017) conclude that the psychometric literature does not, for the most part, theorise the concept of 'ability' by deriving the assumptions that underpin how it is modelled mathematically from substantive theories of learning, teaching, or cognition. McGrane and Maul (2020) note that the quantitative imperative in human sciences such as psychology and educational assessement derives from a conception of measurement (discovery or estimation of the values of continuous quantities) that is not even universally true in the physical sciences. They argue that in general scientific progress begins with articulation of substantive theories, and the development of methods and models capable of testing aspects of those theories. There is no *a priori* reason to assume that the phenomena of interest will necessarily be usefully treated as having quantitative structure (and disciplines such as molecular biology and quantum mechanics, for instance, bear witness to this).

The papers presented in this thesis are sensitive to these foundational issues for the statistical (psychometric) methods that are commonly used to analyse the outcomes of educational assessment procedures, and that provide the warrants for substantive conclusions about students or learners on the basis of how they respond to certain tasks or activities. As outlined in the following section, each paper explores a separate research question. Taken together, they argue for a narrative that unfolds along the following lines.

1. Not only theoretical considerations, but also empirical evidence, indicate that it is in general implausible that the constructs investigated in educational assessments have the mathematical structure of quantities, even though such an assumption is a pre-requisite for the application of the statistical and psychometric methods commonly used in educational measurement.

2. A re-conception of measurement, drawing largely on van Fraassen's (2008) notion of *location in a logical space,* allows processes such attestation to students' having

attained certain *standards, grades,* or *levels of competence* to be formulated in terms that do not require mapping or reducing partially-ordered information to numerical scores. From this viewpoint, it is not necessary to force qualitative phenomena (*à la* Thurstone) into quantitative Procrustean beds in order to open the possibility of measuring them.

3. Mathematical approaches using order theory and fuzzy logic exist, can be applied to assessment data in the context of this theoretical framework for measurement, and can be developed further. However, they may not displace the deeply embedded quantitative mindsets, methodologies and practices of educational measurement. So a key practical question that is worth investigating is how much is lost in the reduction of quality to quantity. This will depend on the nature of (the theory of) the construct that is the target of the assessment procedure. For some types of constructs, approximation by quantities may not be too reductive. For others, the difference may be substantial.

Some of the possibilities that flow from this narrative, as well as some of its limitations and vulnerabilities, are considered in the discussion that follows the papers themselves, in Chapter 7.

1.2 Outline

Chapter 2 of the thesis sets out in some more detail the background and surveys some of the literature required to engage with the papers that follow. It provides more than a review, however, as it also develops some of the ideas that are explored in more detail in the papers that follow – and hence is presented as a prologue to those, rather than a separate appendix.

In particular, Chapter 2 distinguishes two ways of conceiving of the nature of the phenomena – the so-called *assessment constructs* – that are investigated by means of educational assessment procedures. Approaches to assessment that derive from psychological testing and measurement tend to regard constructs as (latent) cognitive or mental properties of the people being assessed – such as their proficiency or ability in mathematics, for example. On the other hand, curriculum-related educational assessment is often concerned with constructs thought of as specifications of the particular nexus of knowledge, skills, and understanding that characterise (good) attainment in a domain such as *GCSE mathematics*. The discussion in Chapter 2 shows how these two notions of assessment constructs are linked via the use of a particular psychometric model (the Rasch model) to study proficiency, and the so-called compensatory approach to aggregating judgements of attainment. In the context of high-stakes assessment in the UK, this leads to an elision or equivocation between the two conceptualisations that does not seem to have been much studied in the literature.

Chapter 2 also demonstrates how this equivocation depends on an underlying assumption that constructs, or the separate attributes that together describe or instantiate a construct, can be reasonably regarded as having quantitative structure. It discusses how the representational theory of measurement – one of the most formally developed theories of measurement within the philosophy of science – can provide a basis for testing empirically, in certain cases, whether this assumption is likely to hold: a line of enquiry that is taken up in detail in Chapter 4.

The final topic considered in Chapter 2 is the concept of a fuzzy relational system, and the suggestion that there is value in applying it to modelling educational assessment procedures – to conceptualising them as systems of relationships between learners, on the one hand, and construct-relevant attributes, on the other, in which the relations may be more or less fuzzy as to the extent to which they obtain, depending on the

nature of the assessment domain. Some approaches to studying relational systems, in particular the branch of mathematical order theory known as formal concept analysis, are briefly introduced, in preparation for further application to educational assessment data in Chapters 5 and 6.

Chapter 3 is a brief summary of the methodological and technical background that is not covered in a self-contained way within the papers that follow. The papers address three research questions, namely

1. Empirically, does the assumption of quantitative structure hold for educational assessments of the kind used to underpin the award of high-stakes qualifications (GCSEs and A levels) in England?

2. Conceptually, can the quantitative paradigm of measurement that underpins the psychometric approaches summarised in Chapter 2 be extended to accommodate educational assessment constructs whose measuring procedures generate, in general, fuzzy, partially-ordered data?

3. Can non-quantitative methods of measurement (modelling educational assessments as fuzzy relational systems) provide more valid or useful information about learners' attainment or poficiency with respect to educational assessment constructs than quantitative approaches? In particular, can they help shed light on one of the central problems faced when awarding high-stakes qualifications to learners, namely the maintenance over time of the performance standards required for the award of different grades?

Thus Chapter 3 collects together some background material on mathematical preliminaries, such as operations and relations, especially different types of order relations, on sets. It reviews some key ideas from order theory – in particular, formal concept analysis, that are used in Chapters 5 and 6, as well as a brief outline of key aspects of fuzzy

logic, such as modelling and combining truth-degrees of propositions. It also summarises some relevant aspects of measurement theory that underpin the empirical investigation of research question 1, and the conceptual investigation of research question 2.

Chapter 4 [Paper 1] is the first paper: 'Is educational attainment a quantitative phenomenon? Applying stochastic conjoint measurement checks to UK public examination data'. A presentation based on some of this material was accepted for, and presented at, the 23rd annual meeting of the Association for Educational Assessment, Europe, in Dublin in November 2022.

The paper reports the results of an empirical investigation of the extent to which the quantity assumption is tenable for some high-stakes summative assessments in England. It used a methodology that tests stochastically whether pre-conditions for variables to be validly measured on an interval scale are likely to hold. The results suggest that requirements for quantitative structure were generally not met, although the extent of deviation varied by subject. Some challenges of applying the methodology to assessments with polytomously-scored items are discussed. In the light of the findings, a key area for further research is to compare alternative structural approaches to measurement, that do not assume the assessment construct is quantitative, with existing approaches that are generally (if uncritically) accepted by teachers and students.

Chapter 5 [Paper 2] is the second paper: 'Educational assessment without numbers'. This was published in *Frontiers in Psychology* in October 2024.

It explores the application of van Fraassen's (2008) conceptualisation of measurement to educational assessment. After a brief review of some of the main theories of measurement that have been applied to educational assessment, the paper considers in some detail how the specifics of van Frasssen's approach play out in this context. This entails examining how to construct so-called data models and surface models to describe 'levels'

of attainment or competence in a subject, and what counts as persuasive evidence for assigning a student to a particular level.

It suggests that applying this conceptualisation to the assessment of intersubjectively constructed phenomena, such as a learner's proficiency in an inherently fuzzily-defined domain, entails recognising the theory-dependent nature of valid representations of such phenenomena, which need not be conceived of structurally as quantities. Finally, some connections are drawn between this 'qualitative mathematical' view of educational assessment, and the application of techniques from machine learning and artificial intelligence in the area.

Chapter 6 [Paper 3] is the third paper: 'Beyond latent variable models for non-quantitative constructs: applying fuzzy relational methods to the analysis of examination standards'. This builds in part on work by Bartl, Bělohlávek, and Scharaschkin (2018) that applied fuzzy formal concept analysis to the analysis of assessment data from A level examinations in England. It shows how a methodology for analysing assessment outcomes, thought of as a fuzy relation between learners and construct-relevant attributes, can be seen as an extension of traditional latent variable methods for the analysis of quantitative data. It then uses this methodology to outline a new approach to appraising examination grading standards over time, that directly addresses the question: 'is the *kind of attainment* that students have to demonstrate to be awarded a particular grade *qualitatively equivalent* between different examination versions?'. The approach is illustrated using data from high-stakes examinations in English language and mathematics, and its benefits and limitations are discussed.

Chapter 7 is a discussion of key issues raised by the preceding papers, with suggestions for areas where further research could build on the ideas they explore, both empirically and conceptually. It also sketches some potential connections to wider issues of measurement,

theories of logic and structure, and the application of large language models, that draw on vector-space approaches to semantics, to educational assessment.

2 Background and literature survey

2.1 Educational assessment

Educational assessment, broadly conceived, is the discipline concerned with evaluating how well learners are performing with respect to educational aims and objectives. Educational assessment in practice covers a wide range of formal and informal procedures, interactions, and methods. Whenever a teacher observes how a student performs or reacts to an explicit or implicit task or activity, for instance, and decides what interaction to have next with that student, the teacher is carrying out an assessment procedure.

Usually an educational assessment procedure focuses on a specific *domain, subject,* or *curriculum area* (such as 'basic numeracy', 'French', or 'art'), and its purpose may range from providing feedback to help students progress, to awarding qualifications, to holding schools to account for the quality of their teaching (see Newton, 2007, for an analysis of the multiple purposes of educational assessment). Wiliam and Black (1996, p. 540) define educational assessment as a process that requires that 'evidence of performance or attainment is elicited, interpreted, and acted on in some way.'

Familiar examples of educational assessments are tests and examinations. These are assessment procedures that elicit evidence of performance or attainment by giving students certain tasks to do, and then using information obtained from their responses to

summarise the quality of those responses. This summary information is often said to *measure* some feature of substantive interest, such as 'attainment', or 'proficiency', and is often expressed numerically, for example as a score.

2.2 Public examinations as educational assessment procedures

This project takes, as case studies of educational assessment procedures, English public examinations (the externally set and marked GCSE and A level examinations, taken by pupils in England at the ages of 16 and 18 respectively). The tasks that students are required to do in these examinations may range from selecting the correct answer to a multiple-choice question, to creating a relatively extensive essay, artwork, or musical performance.

Public examinations are *curriculum-related* assessments. For each subject in which a GCSE or A level qualification is available, the government specifies what *content* shall be covered, as well as what the so-called *assessment objectives* shall be. The latter articulate what are deemed to be the valued characteristics of how learners engage with the subject content. For example there are five assessment objectives for A level English literature (Ofqual, 2017), including 'analyse ways in which meanings are shaped in texts' and 'explore connections across texts, informed by linguistic and literary concepts and methods'.

Taken together, therefore, the subject content and the assessment objectives jointly specify what is deemed to constitute *quality of educational attainment*, for the subject in question (Scharaschkin, 2017). The (more or less fuzzily articulated) criteria for determining 'what a good examination performance looks like' – or for discriminating between

different levels or kinds of relatively better or worse performances – are intended to be anchored in the content and assessment objectives. Assessment criteria that are clearly derived from the subject content and assessment objectives are said to be *construct-relevant* (Pollit et al., 2008). An examination may be said to have a high degree of *construct validity* (Messick, 1989; Newton and Shaw, 2014) to the extent that its method of discriminating between, classifying, or grading students values precisely those characteristics of students' performances that the assessment designers intended it to value.

The awarding organisations that provide accredited GCSE and A level qualifications in England are required to provide detailed rationales to the exams regulator Ofqual as to how their assessment procedures will ensure that what counts as creditworthy in candidates' responses to assessment tasks properly reflects these specifications of the content and assessment objectives for each subject (Ofqual, 2022; 2017).

2.3 The information elicited in educational assessment procedures

2.3.1 Tests and examinations

Assessment procedures such as examinations generate information about each learner. For example typically a mark or score is assigned to learners' responses to each item (task or question) in an examination. The mark may be generated automatically, for instance in the case of items that require candidates to choose the correct answer to a multiple-choice question; or it may be generated by means of expert human judgement, as in the case of items that require candidates to write a description or explanation of something, or write out a mathematical derivation or proof, or draw a diagram. The latter types of items (the most common in English public examinations), known as *constructed response*

items, are usually marked by assessors who have been trained to apply a classification schema known as a *mark scheme* or *rubric*.

Bramley (2011, p. 30) noted that for educational assessments of this kind,

> It has become so natural for us to think of the data arising from testing as quantitative that we can sometimes lose sight of the fact that the "raw data", as it were, usually consists of written answers to written questions. Where do the numbers come in? The mark scheme can be thought of as a coding scheme that assigns numerical values (usually integers) to examinee responses according to a certain rationale

Mark schemes give instructions to assessors about how to classify the response to each item into a specific ordered category. For example, the possible categories for the response might be {`correct, incorrect`}, or {`5-mark-response, 4-mark-response, ..., 0-mark-response`}, or {`level 4, level 3, ..., level 0`}. The instructions for classifying responses are typically

- **Rules-based** (by means of *necessary and sufficient conditions* for a response to fall into a particular category), or
- **Pattern-based** (by means of *best fit* to exemplars, descriptions, or protypes of the possible categories of performance quality).

Applying mark-scheme instructions to students' responses to an examination leads to the *formation of concepts* of *value* or creditworthiness: what kinds of performances count as 'pass/fail', or 'worth-one-mark/worth-two-marks/worth-three-marks', etc. This process of intersubjective concept formation is called *marker standardisation* by the organisations that develop and administer public examinations.

Rules-based and pattern-based methods of classification correspond to different cognitive theories of concept formation. Rules-based classification methods correspond to the so-called *classical theory* of concepts, and pattern-based methods to the *prototype theory* (Machery, 2011; Laurence and Margolis, 1999; Hampton, 2006). Rules-based mark schemes are often used for closed-form or selected-response tasks, and pattern-matching schemes for open or constructed-response tasks, as shown in Figure 2.1.

These intersubjectively constructed concepts of value or creditworthiness are often 'inexact' or fuzzy: but that does not preclude the possibility of being able to reason with them (Goguen, 1968), and to use them to support inferences about the quality of learners' educational attainment. Chapter 6 of this thesis explores in detail the mathematical and logical analysis of the information generated by examinations, where candidates' responses to items are derived from mark schemes applied by human experts, separately from the question of mere manipulation of the item marks as if they really were *numbers,* rather than *labels* for different kinds of performances. Chapter 7 discusses briefly the relationship between such analyses and theories of concept formation, as well as some of the implications of the development of large language models in machine learning for the application of broad valuation criteria expressed in natural language to students' responses to assessment tasks.

The results of the assessors' application of the mark scheme to a student's response is thus a *valuation* (an ordinal classification) for each attribute of performance that is covered by the mark scheme (each assessment item). If there are m performances and n attributes, then the information generated by the examination can be represented as an $m \times n$ matrix of ordinal values, in which entry (i, j) is the valuation of the ith performance with respect to attribute j – in the usual terminology for examinations, the mark that candidate i got on item j.

Figure 2.1: Generating outcome data from assessment tasks

Rules based
Question: 'What is 274+65?'
Mark scheme: `If` response='339' `then` classification='1 mark'
`else` classification='0 marks'.

Pattern-based
Question: 'Discuss how Shakespeare presents the relationship between Othello and Desdemona.'
Mark scheme: `If` response `definitely-has` attributes {'assured argument', 'perceptive relation to historical context', 'relevant use of extracts from text', ...} `then` classification=level 5
`else if` response `almost-has` (these attributes) `then` classification=level 4
...
`else if` response `definitely-does-not-have` (these attributes) `then` classification=level 0.

2.3.2 Assessment procedures more generally

2.3.2.1 Generating information about ordered attributes

The structural aspects of examinations as information-generating procedures summarised in the preceding section apply to educational assessment procedures more generally. Modes of assessment can vary widely. Assessment may involve observation, or interviews, or learners engaging with gamified activities, or portfolios of artefacts being appraised, for instance. In public examinations in England candidates are asked to repond to questions such as 'To what extent was England's government fundamentally transformed in the years 1509 to 1547?', or 'Show that the exact value of $\int_0^{\sqrt{3}} x \tan^{-1}\left(\frac{x}{\sqrt{3}}\right) \mathrm{d}x$ is $p\pi + q$,where p and q are rational numbers'. But in all cases (consonant with Wiliam and Black's (1996) definition of assessment as a process in which 'evidence of performance or attainment is elicited, interpreted, and acted on in some way'), the learners being assessed are (explicitly or implicitly) assigned tasks to do; their responses are collected

or observed; and certain attributes (features, properties, or characteristics) of those responses are extracted and evaluated.

For instance an attribute of interest might be the time taken for a learner to complete a task (e.g. in a physical education assessment), or the number of attempts taken to solve an problem presented in a video game. There may be syntactic or semantic features that can be extracted from written responses to tasks that are of interest, such as relative richness of vocabulary based on standard word-frequency information, or lengths of answers. In the case of tests or examinations made up of questions, the attributes are normally the results (marks, scores) of the examinees on the items.

A fundamental feature of the attributes that are studied in educational assessment is that they have *ordered* values, reflecting a notion of ***betterness*** that obtains between responses to tasks. Thus if there are only two values for an attributes, such as {`incorrect`, `correct`}, then the correct responses are regarded as better than the incorrect responses, with respect to the attribute in question. If there are five values, which for an examination question would normally be labelled by integers drawn from $\{0, 1, 2, 3, 4\}$, then the responses that were classified '3' (scored three marks) are regarded as better than those classed (scored) as '2', which in turn are better than those classed '1', etc.

Normally – and certainly in examinations such as those considered in this project – one sees *total orders* on the value-sets for each attribute (that is to say, the possible values for an attribute can be *ranked*); but in the general case they could be *partially ordered*[1].

[1]See Chapter 3 for definitions of total and partial order relations. In essence, when objects are totally ordered with respect to a particular feature, they can be ranked with respect to that feature: it is always possible, given any two objects, to compare them. When objects are partially ordered, there may exist pairs of objects that are not directly comparable, and the objects cannot necessarily be placed in a single linear sequence (a ranking) with respect to the feature of interest.

In practice objects often have several different features of interest, as in the case of assessments where learners respond to a number of different tasks. If objects can be totally-ordered with respect to each feature, then in general that generates an overall partial order, in respect of how the objects can be compared with respect to the totality of features of interest. A question of interest in educational assessment is then: when can such an overall partial order be approximated or replaced by a total order, i.e. under what circumstances is it valid to summarise the information collected about the

There may also be additional structure, over and above ordering, to the value-sets for attributes in some cases (for instance, 'time taken to respond' is not only an ordered attribute, it is a *quantity*). To the extent that there *is* more structure, that structure may be used to support valid inferences from the assessment. However, as discussed further below in §2.4.5, it is necessary to establish whether the assumption of additional structure (for example, that outcomes are not only ordered, but quantitative), before being warranted in using approaches to the analysis of assessment information that take such structure for granted. Chapter 4 of this thesis examines empirically this question of the structure of attributes using GCSE and A level examination data.

2.3.2.2 Information as truth degrees

If m learners take an assessment in which the n attributes of interest are all *dichotomous* (that is to say, their value-set consists of only two possible outcomes, such as {0,1}, or {fail, pass}: this is the case, for instance, if the assessment is a test made up of n multiple-choice questions), then the $m \times n$ matrix that holds the information generated by the assessment is a tabulation of 0s and 1s. Entry (i, j) in the matrix is 0 if learner i answered item j incorrectly; it is 1 if they answered the item correctly. Now, as Michell (2009, p.45) notes,

> Tabulated numbers are shorthand for a set of propositions that tell where the numbers came from. Furthermore, deductions from a data set are inferences from those propositions.

In the case of an assessment consisting of dichotomous items, the matrix entry (i, j) provides the *truth value* of the proposition 'learner i answered item j correctly'. If

different aspects of performance of candidates in an assessment by a simple ranking?

matrix entry (i, j) is 0, then the proposition 'learner i answered item j correctly' is false. If the (i, j) entry is 1, then the proposition is true.

As discussed further below in §2.4.4, and examined in more detail in Chapters 5 and 6 of this thesis, the claim that a learner's response to an assessment task demonstrates a certain attribute at a certain level (for example, to assert that it is 'worth' 3 out of a possible 4 marks), can be modelled as a claim about the truth value of a proposition. In this case, however, it is helpful to go beyond the restrictions of traditional (Boolean) propositional logic in which the only possible truth values that a proposition can take are 'true' or 'false'. It has long been recognised that bivalent logics can be unduly restrictive for modelling situations in which there is inherent fuzziness, vagueness, or uncertainty (see e.g. Goguen, 1969; Goertz, 2006; Belohlavek, Dauben, and Klir, 2017). *Fuzzy logic* (Hajek, 1998; Belohlavek, Dauben, and Klir, 2017) allows propositions to have truth values drawn from ordered sets of *truth degrees*, that can be more extensive than just {`false, true`}. For example, possible sets of truth degrees could be {`false, partly-true, true`}, or {`clearly-false, partly-false, partly-true, clearly-true`}. To be used in models of logics, sets of truth degrees need some ordinal structure (for example, the sets in the preceding two sentences are listed in order of 'increasing true-ness')[2].

It is suggested in this thesis that the possible levels or values of an attribute can be regarded as providing information about the truth degree of propositions concerning that attribute. In particular, suppose attribute j has three levels (such as {0,1,2} or {fail, pass, merit}). Then inferences can be drawn regarding propositions with three degrees of truth (they could be labelled {`false, middle, true`}). If learner i's response to the assessment is classified as 0 on attribute j, then the proposition 'learner i displayed attribute j' is false. If learner i's response is 1, then the proposition is not false, but

[2] As well as requiring an ordinal structure, truth degrees also require a compositional structure, to allow truth degrees to be logically combined or aggregated across propositions. This is discussed further in Chapter 3, and in Appendix 1 of Chapter 6.

neither is it fully true: it is partly true, or true to an intermediate degree. If the learner's response is 2, then the proposition is (fully) true.

2.3.2.3 Fuzzy relational systems

The preceding discussion provides the background to the general framework used in this thesis to conceptualise the information generated by assessment procedures. Item marks or scores in examinations or tests are a specific case of this more general framework. Namely, assessments provide information about certain attributes of interest derived from learners' responses to assessment tasks. As discussed in §2.4.2, an important aspect of assessment design is ensuring that the attributes about which information is collected are *construct-relevant,* that is, that collectively they support inferences about the overall property or phenomenon that the assessment procedure is intended to measure.

In an assessment context with m learners and n tasks, the information can be collected in an $m \times n$ matrix, in which each row represents a learner and each column an attribute. Thus assessments may produce information such as

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix}, \text{ or } \begin{pmatrix} A & B & C & C & B \\ C & C & A & B & C \\ A & A & B & B & C \end{pmatrix}, \text{ or } \begin{pmatrix} 7 & 5 & 3 \\ 5 & 2 & 1 \\ 10 & 5 & 5 \\ 6 & 3 & 0 \end{pmatrix}, \text{ etc.}$$

In general, a matrix of ordinal-valued entries of this kind can be taken to represent a *fuzzy relational system*, (Belohlavek, 2002) in which entry (i, j) represents the (ordinal, or partially ordered) *degree to which* student i's response to the assessment tasks displays attribute or characteristic j.

Equivalently, the matrix can be taken as representing a (fuzzy) *binary relation* between learners and (construct-relevant) attributes[3]. If the (i, j) entry is k, then learner i's performance is related to attribute j to the extent k. We normally call the binary relation between performances and attributes 'having' or 'possessing', and say that learner i's performance *has* attribute j, to the extent k.

In the special case of assessments with dichotomous attributes, there are only two possibilities for k, and so, for each i and j, either learner i's performance *has* attribute j, or it *does not have* attribute j.

Although it has been applied to other areas of social sciences, as yet this lens for conceptualising and analysing assessment information has largely not been considered in the educational assessment literature. Some recent examples of its application are discussed briefly in §2.6, and in §5.4 of Chapter 5. But the overwhelmingly most common approach to appraising data generated by assessments is based in the psychometric tradition, which is outlined next.

2.4 The quantitative tradition and psychometric approaches to educational assessment

2.4.1 The emergence of psychometrics

Psychometrics is a subfield of psychology concerned with the definition and measurement of mental or psychical constructs, such as anxiety, intelligence, memory, compulsiveness, extraversion, etc. It originated in the work of Francis Galton (1822-1911), Karl Pearson

[3]See Chapter 3 for full definitions and more detail on binary relations.

(1857-1936), and Charles Spearman (1863-1945) in the late nineteenth and early twentieth centuries. Galton's main interest, heralded in his book *Hereditary Genius* (Galton, 1869), was eugenics, including the identification of physical characteristics of 'men of genius' (Gould, 1996). He set up the Department of Eugenics at University College, London, endowing a Chair of Eugenics there on his death (later converted into a Chair of Statistics).

Partly to assist them in pursuing their aims in eugenics, he and his students and followers developed much of the early machinery of statistics as a discipline. Galton introduced the ideas of standard deviation and regression; Karl Pearson, who became Professor of Eugenics in 1911 and was the editor of *Biometrika* from 1902 to 1936, invented the correlation coefficient; and Charles Spearman developed factor analysis.

Spearman's 1904 paper on the measurement of general intelligence (Spearman, 1904) is perhaps the seminal document in the study that evolved into psychometrics as currently practised[4]. In the terminology of the previous section, Spearman's information matrix consisted of scores on various assessment tasks for English school boys aged 9-13. For example he analysed a 22×6 matrix of scores, for 22 pupils tested on classics, French, English, mathematics, pitch discrimination, and music, and concluded that a common or general factor (the 'g factor') of intelligence could be deduced from it.

Spearman's approach was to use the newly-invented correlation coefficient to calculate what would now be called the covariance matrix between the different attributes (tests), and then to argue for the existence of a latent common factor that would explain the pattern of positive correlations calculated from the scores (the latter – although derived in a mix of ways from teacher judgements, written tests, and tasks – being assumed

[4]There is also a strand of evolution that developed from nineteenth century work on psychophysics, with its attempts to measure perceptions of physical phenomena such as frequency and intensity of sound and light: see for example Mosier (1940, 1941). A comprehensive account of the work of Galton, Spearman, and other key figures in the development of psychometrics and measurement practice in the human sciences more generally is given by Briggs (2022).

to be numerical quantities that could be manipulated accordingly). Building on later developments in Spearman's work, Thurstone (1931, 1947) generalised the methodology to provide what became the basis for the modern technique of factor analysis, and related 'latent variable' methods such as structural equation modelling. This broad approach – mathematically based on factorising in certain ways the covariance matrix derived from the assessment information matrix, treated as a collection of quantitative values – has been the workhorse of psychometrics ever since. For instance one of the most important psychometric models, the Rasch model that is discussed further below, can be simply derived from applying factor analytic methods to binary data (Bartholomew et al., 2011: see also Appendix 2 of Chapter 6).

The following sections briefly review the psychometric approach, and the extent to which it is based on the assumption that the phenomena that educational assessment procedures aim to measure are *quantities* (rather than having some other mathematical structure, for example partially ordered sets or networks). Objections to the uncritical reliance on the assumption of quantitative structure are summarised. The question of what *measuring* a (quantitative) phenomenon entails is discussed in relation to one prominent theory, the representational theory of measurement. This leads to a brief indication of how theorems from the representational theory of measurement can be used to examine empirically the question of whether assumptions about quantitative structure of educational assessment constructs are valid.

The chapter concludes with an introduction to the literature on some non-quantitative ('qualitative mathematical') approaches that can be applied to analysing, and drawing inferences from, the information collected in educational assessments viewed as fuzzy relational systems.

2.4.2 The assessment construct, and its relation to items or assessment tasks

2.4.2.1 Constructs

Andrich and Marais (2019, p.3) note that the terms *property, trait, construct, attribute,* and *variable* are used variously in the educational literature, sometimes as loose synonyms, though sometimes with different nuances in different contexts, as ways of describing the phenomena that assessment procedures aim to measure. In accordance with common usage in the context of UK public examinations, the term *construct* will be used henceforth to mean 'what is assessed' in a GCSE or A level examination. Thus examples of possible *assessment constructs,* in this context, include 'GCSE mathematics', and 'A level German'. As noted in §2.2, these constructs are defined in terms of particular subject content and assessment objectives, that are chosen by the government, and set out in the *specifications* for each GCSE and A level published by the exam boards. The specifications thus aim to state (in terms that are sufficiently clear and informative, but not reductively prescriptive), what precisely, from the general possibilities entailed in domains such as 'mathematics', or 'German', will actually constitute the specific nexus of knowledge, skills, and understanding that defines 'GCSE mathematics', or 'A level German', for the time being.

Baird et al. (2017, p.318) reiterate that the term *construct* is normally used in the UK examination context to mean 'the subject matter or domain being assessed', but also note the tension with the definition given by Cronbach and Meehl (1955, p. 283) – in a paper that was foundational for subsequent work on *construct validity* – that a (psychological) construct is 'some postulated attribute of people, assumed to be reflected in test performance'[5].

[5]Much of the discussion of the actual practice of construct validation in Cronbach and Meehl (1955) is couched in terms of standard quantitative techniques of the time (in particular, with respect to

The two conceptualisations of *construct* are often connected by an appeal to hypothesised *latent variables.* Andrich and Marais (2019, p. 4) state that 'the observations from an assessment, often referred to as *responses*, are qualitative. However, as a step towards measurement, these responses have an order which immediately implies more or less of the property to be assessed.' In essence, if the construct, in the first sense, that is being assessed is *GCSE mathematics*, then that consists of specified content and objectives. A good performance on an assessment of this construct is one that demonstrates a high level of quality with respect to the assessment objectives as applied to to subject content[6]. The claim summarised by Andrich and Marais is then that the relation *better than*, in terms of quality of performance, entails a relation *more of*, in terms of an unobserved (latent) property of the examinees – the second sense of *construct* as 'some postulated attribute of people'. In fact, using the term *more of* may also seem to suggest that this postulated property of people can be taken to be a *numerical quantity* – a stronger claim than that the property can be taken to be an *ordered structure,* for instance.

The connection between these two notions of assessment construct is usually taken to be a *causal* one. The variation in students' GCSE mathematics performances, on this account, is caused by variations in their *mathematical proficiency.* Students with more (larger amounts of) proficiency tend to do better on the examination. The examination has been designed to allow exam boards, through the award of a grade, to attest to a student's demonstrating a certain *level of attainment* (such as a 'grade A') with respect to the assessment construct in the first sense. But the information generated by the examination is also used to draw inferences about the latent variable *mathematical proficiency* –

establishing propositions of the form 'x per cent of test variance is accounted for by the construct': p.294). But is interesting to note that the authors do not equate constructs with quantities. They say (p. 283, emphases added) 'we expect a person at any time to possess or not possess a *qualitative attribute ... or structure*, or to possess some *degree of a quantitative attribute* '.

[6] Scharaschkin (2017, p. 455), commenting on Baird et al. (2017), notes that 'implicit in the articulation of any particular construct is a network of relationships, namely notions of "better/worse" with respect to construct-relevant attributes. The construct-relevant attributes are the features that define what is valued, or important, in the subject area: the properties that any assessment of the construct is supposed to value'.

indeed, in many cases, to calculate an estimate of 'how much' proficiency each student has (or had, at the time of the examination).

Goertz and Mahoney (2012) suggest that these two ways of thinking of constructs are aligned with qualitative and quantitative research methodologies in the social sciences. They conclude that

> Qualitative researchers adopt a semantic approach and work hard to identify the intrinsic necessary defining attributes of a [construct]. Quantitative scholars adopt an indicator-latent variable approach and seek to identify good indicators that are caused by the latent variable.

The question of whether the 'quantitative latent variable' approach is convincing as an account of what it means to *measure* educational phenomena is examined in Chapter 5 of this thesis, as part of a wider conceptual examination of how developments in the philosophy of measurement more generally can support a 'qualitative mathematical' (Rudolph, 2013) approach to educational assessment, with the potential to provide new insights into assessment design and validation.

2.4.2.2 Attributes

Henceforth the term *attributes* will be used to refer to the specific aspects, features, or characteristics of learners' responses to tasks, in a particular assessment context, about which information is generated by the assessment procedure. For example each item on an examination provides an attribute of examinees' performances, with the possible ordered values for the attribute being labelled by the possible marks for the item.

Collectively the attributes are supposed to represent, capture, or instantiate, the assessment construct. In order to provide information that is useful for measurement, the

attributes must differ from each other, but not 'too much'. They must all be construct-relevant and must 'cohere' in how they work together to provide appropriate information to *locate* a learner with respect to the construct being assessed. Measurement, in educational assessment, is often taken as requiring location on a continuous line (an interval scale), for example, via the calculation of the value of an hypothesised quantitative latent variable such as 'mathematics proficiency' (Michell, 1990; 1994; 1997). It is argued in Chapter 5 of this thesis that, in fact, a broader notion of location, namely what van Fraassen (2008) calls *location in a logical space,* is more appropriate for conceptualising the idea of measurement in the context of educational assessment. But whether one takes a narrower or a wider view of the meaning of 'location', the requirement holds that the attributes selected to provide the information necessary to measure learners with respect to the assessment construct must, in some sense, cohere. They must capture an appropriate range of different aspects of, or viewpoints on, the phenomenon of interest. This is essential for the construct validity of the assessment procedure. Indeed it is hard to see how most of the kinds of inferences about learners (Newton, 2007), that may be drawn from the results of the assessment (learners' scores or grades), could be valid, if the assessment was merely 'about' an unconnected collection of properties, rather than some kind of unified construct.

The question of the coherence of the attributes investigated in a given assessment context is most commonly studied by calculating correlation coefficients between them (taking them to be numerically-valued properties). Thus the covariance structure of the attributes becomes a key area of interest, and techniques from linear algebra (for example relating eigenvalues of the covariance matrix to possible factorisations of the data matrix) are used to examine the 'dimensionality' of the assessment construct as reflected in the collection of attributes about which data has been generated by the assessment procedure (and, if possible, the location of learners on an underlying quantitative continuum). As Andrich (1985, p.34) puts it, '[t]he main emphasis in traditional test theory is on

the empirical evidence that items (through their correlations and derived indexes such as factor scores) work in the same direction.' Essentially this is a logical evolution of Spearman's (1904) method of using observed patterns of correlation coefficients to argue for properties of an underlying quantitative latent variable (the g factor). It corresponds to the 'quantitative' view of the assessment construct as a trait or property of people, discussed above.

Another style of argumentation is possible, however: what Guttman (1971) called a *structural* one. Rather than focusing on the covariance structure of the attributes, it examines whether structural features of the binary relation between learners and attributes may support inferences about locating learners with respect to the assessment construct. A simple, but important, idealised case is discussed in §2.4.3.1. More general structural approaches are discussed in §2.6. These correspond more closely to the 'qualitative' view of constructs as 'what is assessed'.

2.4.3 Assessment responses with Boolean (dichotomous) attributes

The assessment procedures that are the simplest to study are those in which each attribute is dichotomous, i.e. can take only one of two ordered values (such as 0< 1, or false<true, or fail<pass, or 'does not display quality Q' < 'displays quality Q'). For example, tests consisting of a sequence of multiple-choice items are of this kind. The information generated by such a test is a matrix whose entries are 0s and 1s. The matrix displays a binary relation R between students S and items A ($R \subseteq S \times A$), such that student i is related to item j (i.e. $(i, j) \in R$) iff[7] student i got item j correct. 'Student i is related to item j' is often written iRj rather than $(i, j) \in R$.

[7] The abbreviation 'iff' is used throughout to mean 'if and only if'.

One possible theory that could explain the observed pattern of 0s and 1s generated by such an assessment procedure is that the truth-value of the proposition 'student i gets item j correct' depends on (unobserved, but inferred) properties of students (their *proficiency*), and of the items (their *difficulty*). These properties have ordered values, so that students can be ranked by ability, and items by difficulty. For example this assumption is made in the Rasch measurement methodology (Rasch, 1960/1980; Andrich, 1988), that underpins important educational assessments such as the international PISA tests that are commonly used to compare countries' educational outcomes.

(Student) proficiency and (item) difficulty, under this theory, are *dual* properties in the sense that each must be defined in terms of the other. A student's proficiency is high if she tends to answer difficult items correctly. An item's difficulty is high if it tends to be answered correctly only by proficient students. Proficiency and difficulty are two sides of the same coin. Andrich and Marais (2019, p. 79) go so far as to say that '[t]his same property must *reside in* the items and in the persons' (emphasis added). As will be argued in chapter 7, this is a specific instance of a more general duality between *intensive* and *extensive* ways of measuring educational attainment.

How are these hypothesised properties of students and items derived from the information collected via the assessment procedure? In the case of data that exhibits a so-called *Guttman pattern*, this is straightforward.

2.4.3.1 Guttman structures

Guttman (1940, 1950, 1971) studied the data structure that arises when the binary relation R between students and items in a test is what is known as a *biorder* (Ducamp and Falmagne, 1969). A relation from S to A is called a biorder iff for all $a, b \in S, x, y \in A$, it is the case that $aRx \,\&\, bRy \Rightarrow aRy \,\text{or}\, bRx$. This definition expresses the intuition that

the only way for one student to get a higher total score on the test than another student, is if the first student gets all of the same questions correct as the second student, plus at least one further question. A biorder guarantees a staircase pattern in the data matrix, so that the outcomes from assessments in which students and items are biordinally related look like

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \text{ or } \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \text{ etc.}$$

If a total score is calculated for each student, as the number of items answered correctly, then there is a one-to-one correspondence between the possible mark profiles (types of performances displayed by students), and scores. Let $\#A$ be the number of attributes in A. Then there are precisely $\#A + 1$ possible levels of proficiency, corresponding to precisely $\#A + 1$ possible distinct mark profiles. There is a 'bottom' performance – with a score of zero – in which a student displays none of the attributes. There is a 'top' or 'mastery' performance, corresponding to a score of $\#A$, in which all of the attributes are demonstrated. For biordered data, the intermediate kinds of performances, between 'no evidence of attainment' and 'mastery', are totally ordered: they form a ranking.

So if an assessment generates biordered data, one can talk about the assessment as measuring a single construct – all the attributes cohere in how they describe the construct. It is an ordinal (totally ordered) property of students' performances: one can say that students are characterised by their (ordinal) proficiency, or equivalently, in this case, their ordinal level of attainment. Likewise (dually) items are characterised by their

(ordinal) difficulty (there are $\#A$ possible levels of difficulty) with respect to this single construct.

In Guttman's terminology (in which he calls constructs 'concepts': Guttman, 1944, p.141), '[t]he universe is the concept whose scalability is being investigated . . . The universe consists of all the attributes that define the concept . . . it consists of all of the attributes of interest to the investigation which have a common content'. By *scalability* he means 'possibility of unambiguously characterising learners as to their level of attainment (or proficiency) in respect of the construct'. Andrich (1985, p. 37, emphasis in original) notes that Guttman's approach is underpinned by 'the desire to classify persons into an unequivocal order on the dimension in question. Guttman's requirement for such an ordering is perfect transitivity: If object A is deemed to have a greater value than object B, and B is deemed to have a greater value than C, then A should be deemed with *certainty* to have a greater value than C'.

Guttman does not use the term 'latent variable', and it does not seem necessary to invoke it for biordered data: because of the one-to-one correspondence between *ordered levels of proficiency*, and *ordered kinds of performances* (i.e. mark profiles: the different ways in which learners' responses to the assessment tasks can display the attributes that embody the construct). For biordered data, the two different notions of *construct* are equivalent for the purpose of measurement, i.e., for the purpose of *locating* a student with respect to a level of proficiency, or a level of quality of performance.

However, whenever assessments composed of dichotomous items do not yield biordered outcomes (in practice, almost always), the divergence between 'construct as what is assessed' and 'construct as a latent property of those people being assessed' becomes clear, as discussed in the following section.

2.4.3.2 The Rasch model, 'probabilistic Guttman', and 'compensation'

Modelling latent proficiencies Georg Rasch (1960/1980) developed a model for how students engage with test items. It is grounded in the notion of an assessment construct being a latent property (usually called *proficiency* or *ability)* of the people being assessed, rather than a representation of what is assessed. It assumes that such latent properties can be regarded as numerical quantities. For tests composed of dichotomous items, the Rasch model says that the likelihood (expressed as the natural logarithm of the odds) of a student with proficiency θ correctly answering a question with difficulty δ is simply $\theta - \delta$. If a student has more proficiency than an item has difficulty, the student is likely to get the item right, with the likelihood increasing to the extent that the student's proficiency exceeds the item's difficulty. Similarly, if a student has less proficiency than the item has difficulty, she is likely to get the item wrong.

If probability, rather than log-odds, is used as the measure of likelihood, then the Rasch model for dichotomous items is equivalently expressed as

$$P(x_{ij} = 1) = \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)}, \tag{2.1}$$

where x_{ij} is the response of student i to item j (0 meaning 'incorrect', and 1 meaning 'correct'), θ_i is the proficiency of student i, and δ_j is the difficulty of item j.

Thus, in this account of the interaction between students and test items, students are characterised (measured) by their proficiency values – real-valued quantities – with respect to the property being assessed (and dually, items are characterised by their difficulties).

A fundamental property of the Rasch model is that *total score* on the test (calculated for each student as the number of items she got right) is *sufficient,* in the statistical sense, to estimate the proficiency parameter θ (see, e.g., Andrich, 1988; Andrich and Marais 2019). That is, if the Rasch model is fitted to the data resulting from the application of a test to a collection of students, then the estimated values of θ that are obtained for each student, using a maximum likelihood estimation procedure, depend only on the *number* of items the student answered correctly, and not on the student's *profile* of correct responses across items. For a three-item test, for example, all students scoring a total of two marks would receive the same proficiency score, irrespective of whether their mark profiles were (1,1,0), or (1,0,1), or (0,1,1).

Andrich (1985, p.74) argues also that 'the Rasch model is the natural probabilistic counterpart of the deterministic Guttman scale', because (*i*) the Rasch model entails one location parameter for each person, and one for each item; (*ii*) the probability of a given response pattern, given total score, does not depend on person location; and (*iii*) the Guttman response pattern has the highest probability of occurring for each total score. Michell (2014) argues that the Rasch model is a probabilistic version of the Guttman *model* (not 'scale' or 'data structure', but rather the supposed *law* – which results in biordered outcomes – for whether a person of a given proficiency correctly answers an item of given difficulty, namely that they answer it correctly if $\theta \geq \delta$, where θ is the person's proficiency, and δ is the item's difficulty)[8].

Marking test responses Let us now turn from a consideration of characterising students with respect to an assumed latent *amount of proficiency,* to characterising them

[8] Another, less commonly used, probabilistic development of the Guttman model is Mokken scale analysis (Mokken, 1971; Wind, 2017). This is essentially a non-parametric version of the Rasch approach, in which it is not assumed that the item response functions (the maps between the values of the assumed quantitative latent proficiency variable, and the probability of a correct response, for each item) take the form of equation 2.1, merely that they are monotonic. Usually as well, it is assumed that the item characteristic curves (the graphs of the item response functions) do not intersect.

with respect to the *kinds of performances* they have displayed on an assessment (i.e. mark profiles: the different ways in which learners' responses to the assessment tasks can display the attributes that embody the construct). In other words, let us now consider the question of measuring the quality of students' performances on an assessment of a construct thought of as a particular specification of knowledge, skills, and understanding. In practice, when the assessments are carried out using tests or examinations, this question is addressed through the process of *marking,* as discussed in §2.3.1, usually combined with some method of *grouping* or categorising the resulting collection of mark profiles. The most common method of grouping is *summing marks.* This entails deciding that the operation of summing marks defines an equivalence relation on the set of all performance profiles (all qualitatively distinguishably different kinds of performances, given the marking criteria), such that any two performances with the same total score represent the same level of performance quality.

Totalling marks (or calculating a linear combination of marks, if there is the intention to 'weight' certain parts of the assessment more highly than others when categorising response quality: see §2.4.4) is quite fundamental to how examinations are normally conducted. In the UK context, awarding organisations usually describe it as *compensation*, or the application of a *compensatory model* of assessment. For example Newton (2018, p.7), emphasis in original) says that

> there is no expectation within GQs [general qualifications] that precisely defined criteria need to be satisfied for the award of any grade, ie they are **compensatory**, meaning that a high level of proficiency[9] on one outcome can compensate for a low level of proficiency on another.

This is by contrast to *mastery models,* in which the idea is that all of a number of at-

[9] Newton notes (p.7) that he uses the terms 'proficiency', 'competence', and 'attainment' in this paper 'in essentially the same way'.

tributes have to be demonstrated. So in the compensatory approach to the assessment of students' attainment against an assessment construct, thought of as specification of intended learning outcomes, there is a mapping between performance profiles and total scores. In contrast to the situation with biordered data, this is not a one-to-one correspondence, but rather a many-to-one correspondence. To each type of performance there correspondents a unique total score: but to each total score there may correspond more than one type of performance: more than one 'route' to that score[10].

The fact that total score assumes a fundamental role in both the process of categorising students' qualitative assessment responses (in the compensatory approach), and the process of estimating values of an assumed latent proficiency (if the Rasch model is assumed), means that it is possible to elide the distinction between the two ways of thinking about assessment constructs discussed in §2.4.2. The total mark that defines a student's level of attainment also defines their level of proficiency[11].

But it is necessary to assume an underlying quantitative structure for both of these approaches to be valid (and hence also for the possibility of equivocation between them to exist). For the Rasch approach to measurement, this is because the parameters to

[10] In fact, the UK qualifications regulator Ofqual recognises there are different routes to a mark (it has tended to focus on assessments that offer a degree of choice to students as to which questions they attempt, for example Ofqual (undated); but the point applies even when there is no explicit optionality in an examination). Exam boards are enjoined to ensure that different routes to a mark should be equally demanding (or at least not lead to different likelihoods of gaining a particular grade). Quite how to gain assurance on this point, without question-begging, is not entirely straightforward. Scharaschkin and Baird (2000) demonstrated that different performance profiles with the same total mark can be associated with different levels of performance, according to expert judgement. But if there is no option other than to use sum-scores for the purpose of classifying examination responses, then capturing, or eliminating, such judged differences in the quality of performance associated with different routes may present a problem.

[11] This elision is illustrated in practice in the National Reference Test (Ofqual, 2019) in England – an annual test taken by a sample of students in school year 11, that is intended to detect 'changes in performance over time' in English language and mathematics. It is used to inform the awarding of GCSE examinations in those subjects, including, if necessary, directing awarding organisations to change grade boundary marks they would otherwise set based on their usual sources of evidence. The question of whether there have been changes in performance is answered, according to the National Reference Test, by examining whether their have been changes in the proportion of students with certain estimated levels of proficiency, derived from an item response model.

be estimated are assumed to be real numbers; for compensatory classification because it requires that attributes with ordinal values can be manipulated (added, weighted) as if they were numbers. As discussed in §2.4.5, this assumption is problematic.

2.4.3.3 More general item response theory models

The Rasch model is a particular case of a family of statistical models known as item response theory (IRT) models (van der Linden and Hambleton, 1997). For the case of assessments with dichotomous items, these extend the formulation in equation 2.1 to include more parameters intended to capture aspects of the hypothesised relationship between the proficiency of a learner and aspects of the test item the learner engages with. In general the (unidimensional) dichotomous IRT model can be written as

$$P(x_{ijkl} = 1) = \gamma_j + (1 - \gamma_j) \left\{ \frac{\exp(\alpha_j(\theta_i - \delta_j - \epsilon_k - \zeta_l))}{1 + \exp(\alpha_j(\theta_i - \delta_j - \epsilon_k - \zeta_l))} \right\},$$

where $P(x_{ijkl} = 1)$ is the probability that student i gets item j correct, given that it was in section k of the test and marked by examiner l, and

γ_j is the propensity of item j to guessing;

α_j is the 'discrimination' of item j;

θ_i is the proficiency of student i;

δ_j is the difficulty of item j;

ϵ_k is the challenge that section k of the test presents (it could consist of one or a number of items);

ζ_l is the leniency/severity of examiner l in marking (categorising responses as correct or incorrect).

The most commonly used versions of this model, apart from the Rasch model, are the *two-parameter model*, in which (in addition to the student proficiency parameter) the item difficulty and discrimination parameters δ and α are estimated, the others being set to zero; and the *three-parameter model*, in which additionally the guessing parameter γ is estimated.

When multi-parameter IRT models are used, one no longer has the link between total score and proficiency estimate. Two learners with the same total mark but different mark profiles may have different proficiency values. Calculating estimated proficiencies using an IRT model of this kind does produce a total order on the set of all observed responses to an assessment, because each performance has a quantitative proficiency score associated with it. But this ordering is not the same as the ordering of performances by their quality, if, as in the compensatory approach discussed above, ordered equivalence classes of 'levels' or 'kinds' of performance quality are determined by summing marks.

This observation throws into relief the need to have a clear conception of an appropriate mathematical structure for the construct being assessed. IRT models assume that the construct is a latent property of persons, and that it has the structure of a quantity. Compensatory models assume that the construct is defined by means of a collection of (more or less fuzzily defined) attributes, each of which has an ordered value-set (in the simplest, dichotomous, case, just {0,1}). Students' assessment results give a profile of performance against relevant attributes, and (as discussed further in §2.6) these results are consequently partially ordered. 'Compensation' means imposing a total order on the set of performances, and the most common way of doing this in practice is by deeming any two performances to be of equivalent quality if they have the same total score.

It should be noted, however, that it is possible to create ordered categories of performances without ranking all performances with respect to each other – a point which is germane to the case of public examinations, where the promise that the awarding organisation attests to, when awarding a candidate a qualification, is not what mark they have obtained, but what *grade* of performance they have demonstrated. For instance, let $\preceq$ represent a partial order on a set of examination responses, and suppose we are given four students' responses, $s_1, \ldots, s_4$. Suppose s_1 is a 'bare grade A' performance (a 'borderline A/B performance'), and s_4 is a borderline B/C performance. Suppose $s_2 \preceq s_1$ and $s_3 \preceq s_1$; and also $s_4 \preceq s_2$ and $s_4 \preceq s_3$. Then it follows that both s_2 and s_3 are both grade B performances, even if we cannot say which of the two is the 'better' grade B, or whether they are 'equal' grade Bs. They are simply different examples of the kind of performance that merits a grade B. This point is returned to in §2.6.

2.4.4 Assessment responses with ordered (polytomous) attributes

2.4.4.1 Polytomous IRT models

Most attributes in UK public examinations are not dichotomous. Candidates' responses to tasks are generally ascribed a value (a level) with respect to the attributes of interest by means of assessors (examiners) applying classification schemes (mark schemes). As noted in §2.3.1, the method of classification may be rules-based or pattern-matching-based, with the aim being that examiners who have been trained to apply a mark scheme can classify responses consistently and in such a way that (in the terms of §2.3.2.2) the higher the valuation, the higher the degree to which it is true that the selected attribute has been demonstrated in the response.

Polytomous IRT models (Andrich (1978b); Ostini and Nering, 2005) have been developed to handle assessments in which items can have more than two ordered categories or levels

of valuation (more than two degrees-of-demonstration of an attribute). Such models assume a quantitative proficiency continuum that can be divided into intervals defined by a sequence of threshold proficiency values. For example if an assessment item classifies learners into one of three categories – labelled, say, as `fail, pass, merit` – there is assumed to be a theshold value θ_P such that if a candidate has more than θ_P units of proficiency, she is more likely to pass than fail, and a value θ_M such that if her proficiency exceeds θ_M her most likely result on the item is a merit. In this way polytomous responses can be treated as ordered collections of dichotomies, and dichotomous IRT models can be extended. For instance the polytomous version of the Rasch model (eq. 2.1) is

$$P(x_{ij} = k) = \frac{\exp\left(k(\theta_i - \delta_j) - \tau_{1j} - \tau_{2j} - \ldots - \tau_{kj}\right)}{\sum_{n=1}^{m_j} \exp\left(n(\theta_i - \delta_j) - \tau_{1j} - \tau_{2j} - \ldots - \tau_{nj}\right)},$$

where:

there are $m_j + 1$ possible levels, labelled 0,1,...,m_j, for item j;

x_{ij} is the response of student i to item j;

θ_i is the proficiency of student i and δ_j is the difficulty of item j; and

$\tau_{1j}, \tau_{2j}, \ldots, \tau_{m_j j}$ are the threshold parameters for the different levels of item j.

2.4.4.2 Polytomous items in practice

Polytomous models hypothesise that the thresholds are ordered, such that

$$\tau_{m_j j} > \tau_{m_j - 1, j} > \ldots > \tau_{2j} > \tau_{1j}$$

for each item j. However this ordering might not be consistent with the outcome data from the assessment. It is consistent with an ordering process (that is to say, a marking-scheme instruction) that requires markers to assign higher marks to responses that demonstrate the attribute in question more fully or strongly. This is often the case for items marked in the pattern-based manner exemplified in Figure 2.1. This approach to marking, often applied when item responses take the form of short or long-form prose, is usually called *level-of-response* marking in the context of public examinations.

However, the ordered-threshold hypothesis may not be consistent with an ordering process that arises from instructions to markers to assign partial credit for different features *within* a single attribute. For instance, as noted in §2.4.3.2, polytomous items in mathematics assessments (marked in a rules-based manner) may allow candidates to obtain marks for method and for accuracy. Such items may display 'disordered thresholds' if a polytomous IRT model is fitted, and might be better treated statistically as sequences of (non-independent) dichotomies. That is, they may be better conceived of as separate (but related) attributes, than as 'more of' a single attribute.

Even when it does seem reasonable to think of the levels of an attribute as indicating the degree to which it has been demonstrated in a given response, though, this rationale can be confounded in practice with the use of mark totals to *weight* certain assessment items more highly than others. For instance, in GCSE English language examinations, there are a number of items that require candidates to write essay-style answers, testing different assessment objectives, which thus form the attributes about which information is provided by the assessment procedure. Two example attributes are (A_1) 'communicates clearly and effectively'; and (A_2) 'uses a range of vocabulary and sentence structure with accurate spelling and punctuation'. Each of these attributes can be demonstrated to a greater or lesser degree in a given response, and each item is marked using level-of-response marking with five levels. But because there is a requirement that A_1 should

carry more weight than A_2 when determining the overall quality of a candidate's performance on the examination, the levels-of-response mark scheme for the examination question assessing A_1 is extended by comparison with the mark-scheme for A_2, by superimposing additional discrimination (mark points) within each level in order to increase the proportion of a candidate's total score that is accounted for by this item. This is because weightings are expressed not as orderings but as percentages: e.g. that A_1 is weighted at 30% and A_2 is weighted at 20%[12].

Weighting in this way requires more than ordinal structure to the attributes – for it requires, for example, 5 marks to count for five times as much as 1 mark. It is not obvious that simply producing marking criteria to specify what counts as 1, 2, ..., 5 marks ensures this 'equal interval' requirement. To assume so would be to commit what Michell (2012) calls the pychometricians' fallacy.

2.4.5 The psychometricans' fallacy

The *psychometricians' fallacy* (Michell, 2009; 2012; 2020) is to assume that ordinal structure necessarily entails quantitative structure. Applied to the constructs studied in psychometrics, this assumption permits them to be represented on an equal-interval scale (Stevens, 1946). The extra requirements that require justification in order to infer quantity from order are (*i*) that there is a meaningful notion of 'difference' between the levels (e.g. 'good response', 'satisfactory response', 'poor response') in the value-set for the construct, or attribute, under consideration ; and (*ii*) that these differences themselves have a compositional structure, such that they stand in ratios to each other. Michell takes issues with the latter assumption on several grounds, including that it is unlikely that the difference between one pair of levels for a psychological construct will

[12]The author has observed examples of GCSE examiners in practice – for instance in grade awarding meetings – actually describing weightings as orders rather than numerical percentages, for example stating that 'on this paper the assessment objectives are weighted $A_1 > A_2 > A_3$'.

be entirely homogeneous with the difference between another pair of levels (in his terms, psychological constructs exhibit 'impure differences of degree': Michell, 2012, p. 264). As an example, he notes (2012, p. 265) that

> [T]he differences between cognitive resources needed to solve easy and moderately difficult mathematics items will not be the same as the difference between the resources needed to solve moderately difficult and very difficult mathematics items. This observation suggests that abilities are attributes composed of ordered hierarchies of cognitive resources, the differences between which are heterogeneous.

It is interesting to compare this example with Keynes' (1921) treatment of probability, of which he states 'A degree of probability is not composed of some homogeneous material, and is not apparently divisible into parts of like character with one another' (p. 32)[13]. This is germane to the discussion in Chapter 4 of the nature of item response models that take the (assumed quantitatively structured) *probability* of a learner correctly answering an assessment item as the joint effect of the learner's *proficiency* and the item's *difficulty*, also hypothesised to be quantitative latent variables.

The psychometricians' fallacy highlights that it is conceptually problematic to assume that educational assessment constructs (construed in either of the senses discussed in

[13]The mathematical theory of probability takes as its point of departure axioms developed by Kolmogorov (1933/1954), that posit a particular kind of real-valued mapping (called a probability measure), on a a set-theoretic structure that is motivated by abstracting from the notion of the occurrence or non-occurrence of an event. As a formal mathematical theory, it simply *defines* probability as a quantitative property, and explores what theorems can be deduced on the basis of the axioms. Keynes was interested in the idea or concept of probability in an experiential sense, and how the intuitions about the nature of probability as a psychological entity might be described. He suggested that a partial-order representation, resulting in non-additive, non-linear, interval-valued measures, would provide a better theory of this concept of probability. There is a possible analogy with theories of colour perception as an experiental event, which can suggest that people actually perceive changes in colour along a number of dimensions, by comparison with the physical theory of colours as manifestations of different frequencies of light, i.e. as a quantitative phenomenon (whose values are points on the electromagnetic spectrum).

§2.4.2) have the mathematical structure of a quantity. But can this question be investigated empirically? The following section outlines an approach to doing so in certain cases, using results from a particular theory of measurement.

2.5 Measurement and quantities

2.5.1 The representational theory of measurement

Tal (2020), in his survey of the philosophy of measurement in science, describes how investigations into the possiblity of quantifying phenomena studied in the human sciences, and into the notion of 'scales of measurement', converged in the mid twentieth century in work by Patrick Suppes (1951) that aimed to give an axiomatic description of the notion of measurement. Suppes' work laid the basis for the representational theory of measurement (RTM), which, according to Tal, 'remains the most influential mathematical theory of measurement to date'. Wolff (2020), in a recent structuralist account of quantity and measurement, describes RTM as 'arguably the most developed formal theory of measurement'. Michell (1990a) claimed that RTM 'is the orthodox theory of measurement within the philosophy of science'.

According to the canonical text on RTM (Krantz et al, 1971, p. 9), the representational theory of measurement regards measurement as: *'the construction of homomorphisms (scales) from empirical relational structures of interest into numerical representational structures that are useful'.*

An 'empirical relational structure', in this definition, is thought of as a collection of objects that are observed to stand in certain relations to each other, and possibly also admit of ways of being combined with each other. The coherence of this notion of an

empirical relational system, and its relationship to the notion of *abstraction*, is examined in Chapter 5, with specific reference to the case of educational assessment. The current section restricts itself to a brief sketch of the application of RTM to assessment, in particular to the question of investigating the warrantability of assumptions about quantitative structure for observed attributes of learners' performances, or hypothesised latent traits.

2.5.2 Applying the representational theory of measurement in educational assessment

In assessment, we might take as objects, for example, students' responses to a writing task, and consider a binary relation $\succeq$ of *betterness* as being of interest (as in 'student X's piece of writing is a better response to the task than student Y's piece of writing': $X \succeq Y$) . Or we might be interested in how parts of assessments combine (via a binary operation $\bullet$) to form an overall measure. For example, 'correctly answering questions 3 and 4 represents a higher level of attainment than correctly answering questions 1 and 2': $q_3 \bullet q_4 \succeq q_1 \bullet q_2$. We might then wish to investigate whether these aspects of students' responses to tasks – this empirical relational system – can form the basis of a numerical ordering or scoring system.

Results from RTM can help in carrying out such an investigation. If it is possible to represent, for instance, the empirically observed relation of qualitative *betterness* by means of a numerical scale, then we can obtain valid inferences from an assessment procedure that collects data on *betterness* by drawing inferences within the numerical relational structure that represents it. For instance, comparing the numerical values assigned in this way to three different students' responses to an assessment task would then be sufficient to conclude which response was the best, which second-best, and which the worst. If the numerical relational structure has the form of an interval or ratio scale,

then it is also meaningful to draw conclusions about *how much* better one student's response is than another, and whether a certain student's response has demonstrated a certain 'amount' of quality, represented by a particular point (such as a cut-score, or threshold value) on the scale.

If, on the other hand, it is not possible validly to represent the empirical, qualitative, relational structure of students' responses to assessment tasks and their judged relative quality by means of a certain kind of numerical relational structure, then one can of course draw valid inferences *within the numerical structure*: but there is no warrant for concluding that these particular inferences or derivations tell us anything substantive about the *responses* or the *students*. As noted by Briggs (2013) and Ballou (2009), educational assessment procedures often simply *assume* that the phenomena that are the objects of measurement (e.g. quality of writing) are interval-scaled. Briggs and Ballou demonstrate how this assumption, if not warranted, is particularly problematic when assessments are used to draw inferences about students' progress, or educational growth over time, or for the purpose of appraising the value-added by educational interventions. If the object of measurement is not, in fact, quantitatively structured, then inferences drawn from scale scores may simply represent artefacts of the assumed numerical relational structure, rather than valid substantive conclusions about educational progress or growth.

2.5.3 Conjoint measurement: testing for quantitative structure

Luce and Tukey (1964) introduced *conjoint measurement* as potential method of enabling quantification in situations in which a property, whose values are ordered, arises as the conjunction, or joint effect, of two or more other ordered properties. In assessment, this is what is envisaged when the Rasch model is applied: the probability of a correct response to a dichotomous assessment item is modelled as the joint effect of the proficiency of

the candidate and the difficulty of the item. The key theorem of conjoint measurement (Krantz et al., 1971, p. 257) states five conditions[14] that are sufficient for the empirically observed matrix of ordered values of the properties to be represented by a numerical relational system in way that preserves the observed relationships between the properties. That is, the conditions define sufficient conditions for all three of these properties to have quantitative structure.

Thus, by investigating whether the conditions hold in actual assessment or testing contexts, it is possible in principle to obtain evidence for or against the quantity assumption. Michell (1990a) takes this approach. For example, he re-analysed in this way data collected by Thurstone (1927) in an application of the latter's method of *comparative judgement* to measure the seriousness of crimes. The method of comparative judgement assumes that the property that it is used to measure does have quantitative structure. Michell's conclusion from his re-analyis of Thurstone's data was that 'either seriousness of crimes is not a quantitative variable, or else some other part of Thurstone's theory of comparative judgement is false'.

A practical problem with applying this approach in assessments of any realistic size is that the number of checks that need to be performed rapidly becomes much larger than is feasible to complete in a reasonable time, even with fast processing speeds. For this reason, and also to attempt to allow for 'measurement error' in the observed data, stochastic approaches to testing the conditions have been developed, for the case of assessments consisting of dichotomous items (Karabatsos, 2001; 2018; Domingue, 2014). Chapter 4 investigates the application of such methods to high-stakes assessment data in the UK context, and concludes that the emprical evidence suggests the quantity assumption is problematic. One possible response to this conclusion is to consider whether its is still possible to entertain a coherent and useful conception of *measurement*, if the requirement

[14] See Chapter 4 for details.

for quantitative structure is relaxed or dropped.

2.5.4 Measurement without quantification

2.5.4.1 Measurements as information

In an early critique of the representational theory of measurement, Adams (1966, p.130) makes a case for an informational theory of measurement, in which 'measures of a quantity are not so much "true" or "false" as they are more or less informative about the phenomena they are supposed to be indices of.' His objection to RTM is based in part on a rejection of the idea that the structure that does the 'representing' – i.e. the numerical relational structure – must be *numerical,* noting that 'the ancient Greeks did not have our concepts of rational, much less real numbers, yet it seems absurd to say that they could not measure because they did not assign numbers to objects. In sum, I would say that the employment of numbers in describing the results of measurement is not essentially different from their employment in other numerical descriptions, and that this employment is neither a necessary nor a sufficient condition for making or describing measurement'.

Adams claims that the proper question to ask of a procedure for measuring some quantity (say weight) is not "is it a true measure of the quantity or not?", but "how good an indicator is it of the phenomena it is supposed to give information about?". This seems close to the position with respect to examination scores that is taken by Ofqual, the government regulator of standards in public examinations in England. It explicitly recognises that the extent to which a performance merits a particular mark is ultimately a value-judgement. Its official criteria for reviewing how performances have been marked[15]

[15]See https://www.gov.uk/guidance/regulating-gcses-as-and-a-levels-guide-for-schools-and-colleges-2022/reviews-and-appeals.

state that difference of academic professional opinion between competent judges as to the merit of a performance are not 'marking errors', provided that these opinions have been expressed with reference to the rationale for encoding features of examinee responses (the mark scheme). Reviewers are not asked to determine whether the performance has been given 'the correct mark', but rather to determine whether the original mark could have been given by a marker who properly applied the mark scheme to the response and exercised his or her academic judgement in a reasonable way: whether it is a 'good indicator' of the student's attainment.

Adams goes on to link the idea of information-content to the idea of location with respect to a frame of reference (p. 146):

> If we focus only on the intrinsic informational content of measurement procedures, many procedures can be regarded essentially as ones which locate objects (the objects measured) in frames of reference (the known standards). To locate an object A in a spatial reference frame S is to determine the spatial relation of A to S (e.g., to determine how far A is and in what direction from a given 'origin'). Similarly, to determine the Mohs scale hardness of a rock is to determine its scratching relations to known mineral standards, and to weigh an object on a beam balance is to determine what combination of standard weights it balances. ... The analogy of measurement to determining location in a reference frame somewhat reinforces, I think, the thesis that there is nothing essentially numerical about measurement. ... [R]eference frames are chosen for convenience and informativeness and so are standards and numerical descriptions.

2.5.4.2 Measurements as locations

Adams' characterisation of measurement as location pre-figures in some respects the theory of measurement advanced by van Fraassen (2008), fundamental to which is the notion of *location in a logical space*. van Fraassen's viewpoint includes, as a special case, the traditional conception of measurement as quantification (location of the value of a measurand on a line), but extends it to encompass, for example, measuring procedures that are 'cases of grading, in a generalised sense: they serve to classify items as in a certain respect greater, less, or equal. But ... this does not establish that the scale must be the real number continuum, nor even that the order is linear. The range may be an algebra, a lattice, or even more rudimentary, a poset' (van Fraassen, 2008, p. 172)[16].

The theoretical framework developed in Chapter 5 conceptualises proficiency in an educational domain as a qualitative phenomenon, and suggests that it can be measured in van Fraassen's sense. In order to *locate* a student's attainment within an educational context or domain (for example, to *certify a level of competence* with respect to that domain), it suggests applying elements of mathematical order theory and fuzzy logic to the data generated by educational assessment procedures. This conceptualises the relevant logical space for educational assessment procedures such as public examinations as a certain type of network, or partially-ordered set (a so-called fuzzy concept lattice), based on the suggestion (as set out earlier in §2.3.2.2) that assessment contexts can be modelled as fuzzy relational systems. An application of this approach to the question of grading public examinations in England is then explored in the following chapter.

To conclude this introductory survey, therefore, the following section briefly introduces some of the analytical approaches relevant to the analysis of relational systems. Further

[16] An *algebra* is a type of mathematical structure. See Chapter 3 for definitions of *lattice* and *poset*.

background relevant to the methods used in Chapters 5 and 6 is given in the Methodology chapter that follows this one.

2.6 Non-quantitative constructs: structural approaches to analysing assessment information

2.6.1 Order-theoretic approaches

The analytical techniques of psychometrics and test theory discussed in §§2.4.1-2.4.4 were developed using mathematical tools available in the late nineteenth and early twentieth centuries. They are largely grounded in analytical models for continua (in particular the real numbers and vector spaces over the reals). Over the twentieth century, the mathematics of discrete (non-continuous) structures became increasingly important in applications, for example computing, artificial intelligence, and linguistics (see e.g. Schmidt and Ströhlein, 1993; Partree, ter Meulen and Wall, 1993). Social science applications of a wider range of mathematical approaches are now becoming more widespread (e.g. Rudolph, 2013), including in particular methods of studying subject matter where the notion of *ordering, comparing* or *ranking* is of interest - topics studied in general terms in mathematical *order theory* (Davey and Priestley, 2002). For example, see Fattore (2016) for applications to the assessment of multidimensional deprivation in international development, and Badinger and Reuter (2015) for applications to fiscal policy analysis. These applications do not assume that the phenomena or features being studied must be representable by real numbers.

With respect to educational assessment, Kane (2008, p.104) notes that 'given an area of achievement that is broadly defined, we are likely to have, at best, a partial ordering [of learners' responses to assessment tasks]' – in other words, that the observed collection

of learners' responses usually forms a partially-ordered set. In the terminology of §2.4.2, each construct-relevant attribute has an ordered set of possible values, and each learner's response can therefore be represented as a vector of ordered values (learner i's response is the ith row of the outcome matrix discussed in §2.3.2.3). Suppose there are three attributes, each taking values labelled 1, 2, or 3 (with the usual ordering of integers). Then a learner with response (2,2,3) has performed better than one with response (1,1,1). But learners with responses (1,3,2) and (3,1,1) cannot be compared with each other (unless some other rules are imposed, such as a specific method of compensation: §2.4.3.2): these learners have produced qualitatively different responses that locate them, within the partially-ordered set of all possible response profiles, at different places, of which it cannot be said *a priori* that one is 'higher' or 'better' than the other.

2.6.2 Knowledge space theory

The theory of *knowledge spaces*, originated by Doignon and Falmagne (1985), aims to study partially-ordered assessment response data directly. It aims to link structural features of the data, investigated using methods and results from order theory, to psychological theories of cognitive processing. Assessment constructs (in the sense of 'what is assessed': §2.4.2) are represented as partially-ordered collections of 'knowledge states', each of which is defined in terms of the 'atomic' construct-relevant attributes that collectively embody the construct. The aim of analysing data from an assessment procedure is then to determine each learner's knowledge state at the time of their being assessed – that is, to locate each learner, to an acceptable degree of probability, within the partially-ordered 'knowledge space' relevant to the assessment construct. Different learning pathways through the knowledge space, from 'no knowledge' to 'mastery' of the given construct, may be appropriate for different learners, depending on their current knowledge states, and in this way, knowledge space theory may be used as the basis of

an assessment and learning system.

The main substantive application of the theory to date is the ALEKS system[17] run by McGraw Hill, which offers knowledge-space-theory-based assessements and learning materials in a range of primary, secondary, and tertiary level subjects such as mathematics, sciences, and accounting. It would appear to be more suited to assessment constructs that can be defined quite 'atomistically', often via collections of dichotomous attributes, rather than attributes that are fuzzier, whose assessments are more likely to require pattern-based categorisation (§2.1).

2.6.3 Fuzzy logic and formal concept analysis

The idea of *fuzzy logic* (Hajek, 1998; Bělohlávek et al., 2017) was introduced in §2.3.2.2. In fuzzy logics, propositions are not restricted to being either true or false, but may have more degrees of truth. Often truth-degrees are represented by numbers between 0 (false) and 1 (true) – although these should be regarded as ordered labels, not numerical quantities. It should also be noted that degrees of truth (in logic) differ from degrees of belief (in probability). Truth degrees are used to model notions that are inherently vague or inexact. For instance, one might ascribe a truth degree of 0.75 to the statement 'this student's performance demonstrates persuasive writing'. On the other hand, degrees of belief are used to model uncertainty resulting from a lack of evidence in Bayesian probability theory (Gelman et al., 2013), and associated fields such as possibility theory (Dubois and Prade, 1988) and Dempster-Shafer evidence theory (Shafer, 1976). For example, one might ascribe a probability (degree of belief) of 0.75 to the statement 'this student will pass the exam tomorrow'.

[17] www.aleks.com

Bartl and Bělohlávek (2011) have extended the methods of knowledge space theory to admit fuzzy attributes. Their work forms part of a wider project to link knowledge space theory to another subfield of order theory known as *formal concept analysis* (Ganter and Wille, 1999). Formal concept analysis studies relational systems in which there is a collection of objects, each of which may have one or more attributes. Thus it is directly applicable to educational measurement procedures viewed as collections of relationships between learners and construct-relevant attributes, as suggested in §2.3.2.3. *Fuzzy* formal concept analysis allows such relationships to hold with variable degrees of truth. Further background and details are summarised in Chapter 3.

Formal concept analysis allows the information in data matrices such as those in §2.3.2.3 to be represented as a partially-ordered structure, known as a *concept lattice*. Rather than thinking of each node in this structure as a knowledge state, as in knowledge space theory, they can be thought of as representing the different *concepts* that can be formed given the objects and attributes for which the data has been collected. It is argued in Chapter 5 that such concept lattices enable the measurement of the value of a student's attainment, with reference to a notion of standards, without assuming an underlying quantitative structure.

To date, few applications of formal concept analysis have been considered in the educational assessment literature. Bedek and Albert (2015) report on its application to visualising classroom performance, as part of a three-year study to develop a 'learning analytics toolbox', known as LEA's Box[18] that was trialled in schools in Austria, the Czech Republic, and Turkey. Bartl, Bělohlávek, and Scharaschkin (2018) examined the use of fuzzy formal concept analysis as an alternative to traditional quantitative factor analysis for the identification of core features of students' performances in a UK A-level context. Perez Gamez *et al.* (2020) investigated using formal concept analysis to explore

[18] http://leas-box.cognitive-science.at/index.html

hidden knowledge in the assessment of a university mathematics course. In general, however, the possibilities that are opened up by treating educational assessment procedures as fuzzy relational systems have yet to be fully explored. This dissertation aims to make a contribution to the area.

2.7 Conclusion

The literature on educational assessment, especially when presented as educational measurement, tends for the most part to conceive of the phenomena – the assessment constructs – that are investigated in assessment procedures as quantitative latent traits. There is an alternative tradition that focuses more on the structure of constructs as assemblages of attributes, which pertains especially to the context of public examinations in the UK context[19]. The two viewpoints are linked by means of the identification

attainment valuation with *compensation* (via sum scores)

$$\updownarrow$$

proficiency estimation (via the Rasch model).

However, the validity of conclusions drawn using either of these conceptions, and the warrantability of their identification using total scores, depends on the reasonableness of the assumption that the construct can be treated as if it has the structure of a quantity (or as if each attribute can be treated as a quantity, and composed with other attributes accordingly, for instance via addition). The quantity assumption can be tested, but has not previously been explored with actual assessment data in the UK context. Nor

[19] There is a flavour of this in the literature on standard setting and maintenance in the UK high-stakes examination context, for example in the description of 'weak criterion referencing' of examination grades in Baird, Cresswell, and Newton (2000), and Newton's (2020) notion of 'attainment referencing', in which grade boundaries (cut-scores) are referenced to the nature of candidates' attainment, rather than threshold values of an underlying latent variable. The application of fuzzy formal concept analysis in Chapter 6 builds on these ideas.

have the consequences for its not being met been explored with respect to theorising educational assessment as a form of measurement. The papers that follow aim to advance knowledge in these areas.

3 Methodology

This chapter provides a brief summary of the methodological and technical background assumed in the papers that follow. Each paper addresses a separate research question, and the following sections cover background material relevant to each question, that is assumed in the methodology discussion in the corresponding paper.

Because this is an integrated thesis (a collection of separately self-contained papers), each paper contains its own topic-specific methodology discussion. Hence the material below provides some additional technical background, and reiterates some of the key themes from the separate papers.

3.1 RQ1: Empirically, does the assumption of quantitative structure hold for educational assessments used in high-stakes qualifications in England?

The investigation of this research question is set out in the paper included here as Chapter 4. It relies methodologically on the use of results from the representational theory of measurement (Krantz et al., 1971), which itself is predicated on the set-theoretic concepts of *relation*, *mapping*, and *operation*. It focuses particularly on *order* relations, and on *structure-preserving* mappings. These terms are defined here.

3.1.1 Sets and relations

One may think of a *set* as a collection of objects (also called members or elements). The notation $x \in X$ means 'x is a member of the set X'. Let X and Y be sets. Their *product* $X \times Y$ is the set of all ordered pairs (x, y), where $x \in X$ and $y \in Y$. A (binary) *relation* R from X to Y is a subset of $X \times Y$. Thus, R picks out particular pairs of elements (x, y). It captures the notion of one thing (x) being *related to* another thing (y). For example, if X is a collection of people, and Y is a collection of countries, we could define a relation R from X to Y by saying that person x is related to country y iff x has visited y. Then R is the relation 'has visited', and it is fully specified in this context by listing all pairs (x, y) where x has visited y. Hence the relation may be identified with the set of all such pairs. If x is related, by R, to y – i.e., if $(x, y) \in R$ – then one often writes xRy.

Given a set X, a binary relation on X is simply a relation from X to X, i.e. a subset of $X \times X$. For instance one might think of X as the collection of students' responses to a writing task, and define a relation $\succeq$ on X by defining $x_1 \succeq x_2$ iff response x_1 is judged to be *better than* response x_2 by a teacher who has been asked to assess this collection of responses with respect to some criterion of quality of writing.

A *mapping* (also called a *map* or a *function*) from X to Y is a relation from X to Y which has the property that it relates every $x \in X$ to a unique $y \in Y$. If m is a mapping from X to Y, we write $m : X \to Y$, and if $(x, y) \in m$, we write $y = m(x)$. One can think of a mapping $m : X \to Y$ as taking inputs from X, and producing, for each input x, a unique output $m(x)$.

A binary *operation* on a set X is a mapping from $X \times X$ to X. That is, a binary operation on X takes pairs of elements of X as inputs, and produces elements of X as outputs.

Operations may be thought of as capturing the notion of *combining* elements of a set[1]. For example, the operation of addition on the integers combines two integers to produce another integer (their sum). Or consider an assessment that consists of two parts, such that a candidate must pass the first part (a so-called hurdle), in order to get a grade (fail or pass) based on their performance on the second part. Let x be a candidate's grade on the first part, and y their grade on the second part. Then grades are combined via the binary operation $\bullet$ defined by

$$x \bullet y := \begin{cases} \texttt{fail} & \text{if} x = \texttt{fail} \\ y & \text{otherwise} \end{cases}$$

3.1.2 Order relations and structure-preserving mappings

Let $\sim$ be a binary relation on a set X. Then $\sim$ is said to be:

- *reflexive* if $x \sim x$, for all $x \in X$;
- *transitive* if for all $x, y, z \in X$, whenever $x \sim y$ and $y \sim z$, then $x \sim z$;
- *anti-symmetric* if for all $x, y \in X$, if $x \sim y$ and $y \sim x$, then $x = y$; and
- *strongly connected* if for all $x, y \in X$, either $x \sim y$ or $y \sim x$.

A binary relation that is reflexive, transitive, and anti-symmetric is called a *partial order.* A *partially ordered set,* or *poset,* is a set with a partial order. A partial order that is, in addition, strongly connected is called a *total order.* These definitions aim to capture

[1] Note that because operations are mappings, and mappings are relations, it follows that operations in general can be thought of as kinds of relations. Define an n-ary relation between the sets $X_1, X_2, ..., X_n$ to be a subset of $X_1 \times X_2 \times X_n$. An n-ary operation is then an $(n+1)$-ary relation. For example a binary operation on X is a ternary relation on X, because the binary relation is a map $X \times X \to X$, i.e., a subset of $(X \times X) \times X$.

the core structural features of the notions of *comparing* or *ordering* objects in some way. For instance the relations of *being better than*, or *being at least as good as*, in some sense or with respect to certain criteria – which, as noted in §2.3.2, are fundamental to the enterprise of educational assessment – are examples of order relations. As discussed in §2.3.2 and §2.6, in many, if not most, educational assessment contexts, it is clear that the information generated by assessments gives rise to partially-ordered sets of responses. In some cases, for example if responses to a task are being appraised with respect to a single attribute, rather than a more complex construct that involves multiple attributes, the responses may naturally be totally ordered – that is to say, ranked – with respect to that attribute.

There are obvious order relations on well-known abstract mathematical structures. For instance the positive integers are totally ordered by the usual 'greater than or equal to' relation $\geq$. If one considers ordered pairs of positive integers, then the so-called product order $\gtrsim$, defined by

$$(a, b) \gtrsim (c, d) \text{ iff } a \geq c \text{ and } b \geq d$$

is a partial order.

A key concern of the representational theory of measurement (RTM) is the question whether such abstract relational structures can be used as models for, or representations of, empirically observed relations between objects. RTM conceives of the idea of *representing* one structure by another in terms of *mapping* the first structure into the second, in such a way that structural features of interest are preserved. To be precise, if $(A, \sim, \bullet)$ is a triple consisting of a set A on which is defined a binary relation $\sim$ and binary operation $\bullet$, and if $(B, \approx, \circ)$ is another such triple, then a map $h : A \to B$ is said to *preserve structure*, or to be a *homomorphism* from A to B iff, for all $a_1, a_2 \in A$, it is the case

that $a_1 \sim a_2 \Rightarrow h(a_1) \approx h(a_2)$ and that $h(a_1 \bullet a_2) = h(a_1) \circ h(a_2)$. This definition can be extended in the obvious way to cases involving larger numbers of relations and/or operations.

3.1.3 Homomorphisms of relational structures: representation theorems

RTM is concerned in particular with representing structures of interest (the so-called *empirical relational structures:* see §§2.5.1-2.5.2 and Chapter 5) by means of *numerical relational structures.* The latter are (Krantz et al., 1971, p.8) tuples of the form $(\mathbb{R}, S_1, \ldots, S_n)$,where $\mathbb{R}$ is the set of real numbers, and $S_1, \ldots, S_n$ are relations on $\mathbb{R}$. RTM aims to establish *representation theorems*, which assert that if a given (empirical) relational structure of interest satisfies certain axioms, then a homomorphism into a certain numerical relational structure can be constructed. As Krantz et al. note (1971, p.9):

> A homomorphism into the real numbers is often referred to as a scale in the psychological measurement literature. From this standpoint, measurement may be regarded as the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful. Foundational analysis consists, in part, in clarifying (in the sense of axiomatizing) assumptions of such constructions.

The basic example of an empirical relational structure that is always given in introductions to RTM is a collection of rods, that can be compared with each other in terms of their lengths. So, for example, we can observe (by placing rods next to each other) that rod 1 is *longer than* rod 2 (written $r_1 \succeq r_2$); and that rods can be combined together (by laying end-to-end) such that we may observe that rod 3 is the same length as rod 1 combined together with rod 2, written $r_3 = r_1 \bullet r_2$. RTM then defines a *measure* of (or a

scale for) the property of *length* of rods as a homomorphism L from the set of rods into the real numbers, i.e. a mapping that associates a number $L(r)$ with each rod r, such that:

$L(r_1) \geq L(r_2)$ iff $r_1 \succeq r_2$; and

$L(r_3) = L(r_1) + L(r_2)$ iff $r_3 = r_1 \bullet r_2$.

In this way, comparing or combining empirically-observed properties is replaced with comparing or combining (adding) numbers. The number $L(r)$ measures the length of rod r with respect to the scale L.

There may be multiple homomorphisms from an empirical relational structure to a numerical one – for instance, lengths could be measured in feet or metres. One area of interest in RTM is establishing how such homomorphisms are related to each other. For example, if L_1 measures length in metres, and L_2 measures length in feet, then $L_1 = 0.305L_2$; and in general any two homomorphisms L_1 and L_2 that provide measurement scales for length are related via a transformation of the form $L_1 = aL_2$, for some constant a. This is because the property of length has quantitative structure, and measuring an object's length locates the object at a particular position on a *ratio scale* (see §4.2.1). A property of ratio scales is that any two homomorphisms that provide measures of the ratio-scaled phenomenon stand in the relation that one is some constant multiple of the other.

3.1.4 Conjoint measurement and empirical tests of quantitative structure

Michell (1990) sets out how representation theorems can in some cases form the basis of a methodology for testing whether a particular phenomenon has quantitative structure. Because such theorems establish necessary and sufficient conditions (the relevant axioms) for real-valued homomorphisms to exist for phenomena that are assumed to give rise to

particular relational structures, it may be possible to examine whether those conditions hold in practice, and hence to conclude as to the warrantability of the assumption of quantitative structure for phenomena of interest. This does not necessarily entail accepting RTM as an adequate theory of measurement (indeed Michell himself does not accept it as adequate: see Michell, 2020), but rather simply using the logical entailments established by the theorems in certain contexts to check whether or not necessary conditions for quantitative structure obtain[2].

Particular contexts that are relevant to educational assessment are those that give rise to so-called *conjoint structures* (Luce and Tukey, 1964). These arise when an attribute or property of interest is conceptualised as the *conjunction*, or joint effect, of two or more other other properties. Each property in this scenario is assumed to have possible levels or values that can be ranked (i.e. that form a totally ordered set). As discussed in §2.4.3.2, this conceptualisation is relevant to the approach to educational measurement instantiated in the Rasch model, which postulates that performances on a test or assessment are measurements derived from the proficiencies of persons and the difficulties of items.

The key representation theorem of conjoint measurement (Krantz et al., 1971, p. 257) describes conditions that are required for the empirically observed matrix of ordered values of the properties to be represented by a numerical relational system in way that preserves the observed relationships between the properties. That is, the theorem defines necessary and sufficient conditions for all three of these properties to have quantitative structure. The operation of conjunction in the conjoint structure may be mapped onto the numerical operation of addition, in which case the process is known as additive conjoint measurement, or onto the operation of multiplication, in which case one has multiplicative conjoint measurement.

[2]Section 2.2.2 of Chapter 5 develops this point further.

The methodology used in Chapter 4, and set out in more detail there, focuses on two of the conditions that the representation theorem establishes as necessary to allow additive conjoint measurement, namely the so-called single and double cancellation axioms, and tests whether they hold for the results of some public examinations in England.

As explained further in Chapter 4, the number of tests that must be performed to check these axioms grows geometrically with the number of items in the assessment. Thus in practice a sampling approach is used that examines a sample of the complete collection of cases that must be tested, in order to conclude as to the likelihood of the axioms to be satisfied in general. And in fact, such sampling approaches do not make the assumption that the axioms should be satisfied in *all* cases, as would be required if the observed data is assumed to be been generated from a deterministic conjunction of learners' proficiencies and item difficulties. Rather, as set out in the methodology section of the paper, a stochastic approach is used that aims to make allowance for the existence of measurement error.

The concept of measurement error is problematic within RTM. Krantz et al. (1971) describe it as 'poorly understood', and Luce and Narens (1994) note that 'alternative approaches to random variable ideas of error may be appropriate for extensions of RTM. In particular, applications of Boolean-valued and other multi-valued logics and fuzzy logics seem potentially interesting'. The notion of error is discussed further in Chapter 7. For the purposes of the analysis discussed in Chapter 4, the idea is essentially captured by supposing that observed results will depart from the 'true' values or levels of variables with some (ideally small) probability.

3.2 RQ2: Conceptually, can the quantitative paradigm of measurement that underpins psychometrics be extended to accommodate educational assessment constructs whose measuring procedures generate, in general, fuzzy, partially-ordered data?

This question is addressed in the paper presented as Chapter 5, published in *Frontiers in Psychology* in October 2024. It is a conceptual paper that does not rely on any technical material not already covered above, apart from (*a*) some basic notions from fuzzy logic, which are summarised in the Appendix to the paper, and explained in slightly more detail below, and (*b*) the elements of the theory of formal concept analysis, covered in §§3.3.1-3.3.2.

3.2.1 Lattices

Let $(L, \leq)$ be a poset and let S be a subset of L. An *upper bound* for S is an element $u \in L$ such that $s \leq u$ for all $s \in S$. u is the *least upper bound* or *supremum* of S, denoted $\vee S$, iff $u \leq y$ for all upper bounds y of S. Dually, a *lower bound* for S is a $l \in L$ such that $l \leq s$ for all $s \in S$, and l is the *greatest lower bound* or *infimum* $\wedge S$ of S iff $x \leq l$ for all lower bounds x of S. A *lattice* is a poset for which every two-element subset $\{x, y\}$ has a supremum $x \vee y$ and an infimum $x \wedge y$. A *complete lattice* has a *bottom* (least) element, usually denoted 0 or $\bot$, and a *top* (greatest) element, denoted 1 or $\top$, such that $0 \leq x \leq 1$ for all $x \in L$.

It is suggested in Chapters 5 and 6 that lattices arise naturally in studying the information generated by educational assessments, if one does not make the assumption that

assessment constructs must have quantitative structure. The possible values, or levels, or discriminable kinds, of a construct such as *GCSE mathematics attainment*, form a lattice, rather than a continuous line, in general. Note that a *ranking* and a *continuum* are both particular special cases of the more general notion of a lattice.

3.2.2 Truth degrees and fuzzy sets

The simplest lattice is a two-element poset with just a bottom and a top element. It can be represented as {0,1}, with $0 \leq 1$, or as {`false, true`}, with `false`$\leq$`true`. Because the structure is a lattice, the supremum $\vee$ and infimum $\wedge$ operations are defined, and, necessarily, `false` $\vee$ `true` = `true`, and `false` $\wedge$ `true` = `false`. So if we think of the elements `false` and `true` as labelling the truth-values of propositions in traditional (Boolean) propositional logic, then the operation $\vee$ corresponds to the logical connective 'or', and the operation $\wedge$ corresponds to the connective 'and'. For if the proposition P has the truth-value `false` and the proposition Q has the truth-value `true`, then the proposition 'P and Q' is false (it is not the case that both P and Q are true: symbolically $P \wedge Q =$ `false`); and the proposition 'P or Q' is true (it is the case that at least one of P and Q is true: symbolically $P \vee Q =$ `true`).

This idea of representing truth-values as lattices can be extended to the case of *fuzzy* or *multivalent* logics (Hajek, 1998; Bělohlávek et al., 2017). Such logics allow propositions to be drawn from ordered set of *truth-degrees*, that can be more extensive than just {`false,true`}. For example, possible sets of truth-degrees could be {`false, partly-true, true`}, or {`clearly-false, partly-false, partly-true, clearly-true`}. To be used in models of logics, truth-degrees need some ordinal structure (for example, the sets in the preceding two sentences are listed in order of 'increasing true-ness'). They also need some compositional structure, to allow truth-degrees to be logically aggregated across propositions (for instance if the proposition P_1 is true to degree t_1, and P_2 is true to

degree t_2, then one wishes to express the truth-degree of the proposition 'P_1 and P_2' in terms of a composition of the truth-degrees t_1 and t_2).

To accommodate these notions of order and composition on truth-degrees in a fairly general way, it is normally assumed that truth-degrees have the structure of what is known mathematically as a 'complete residuated lattice' (see e.g. Bělohlávek et al., 2017). This allows them in general to be only partially, rather than totally, ordered, and to have sufficient compositional structure to allow the evaluation of truth-degrees of conjunctions, disjunctions, and implications from the truth-degrees of their component propositions.

For the purposes of this study, and the methods of analysis of educational assessment information developed further in Chapters 5 and 6 in which truth-degrees are used to represent the degree to which learners' performances on assessment tasks have construct-relevant attributes, it is assumed that truth-degrees form totally ordered sets. In practice such sets are often represented by numbers. So in Boolean logic (relevant to the case of dichotomous attributes in an assessment context) one often uses {0,1} instead of {`false`, `true`}. A four-valued set of truth degrees might be represented by $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. This is a convenient way of representing the ordinal structure of truth-degrees by means of the standard order relation on numbers, but it must be remembered that the compositional structure does not necessarily correspond to the standard multiplication or addition operations on numbers.

If L is a lattice of truth-degrees, then an *L-set* or *L-fuzzy set* in a universe set U is a map $A : U \to L$. The value $A(u)$ is the degree to which u belongs to A.

3.2.3 Combining truth degrees

If the lattice of truth-degrees is taken to be the unit interval $\{x \in \mathbb{R} : 0 \leq x \leq 1\}$, then for propositions P_1 and P_2 with truth-degrees t_1 and t_2 respectively, one can evaluate the truth-degree of 'P_1 and P_2' as $\min\{t_1, t_2\}$; of 'P_1 or P_2' as $\max\{t_1, t_2\}$; and of 'not P_1' as $1 - t_1$. This is a common approach to applying fuzzy logic in contexts where one conceptualises truth-degrees, analogously to (Kolmogorovian) probabilities, as taking values between 0 and 1. For example, it is used in the 'fuzzy linguistic variable' approach to psychological constructs adopted by Buntins, Buntins, and Eggert (2016b) that is discussed in Chapter 5. More generally, other ways of representing logical connectives can be used, when the lattice of truth-degrees is taken to have another structure, such as a discrete totally ordered set: see Appendix D for more detail.

3.3 RQ3: Can non-quantitative methods of measurement provide more valid or useful information about learners' attainment or poficiency with respect to educational assessment constructs than quantitative approaches?

Formal concept analysis (Ganter and Wille, 1999) is an important development of lattice theory, that has been applied extensively to fields such as linguistics, political science, information sciences, medicine, and genetics. As noted in §2.6.3, applications to educational assessment have been limited to date, although the development of knowledge space theory can be seen as a special case. Chapter 6 explores its application to the kinds of assessments used to award GCSE and A level qualifications in England, building on the theoretical framework developed in Chapter 5.

Much of the discussion in Chapters 5 and 6 is self-contained, but some background on formal concept analysis is summarised here.

3.3.1 Dual representations of concepts: extents and intents

Formal concept analysis (FCA) can be thought of as a way of representing the information that is implicit in a matrix that relates objects to attributes, of the kinds illustrated in §2.3.2.3. FCA provides methods to to extract the concepts and implications that can be deduced from the dataset, and introduces a logic to reason and infer new knowledge. In this respect FCA is analogous to other knowledge-discovery techniques applied to datasets, such as machine learning. However, because of its logical underpinning, FCA is more suited to providing explainable results.

In FCA, a (formal) *concept* has both an *extent* (or *extension*) and an *intent* (or *intension*). The extent of a concept is the collection of objects that the concept describes. The intent of a concept is the collection of attributes that describe the concept. For instance in an educational assessment context, a concept of interest might be the concept of *grade A performance.* The extent of this concept is the collection of all grade A performances in the context in question. The intent of the concept is the set of attributes that characterise those performances (such as {`displays-knowledge-of-texts, displays-technical-accuracy, argues-convincingly, ...`}).

The extent/intent distinction is an old one in logic. Kneale and Kneale (1962) note how the seventeenth century *Port Royal Logic,* which remained a standard text up to the mid-nineteenth century, introduces a difference between 'comprehension' and 'extension'. Frege (1892) famously discussed the *Sinn* (sense) and *Bedeutung* (reference) of propositions, in his pioneering work on the foundations of mathematics, though Carnap (1947) was the first to use the terminology 'intension' and 'extension' systematically.

When classical (Boolean) logic is used to describe the extent to which attributes apply to objects (for instance in assessment contexts with dichotomous items), formal concept analysis can model the classical (rules-based) theory of concept formation as instantiated in marking schemes (see §2.3.1). When the underlying logic allows membership of extents/intents to be fuzzy, formal concept analysis may accord better with the prototype (pattern-based) theory (this claim is discussed from various viewpoints in Bělohlávek and Klir, 2011). In the fuzzy case, a concept such as *grade A performance* or *level 2 response* is not characterised by a list of necessary and sufficient criteria (explicit marking rules), but rather through a collection of attributes that tend to be characteristic (with varying degrees of truth) of the performances classed as *grade A* or *level 2*. This seems to accord with Wittgenstein's well-known 'family resemblance' theory of concepts, which repudiated the idea the concepts have 'sharp edges', and that precise specification is a requirement of meaning (Wittgenstein, 1953, §74).

3.3.2 Concept lattices

Formal concept analysis thus models a concept as a pair (extent, intent) of mutally-related sets[3]. Applied to a data matrix such as the output of an educational assessment procedure, it extracts all possible such (extent, intent) pairs. Moreover, using the usual set-theoretic relation of *inclusion*[4] orders these formal concepts from *most general* to *least general*. In general the order relation is a partial rather than a total order, and the formal concepts extracted from a data matrix form a lattice.

When the data being studied is derived from learners' responses to assessment tasks, classified as to their value with respect to the collection of attributes that have been

[3] In standard formal concept analysis, with binary (dichotomous) attributes, each element of the (extent, intent) pair that represents a concept is a classical set In fuzzy formal concept analysis, with fuzzy attributes, extents and intents are fuzzy sets.

[4] A set A is *included in,* or a *subset of,* a set B, written $A \subseteq B$, iff for all $a \in A$, it is also the case that $a \in B$.

deemed relevant to provide valid information about the construct being assessed, it is helpful to consider the relation *better than,* rather than *more general than.* The *most general* concept is one that has the largest extent (and smallest intent) in the collection. This corresponds to the concept of *bottom performance* or *worst performance,* since every other performance has displayed at least some additional attribute. The *least general* concept has the smallest extent (and largest intent), and therefore corresponds to the concept of *top performance* or *best performance.*

Each formal concept represents a way in which learners' responses can demonstrate value, given the parameters of the assessment procedure (the tasks and mark schemes). Some of these may represent concepts of substantive interest (such as the concepts of *top* and *bottom* performance), while others may not. When the attributes are not binary but fuzzy, this notion of formal concepts representing possible outcomes of an assessment formalises Cresswell's (2003) idea that examination outcomes are fuzzy sets[5].

This leads the question of construct validity. If certain responses are classified by the assessment procedure as being worthy of a certain level or grade, to what extent do the formal concepts associated with those responses match up with the intentions of the assessment designers, in terms of the *qualitative performance standards* intended to broadly characterise performance at that level? This is the kind of question it is hard to study using standard quantitative, psychometric, methods. Chapter 6 shows how a method of matrix-factorisation for truth-degree data, somewhat analogous to principal components analysis for quantitative data, can help address it.

[5] Incidentally, the so-called *sorites paradox* (or 'paradox of the heap'), whose ethical consequences for grading public examinations were considered by Cresswell (2003) – in an analysis that ultimately led to the 'comparable outcomes' framework that is used in grading GCSEs and A levels, and discussed in Chapter 6 – is resolved through the application of fuzzy logic (Bělohlávek et al., 2017, p. 333).

3.3.3 Quantitative linear models, linear operators, and eigenvectors

Many quantitative statistical models used in psychometrics and the analysis of test and examination data are *linear*, that is, they postulate that the values of (possibly transformed versions of) the dependent variables are linear combinations of the values of the independent variables. For example, given the scores $y_1, y_2, \ldots, y_m$ of m students on a test question, a linear model might postulate that these scores are related to n explanatory values or parameters $x_1, x_2, \ldots, x_n$, plus or minus some variation due to 'random error', by the equations

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ &\vdots \\ y_m &= a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_{ni}. \end{aligned} \qquad (\star)$$

Equations such as $(\star)$ are more conveniently written in matrix form. An $m \times n$ real *matrix* A is an array of real numbers with m rows and n columns. The entry of A in row i and column j is denoted a_{ij} or A_{ij}. A $1 \times m$ matrix is called a (column) *vector*. Two matrices A and B that have the same number of rows and columns are added by setting $(A+B)_{ij} := a_{ij} + b_{ij}$. Two matrices of shapes $m \times k$ and $k \times n$ are multiplied by setting $(AB)_{ij} := \Sigma_{l=1}^{k} A_{il}B_{lk}$.

Thus we can write equations $(\star)$ as $y = Ax$, where y is the vector $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$, x is the vector $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$, and A is the matrix with (i,j) entry a_{ij}. In general there is a one-to-one correspondence between linear functions (often called linear *operators* when they relate vectors rather than numbers) and matrices. Properties of linear operators can therefore often usefully be studied using concepts often initially introduced in statistics textbooks

with reference to matrices (and conversely, properties of matrices can be thought of more generally as properties of linear operators).

Geometrically one can think of an $m \times n$ matrix A as a mapping from n-dimensional to m-dimensional Euclidean space (from $\mathbb{R}^n$ to $\mathbb{R}^m$), such that each vector $x \in \mathbb{R}^n$ is mapped to a vector $Ax \in \mathbb{R}^m$. If $m < n$, this can be thought of as a reduction of the dimensionality of data consisting of a collection of vectors in $\mathbb{R}^n$, such as a collection of the vectors of scores of students on the items comprising an n-item test. As noted in Chapter 6, quantitative latent variable methods in practice often boil down to data reduction of this kind, with the information generated about students by an assessment—their item scores—being replaced with a low-dimensional vector, or often a one-dimensional vector (a number).

For square matrices (for which the number of columns equals the number of rows) an important concept is the notion of an *eigenvector*. These are vectors which are unchanged by the operation of the matrix, save for being *scaled* by some amount, known as the *eigenvalue* associated with that eigenvector. That is, if v is an eigenvector of a square matrix A with associated eigenvalue λ, then $Ax = \lambda x$. Geometrically, the vector x may change in length (and/or reverse direction, if λ is negative), but is otherwise unchanged by the transformation described by A. If the $n \times n$ matrix A is additionally *symmetric* (i.e. $a_{ij} = a_{ji}$: its enries are reflected in its main diagonal), then it is a theorem in linear algebra that A has precisely n real eigenvalues (and associated eigenvectors). Chapter 6 discusses the importance of this result for a particular instance of symmetric matrices, the covariance matrices for item responses in tests scored quantitatively, and considers a generalisation to matrices whose entries are taken to be truth values, rather than real numbers.

3.4 Ethical considerations

3.4.1 Data use

This study analysed historical, anonymised, student response data from some high-stakes (GCSE and A level) examinations in England. The datasets were provided by AQA Education, and it was not possible for the researcher to identify individual students. Application for the use of the data was considered by the Research Ethics Committee of the Department of Education at the University of Oxford, in accordance with the procedures laid down by the University. The research was judged as meeting appropriate ethical standards, and accordingly, approval was granted (ref. CUREC-ED-CIA-18-185).

3.4.2 Researcher positionality

The author of this study has worked for many years in the educational assessment sector in England, initially as a quantitative specialist contributing to the design and evaluation of assessment instruments and procedures, and for the last ten years as a senior executive at a large awarding organisation (itself an independent education charity), and regular interlocutor with policy makers, government, and the media. He has been influenced by the need, in this context, to balance theory (e.g. accounts of what might constitute the optimal approach to a particular assessment issue) and practice (e.g. in relation to the exigencies of operational delivery, especially of high-volume, high-stakes qualifications under time pressure). As an exam board 'insider researcher', to use the terminology of Gray (2020), he has at times been constrained by commercial and/or national political interests in fully disclosing the knowledge generated through his research.

The author's original training in mathematics, rather than a social science, has most likely shaped his views as to the nature of *representation* of constructs such as *proficiency in*

an educational domain. He has been influenced both by realist (e.g. Michell, 1999a) and essentially anti-realist (e.g. van Fraassen, 2008) accounts of scientific measurement, and his own position with respect to the ontological status of psychological and cognitive phenomena, and the relation between mind and brain, continues to evolve.

4 [Paper 1] Applying stochastic conjoint measurement checks to UK public examination data

ABSTRACT

Educational assessment data is usually analysed on the assumption that it reflects underlying proficiency or attainment factors that have quantitative structure. This study investigated empirically the extent to which such an assumption is tenable for some high-stakes summative assessments in England. It used a methodology that tests stochastically whether pre-conditions for variables to be validly measured on an interval scale are likely to hold. The results suggest that requirements for quantitative structure were generally not met, although the extent of deviation varied by subject. Some challenges of applying the methodology to assessments with polytomously-scored items are discussed. In the light of the findings, a key area for further research is to compare alternative structural approaches to measurement, that do not assume the assessment construct is quantitative, with existing approaches that are generally (if uncritically) accepted by teachers and students. The trade-offs between validity and pragmatic useability are likely to vary considerably, depending on the assessment domain.

4.1 Introduction

4.1.1 The assumption of quantitative structure in educational measurement

For many educational assessment procedures, including end-of-course examinations of the kind considered in this study, the observed data consists of item marks (scores): ordinal categorisations of the quality of student's responses to tasks. These categorisations reflect the judgements of experts, having regard to agreed criteria (so-called *mark schemes* or *rubrics*). In some subjects (such as mathematics) the intersubjectivity of these judgements may be close to maximal, so that in practice the item scores can be regarded as 'objective' ordinal classifications. This is also the case when the assessment tasks presented to the student are items that can be scored without the applicaton of any expert judgement, such as multiple-choice questions, where the simplest possible ordinal classification (correct *vs* incorrect) obtains. In other subjects (for example English writing assessed by means of essay questions), it is recognised that there is less inter-rater agreement, and the item scores are thought of as more subjective or 'fuzzy'.

A student's ability score (an hypothesised latent quantity) may be estimated from these observed multi-ordinal data, by applying a psychometric model such as an item response model. In this way students' *qualitative* responses to the totality of the assessment tasks are, apparently, converted or reduced to numerical *quantities*. If (making the assumption that ability is a continuous quantity) the Rasch model is applied (Rasch (1960), Andrich and Marais (2019b)), then students' abilities correspond one-to-one with total scores, which justifies the familiar method of classifying students by their total score on the assessment. A consequence is that qualitatively different performances on an assessment (for example performances displaying different profiles of strengths and weaknesses) are considered to be equivalent, provided their item marks sum to the same total.

Yet Scharaschkin and Baird (2000) found that there was a significant effect of *how* students had arrived at a particular mark total on assessors' holistic judgements of the merits of students' responses (the nature of their item-mark profile, over and above the total mark itself, was significant). They found that qualitatively different performances with the same total score could be regarded as warranting different measures of attainment. This is perhaps a symptom of a more fundamental objection to the uncritical use of quantitative and psychometric methods in educational and psychological measurement raised particularly by Joel Michell (e.g. 1997, 1999, 2002, 2005, 2006, 2009, 2012, 2013), but also more widely in the literature (e.g. Heene (2013a), Kyngdon (2011), McGrane and Maul (2020)).

The objection is that one needs to demonstrate that it is reasonable to suppose that the phenomenon of interest – such as *ability* in a subject area or curriculum domain – has the particular mathematical structure of being a *quantity*, before one is justified in applying techniques, such as item response theory, that simply assume it is. For instance Michell notes the implausibility, on the grounds of cognitive theory, of supposing that mathematics ability is a quantity, suggesting rather that 'abilities are attributes composed of ordered hierarchies of cognitive resources, the differences between which are heterogeneous' (2012, p. 265). As observed by Baird et al. (2017), the psychometric literature does not, for the most part, theorise the concept of 'ability' by deriving the assumptions that underpin how it is mathematically modelled from substantive theories of learning, teaching, or cognition.

Indeed Maul (2017b) notes that the use of quantitative methods in the social sciences is often grounded in a conception of *measurement* (namely, the discovery or estimation of the values of continuous quantities: the location of measurands on a numerical continuum) that is not even universally true in the physical sciences. He argues that in general scientific progress begins with articulation of substantive theories and development of

methods and models capable of testing aspects of those theories, and there is no *a priori* reason to assume that the phenomena of interest will necessarily be usefully treated as having quantitative structure.

van van Fraassen (2008) generalises the notion of measurement, from *location on a numerical continuum*, to *location in a logical space*. This provides a potential way forward in the event that the phenomenon of interest is not necessarily theorised as a quantity. For example, in order to measure educational attainment in the sense of locating it within a logical space describing an educational context or domain (for example, to certify a *level of competence* with respect to that domain), Scharaschkin (2017) suggests applying elements of mathematical order theory and fuzzy logic to the data generated by educational assessment procedures. This conceptualises the relevant logical space for an assessment procedure as a certain type of network, or partially-ordered set (for example, a so-called fuzzy concept lattice: see Bartl, Belohlavek, and Scharaschkin (2018)). From a purely conceptual viewpoint, therefore, it is neither necessary, nor necessarily credible, to assume that the phenomona studied in educational assessment procedures have quantitative structure, even though in practice this assumption is rarely challenged.

4.1.2 Aims of this study

Leaving aside concerns about the theoretical and conceptual adequacy of an uncritical acceptance of assumptions about hypothesised quantitative latent variables, however, it is possible in some circumstances to investigate empirically the question whether properties of interest have quantitative structure. This can be done using results from so-called representational measurement theory (Krantz et al., 1971), specifically, theorems regarding conjoint measurement. The following section summarises the basic ideas that allow the development of a methodology for testing whether a property is quantitative, and hence, in educational measurement terms, can validly be measured on an interval scale.

The aim of this study was then to apply that methodology to data from high-stakes summative assessments in England (A level and GCSE examinations).

The development of stochastic methods for investigating conjoint measurement is fairly recent, and to date there have been no applications to large scale assessments such as A level examinations in England. The literature (e.g. Karabatsos (2001b), Karabatsos (2018), Domingue (2014b)) has tended to focus on simulated data or on relatively small datasets, for illustrative or exploratory purposes. As well as exploring the research question of whether the common (if often implicit) assumption that educational attainment can be measured on an interval scale is likely to be valid, for these high-stakes assessments, a second question of interest was therefore to explore the challenges of testing for quantitative structure in properties that are assessed using constructed responses. The literature to date has focused on assessments made up of dichotomous items (actually rather rare in practice, in the UK context), so theoretical and practical implications of extending the approach to polytomous items were also investigated.

4.2 Theoretical framework

4.2.1 Quantities, continua, and scales

A *property* or *attribute* is a way of assigning descriptions, levels, or values to objects, such that the descriptions or values may admit relations of comparison, and/or operations of combination. A property is *quantitative* if it describes objects in a way that allows them to be compared in a consistent way as to their magnitudes or levels, and for their magnitudes or levels to be combined in a way that mirrors numerical addition. A precise technical definition of what is meant by the statement that a property is quantitative can be given quite concisely (Michell (1990a)), namely: a property is quantitative if its

value-set is *totally ordered* and *additive*. A quantitative property is also *continuous* if its possible values form a continuum with no 'gaps'. Formally, this notion is captured by requiring the value-set of a continuous quantity to be *dense* and *complete*. These technical definitions are unpacked in more detail in Appendix A.

It can be shown that any system of objects and relations satisfying the technical conditions required for the values of a continuous quantity is isomorphic (structurally identical) to the real number system (see e.g. Munroe, 1965). And of course the real numbers, geometrically, model the Euclidean line.

So if a property is a continuous quantity, it can be thought of geometrically as a continuous line. Obtaining the *value of a quantity* (for example the length of a rod, the volume of a ball, or the voltage of a battery) is equivalent to marking a specific *location on a line*. Whereas obtaining the value of a property that is not a quantity is equivalent to determining a specific location in a different kind of space (for example, the spin of an electron can take one of two possible values, so specifies a location (one or other of the two points) in a two-element space).

As is well known, Stevens (1946a) introduced the terms nominal, ordinal, interval, and ratio scales, to describe the structures of certain possible logical spaces (to use van Fraassen's terminology) in which the values of the kinds of properties typically studied in psychology could be located. A *ratio scale* allows the estimation of ratios between a magnitude of a continuous quantity and a unit of measurement of the same kind. As Michell (2021b) notes, it was demonstrated at the beginning of the 20th century by Otto Hölder that the system of all ratios of magnitudes of an unbounded continuous physical quantity (such as length or mass) is isomorphic to the system of positive real numbers. An *interval scale* represents what in mathematics is called the affine line – effectively 'the real numbers without the origin (zero-point)'. In an interval scale, differences (intervals) between points are defined, and the collection of all differences is also isomorphic to

the positive reals. Thus, by stating (or assuming) that a property is 'ratio-scaled' or 'interval-scaled' (often called 'equal-interval scaled'), or that a ratio- or interval-scale representation exists for the property, one is in fact claiming (or assuming) that the property is a continuous quantity.

4.2.2 Additive conjoint measurement

Luce and Tukey (1964b) introduced the notion of *conjoint measurement* as a potential method of enabling quantification in situations in which a totally-ordered property[1] arises as the conjunction, or joint effect, of two or more other properties.

In educational assessment an example could be the *quality of a student's response to a task* (the *level of attainment* that their response merits or demonstrates), conceived of of as the joint effect of a *property or attribute of the student*, that might be called 'ability', and a *property of the task*, that might be called 'difficulty'.

In general, as set out in Krantz et al. (1971), conjoint measurement studies structures that, in the two-component case, are modelled as a pair (X, Y) of sets, thought of the possible values, or levels, of each of two attributes or properties $\mathscr{X}$ and $\mathscr{Y}$,together with a total order relation $\succeq$ on the set $X \times Y$ of ordered pairs of values of the attributes. The aim of additive conjoint measurement (ACM) is to represent the 'empirical relational structure' $(X, Y, \succeq)$ by means of an additive numerical relational structure, i.e., to establish conditions under which it is possible to construct order-preserving maps $\phi_1 : X \to \mathbb{R}$ and $\phi_2 : Y \to \mathbb{R}$ such that

$$(x_1, y_1) \succeq (x_2, y_2) \Leftrightarrow \phi_1(x_1) + \phi_2(y_1) \geq \phi_1(x_2) + \phi_2(y_2). \tag{4.1}$$

[1] A property is said to be totally ordered if there exists a binary relation $\succeq$ on its value-set (i.e. there is a way of comparing its values), such that (i) it is always possible to compare any two values (either $x \succeq y$ or $y \succeq x$, for all values x and y); (ii) comparisons are transitive (if $x \succeq y$ and $y \succeq z$, then $x \succeq z$); and (iii) comparisons are antisymmetric (if $x \succeq y$ and $y \succeq x$, then $x = y$).

If a representation of the form 4.1 exists, then properties $\mathscr{X}$ and $\mathscr{Y}$, as well as the property of their joint effect, can be regarded as real numbers. A measure of the joint effect of $\mathscr{X}$ and $\mathscr{Y}$can be obtained by summing their values. The key representation theorem of additive conjoint measurement sets out necessary conditions (sometimes referred to as axioms for conjoint measurement) for a representation of the form (4.1) to exist. These axioms, therefore, provide a basis for testing, empirically, whether phenomena such as cognitive ability or educational attainment have quantitative structure, and hence whether it is warranted to draw conclusions by subjecting such phenomena to analytical methods that assume they can be represented on interval scales. Essentially the axioms isolate the minimum structural features necessary in the empirical relational system for an additive representation of the kind defined in equation (4.1) to exist.

The five ACM axioms are summarised in Appendix A. It is proved in Krantz et al. (1971a) that if these axioms hold, then interval scales exist for X, Y, and $X \times Y$.

So if one is given three properties constituting an empirical relational system in which the third property is conceptualised as the joint effect of the other two (for instance the probability of a correct response to an assessment item as the joint effect of the ability of the candidate and the difficulty of the item), then the axioms define sufficient conditions for all three of these properties to have quantitative structure.

In other words, if $\mathscr{X}$and $\mathscr{Y}$are properties whose values are in the sets X and Y respectively, then *testing whether the axioms of additive conjoint measurement hold* amounts to *testing whether $\mathscr{X}$,$\mathscr{Y}$, and the property that represents their 'joint effect', are all quantitative*. Empirical evidence as to whether a given property is quantitative, therefore, can be obtained via such tests.

4.2.3 Stochastic checks of the axioms

The first application of additive conjoint measurement to empirical educational assessment data appears to be the study conducted by Perline, Wright, and Wainer (1979). This directly tests whether some of the ACM axioms hold for a small dataset (490 respondents to 9 items). Michell (1990a) gives a careful account of applying conjoint measurement checks in a number of psychological measurement contexts, including a re-analysis of Thurstone's (1927) application of the method of comparative judgement to measure seriousness of crimes[2].

It is not feasible, however, to perform all the necessary checks for conjoint arrays of the size encountered in many practical applications. The A level economics data array analysed below, for instance, has 31 rows and 30 columns. For an $n \times m$ array, there are $\binom{n}{3}\binom{m}{3}$ tests of the double cancellation axiom (axiom 3 in Appendix A). For the 31-by-30 case, that amounts to over 18 million tests: too large a number to check individually. For this reason, and also to take account of presumed 'random error' in the observed data, stochastic methods have been developed to sample from the universe of all possible tests.

Domingue (2014b) considers the case of educational tests consisting of a sequence of n dichotomous items, so that each test-taker can obtain a total score drawn from $\{0, 1, ..., n\}$. The results of such a test can be summarised as an $n \times n$ matrix in which the entry p_{ij} in cell (i, j) is the proportion of correct responses to item j from students who obtained a total score of i. Such a matrix is regarded as providing an empirical sample estimate of the 'underlying' (unobservable) probabilities that students with a level of (latent, unobservable) ability proxied by a total score of i correctly answer item j. That is, the matrix of observed proportions of correctly answered items is taken as an empirical relational

[2]Michell's conclusion was that 'either seriousness of crimes is not a quantitative variable, or else some other part of Thurstone's theory of comparative judgement is false'.

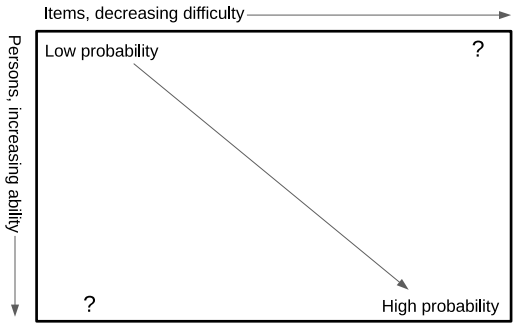


Figure 4.1: Data matrix for conjoint measurement checking

system, in which each proportion p_{ij} is assumed to arise as the conjoint effect of the two properties of student ability (proxied by total score) and item difficulty (proxied by the overall proportion of responses to the item that were correct).

Figure 4.1 (from Domingue, 2014, p.4) shows the set-up. Rows are ordered by increasing ability, so the lowest ability individuals are at the top. Columns are arranged by decreasing difficulty, so the hardest items are at the left. Consequently, the lowest probabilities of correct responses occur at the top left, and the highest probabilities at the bottom right. If ACM axiom 2 (single cancellation) holds, then the probability of a correct response must increase as one moves on any path down and right through the cells. However, it is less clear how probabilities would be expected to change if one moves at right angles to the main diagonal. If ACM axiom 3 (double cancellation) holds, then certain orderings must obtain between these probabilities. If it is found that they do, then that provides evidence in support of the hypothesis that ability and difficulty are quantitatively structured, and can be measured on interval scales. If the orderings do not hold, then the hypothesis of quantitative structure for these properties is not supported.

Domingue's stochastic method tests for violations of single and double cancellation, simultaneously, given an input matrix of the form of Figure 4.1. It uses a Bayesian ap-

proach, building on the methodology initially developed by Karabatsos (2001b). The approach is to enforce the ACM axioms stochastically, and then to determine whether the resulting estimates for the probability of a correct response are reasonable, given the observed data. The ACM cancellation axioms need to hold for every 3×3 submatrix of the full data matrix (because the double cancellation axiom has to hold with respect to any three values of $\mathscr{X}$ and $\mathscr{Y}$ in the conjoint structure: in this case, any three values of ability and difficulty). As noted above, for realistic numbers of items and ability categories, there are potentially hundreds of millions of possible 3×3 matrices that would need to be checked. Hence, tests are carried out on a sample of all possible 3×3 submatrices of the full matrix.

In order to assess the extent to which the results of this sample-based testing support a hypothesis of quantitatively-structured variables, Domingue calculated benchmark values for the proportions of sampled submatrices in which cancellation axioms would fail to hold simply due to presumed measurement error in the observed data. These are obtained by using simulated data that is constrained to fit the Rasch model, and hence, necessarily, satisfies the ACM axioms[3], except to the extent that random error enters the set-up (that 'sum scores are not perfectly mapped to the [assumed] true underlying abilities', as Domingue puts it). Running the stochastic checks methodology on Rasch baseline data provides values for the proportion of cases in which ACM cancellation axioms are not met. Thus if in running the checks on observed empirical data, it is found that the proportion of cases in which the cancellation axioms are not satisfied are similar to these baseline values, then there is no empirical reason to reject the hypothesis of un-

[3]The Rasch model says that the conjoint effect of real numbers called ability and difficulty, namely the log-odds of a student with a given ability corectly answering an item with a given difficulty, is an additive function of the student's ability and the item's difficulty (in fact simply the difference between the two). Constraining simulated data to fit the equation that defines the Rasch model therefore necessarily generates an additive conjoint system, modulo some random error. That does not mean, as some have claimed (e.g. Brogden (1977)), that the Rasch model is a 'probablistic version' or 'instantiation' of additive conjoint measurement (see, e.g. Kyngdon (2011)). Relevant inferences that can be drawn from fit, or lack of fit, of a given set of data to the Rasch model are discussed in section 5.

derlying quantitative structure (equivalently, that the observed data matrix is consistent with variables that form an additive conjoint system). If, on the other hand, the empirically observed proportions of submatrices for which the axioms fail are substantively larger than these baseline values, then the hypothesis that the observed data results from variables with quantitative structure is unlikely to be warranted.

Two approaches to sampling are implemented in Domingue's methodology. The first, which is relatively time-intensive, is to sample at random from the full matrix. The second is to consider only 3×3 matrices that are formed using adjacent rows and columns (when the full matrix is ordered as in Figure 4.1). The idea behind this 'adjacent' method of sampling is that, in general, the most stringent single-cancellation tests are those formed by considering adjacent cells.

Domingue's method has been implemented in the statistical programming language R, as the package ConjointChecks (Domingue, 2012), and it is this package that was used in the applications that follow.

4.3 Methodology and data

The data considered in this study were taken from June 2018 public examinations in England run by the awarding organisation AQA. These examinations are used to award qualifications known as A levels ('advanced levels'; mainly taken by pupils aged 18) and GCSEs ('general certificate of secondary education', mainly taken by pupils aged 16). Two A level subjects were considered: physics and economics. Two GCSE subjects were also considered: English language and mathematics.

4.3.1 A level assessments

Physics and economics were chosen at A level because (relatively unusually in the the UK context), the assessments are composed in part of multiple-choice items. Both of these subjects are assessed by means of three two-hour written examination papers. Paper 1 of A level physics consists of 25 multiple-choice items, and a sequence of short constructed-response items for which the total possible mark is 60. The paper 1 multiple-choice items thus form a subtest that is intended to have a weighting of about 29% of the paper, which is equivalent to 10% of the assessment as a whole. For economics, paper 3 consists of 30 multiple-choice items and 50 marks worth of short constructed-response items. Thus in this case, the paper 3 multiple-choice items form a subtest with an intended weight of 37.5% of the paper, or 12.5% of the assessment as a whole.

There were 18,981 complete responses to the 25-item physics assessment, and 11,463 responses to the 30-item economics assessment. Because these assessement components consist of dichotomous items, they are amenable to analysis using the ConjointChecks procedure as described in section 2.3, and the results of these analyses are summarised in the following section.

The tests were designed to cover the content and assessment objectives for these subjects as defined by the government, and items were written with that aim in mind, and with the intention of providing a reasonable spread of difficulty as well as coverage of the syllabus. They were not pre-tested (it would not be permissible for awarding organisations to pre-test items for public examinations under current regulations in England in respect of confidentiality and predictability of examination content), nor were the tests constructed with the aim of obtaining a good fit to any particular psychometric model.

In the interests of exploring potential quantitative structure, however, Rasch models were fitted, and misfitting items were systematically removed until a reasonable fit was

obtained. Conjoint measurement checks were then performed on these 'pruned' assessments, and the results compared with those for the full assessments.

4.3.2 GCSE assessments

GCSE English language and mathematics were chosen because these qualifications are taken by nearly all 16 year olds in England (in total over 600,000 each year), for most of whom, including many of those who go on to take A level qualifications, the course of study leading to a GCSE in English or mathematics is the end of their formal study of the subject. GCSE qualifications are graded on an ordinal scale with nine categories, from grade 1 (lowest) to grade 9 (highest).

These assessments do not consist of sequences of dichotomous items. In this respect they are much more representative of the bulk of GCSE and A level summative assessment. A trained examiner, who has been 'standardised' with respect to the criteria for assigning credit to responses that are summarised in the mark scheme (rubric) for the item, marks each candidate's response to each task, that is, assigns it to one of a number of ordered categories labelled by integers. In GCSE mathematics assessments the tariff (number of possible evaluation categories) for an item is normally low (for example, 1, 2, 3, or 4 marks), and the mark scheme for the item gives rules for assigning partial credit to responses. Tariffs for GCSE English language items are generally high (for example, between 8 and 24 marks), and the mark scheme may indicate typical features or attributes of a response that would tend to place it in one of a small number (typically 4) 'levels', with a final mark then being awarded based on the examiner's judgement of whether the response is, for example, a good, average, or borderline example of attainment at that level.

The mathematics examination consists of three papers of one and a half hours each,

each with a total possible mark of 80. The English language examination consists of two papers of one hour and 45 minutes each, each with a total possible mark of 80.

The mathematics examination is tiered, that is, offered in a less demanding (foundation tier) version targeted at candidates expecting to get up to a grade 5, and a more demanding (higher tier) version targeted at candidates expecting to get up to a grade 9. The higher tier papers cover more content than the foundation tier. Vertical scaling between each foundation tier paper and the corresponding higher tier paper is carried out using items common to both tiers, in order to guide the awarding committee for the examination in setting grade boundary marks (cut scores) for the two tiers. Currently the approach to scaling is to use chained equipercentile equating. In past years, an item-response theory based scaling method was used, but was found to generate suggested grade boundaries that were not always credible to the expert examiners responsible for standard setting and maintenance.

This study focused solely on paper 1 for each subject. The procedure for checking the conjoint measurement axioms is computationally intensive, and there is no particular benefit in using the complete assessment when a self-contained component is sufficient to explore the issue.

There were 487,677 responses to the English paper, 117,587 responses to the foundation tier mathematics paper, and 89,416 responses to the higher tier mathematics paper.

The mathematics assessments at both foundation and higher tier consisted of sequences of short-answer tasks, each testing a particular content area with respect to a particular assessment objective.

The foundation tier paper had 40 items, and the higher tier paper 38 items, with the distributions of items by tariff (total possible mark for the item) shown in Table 4.1.

Table 4.1: Structure of GCSE mathematics papers

	Foundation	Higher
Tariff	Number of items	Number of items
1	13	15
2	18	8
3	7	12
4	1	2
5	0	1
6	1	0

Table 4.2: Structure of GCSE English language paper

Question	Tariff	Assessment objective
1	4	Identify and interpret
2	8	Explain and comment
3	8	Explain and comment
4	20	Evaluate text critically
5 (comms)	24	Communicate clearly
5 (accuracy)	16	Technical accuracy

To allow for vertical scaling there were nine common items (accounting for 21 possible marks in total) between the two papers.

The English language paper comprised five tasks, testing five assessment objectives, with mark allocations as in Table 4.2. Question 1 required students to list four things that could be inferred from a source text. The other questions required students to write connected prose responses. The final question required students to produce a piece of creative writing in response to a prompt. It was marked separately as to its quality in respect of the 'communicate clearly' objective, and as to the 'accurate use of language' objective.

The initial approach taken to analyse the polytomous responses for GCSE mathematics and English language, was to treat each polytomous item as a sequence of independently judged dichotomies, where 'independent judging' was simulated by drawing separate random samples without replacement from the overall dataset to estimate the proportion 'correct' (i.e., achieving the relevant threshold standard) for each dichotomised item. The ConjointChecks method was the applied to this transformed data.

This approach builds on the insights of Andrich (2013), (2016), expanding on his original derivation of the polytomous Rasch model (Andrich (1978b)), who notes that mathematically the polytomous case can be derived by imagining that the judgement that a response is in one of the n ordered categories for an item can be resolved into $n-1$ independent dichotomous decisions. Because of the relatively large number of entries for the subjects considered here, it was possible to simulate independent dichotomised judgements by firstly creating Boolean variables for each item with tariff greater than one (so that, for instance, a two-mark item is resolved into the dichotomies 'gets at least one mark, *versus* gets no marks', and 'gets two marks, *versus* gets less than two marks '). Then the proportions in each category, for each dichotomised item, was estimated by taking independent random samples without replacement from the full dataset.

As in the A level case, the assessments were not designed with the aim of fitting a particular psychometric model. But in the interests of exploring possible quantitative structure, polytomous Rasch models were fitted, and items with disordered threshold parameters were identified. This in itself enabled conclusions to be drawn about whether the assessment constructs could be reasonably conceived of as quantitative. The implications are discussed further in section 5.

4.4 Results

4.4.1 Physics and economics tests (dichotomous items)

Data from candidates sitting AQA A level physics and A level economics in June 2018 was analysed. The results of applying conjoint measurement checks to these datasets, using the ConjointChecks procedure summarised in the preceding section, are summarised in Table 4.3. The sample size used for the random sampling method was 3,000 for each subject.

Table 4.3: Percentage of sampled submatrices failing to meet ACM axioms

	Random sampling method		Adjacent sampling method	
	Mean (benchmark=2)	Wt. mean (benchmark=1)	Mean (benchmark=18)	Wt. mean (benchmark=16)
Physics	11.7	10.5	34.8	32.6
Economics	9.9	6.8	28.9	23.0

The benchmark figures in the table show the percentages of sampled submatrices that would be expected to fail to meet the ACM axioms, assuming that all variables have quantitative structure, but are measured with some error.

The outcomes in Table 4.3 are clearly considerably higher than the benchmarks, and in fact are of a very similar magnitude to those reported by Domingue (2014b) in his analysis of data generated by the Lexile reading assessment (Stenner and Fisher, 2013). Domingue's conclusion in respect of that analysis was that 'the data do not seem consistent with the axioms of additive conjoint measurement. A great deal of skepticism that this data could be use to create an interval scale seems warranted' (Domingue (2014b), p. 14).

The items that constituted these assessments were not pre-tested. It could be argued that, with the possibility of pre-testing and modification to items to get a better fit to Rasch, the assessment construct could be modified in such a way as to make it more likely to be validly represented as an interval-scaled quantity.

These datasets were therefore 'pruned' by removing items with a relatively poor fit to the Rasch model. Items were deemed to be misfitting if they had outfit values less than 0.8 or greater than 1.0 (these were also the criteria used by Domingue (2014) in his further investigation of the Lexile reading data). Items were removed sequentially, starting with the most over-discriminating item (with the highest outfit statistic). After removing an item, the model was re-fitted and item statistics re-calculated. The process was repeated until all items met the fit criterion. For physics, this required the removal of five items from the original 25-item assessment. For economics, only one item had to be removed.

Table 4.4 shows the results for the pruned data. The percentage of cases in which the axioms are not met is lower, for both subjects and sampling methods, than for the non-pruned data. It appears that the data is closer to what would be required to justify the construction of an interval scale on which students' attainment with respect to the assessment tasks could be located, particularly for the economics assessment, although still above the benchmark values.

Table 4.4: Percentage of sampled submatrices failing to meet ACM axioms (pruned data)

	Random sampling method		Adjacent sampling method	
	Mean (benchmark=2)	Wt. mean (benchmark=1)	Mean (benchmark=18)	Wt. mean (benchmark=16)
Physics	8.6	7.8	27.1	24.6
Economics	7.8	4.7	26.0	19.9

4.4.2 Mathematics and English tests

The data analysed comes from AQA GCSE mathematics and English examinations taken in June 2018. There were 487,677 responses to the English paper, 117,587 responses to the foundation tier mathematics paper, and 89,416 responses to the higher tier mathematics paper. To analyse the polytomous responses for GCSE mathematics and English language using the ConjointChecks procedure, each polytomous item was treated as a sequence of independently judged dichotomies, where 'independent judging' was simulated by drawing separate random samples from the overall dataset to estimate the proportion 'correct' (i.e., achieving the relevant threshold standard) for each dichotomised item. The ConjointChecks method was the applied to this transformed data.

By taking samples, we have smaller numbers of data points, and missing values for large and small total marks, so have to restrict consideration to the middle rows of the matrix, from total marks of 15 to 59 for the higher tier (15 to 58 for foundation tier) (for which the average number of total responses are all $\geq$10). The ConjointChecks procedure was then run using these proportions.

The same method was used for the English assessment, transforming the orginal six items into 80 dichotomous items. Clearly this is a much more radical transformation than for the mathematics assessments Because of the greater number of entries for English (487,677 in total), each of the 80 random samples drawn from the full data, for the purpose of estimating the proportion correct on each dichotomised item, was of size 5,000. That meant there were fewer cases of missing values for low or high total marks, and the ConjointChecks procedure was carried out for total marks from 4 to 70.

Table 5 shows the results. It can be seen that all results are well above the benchmark values.

Table 4.5: Percentage of sampled submatrices failing to meet ACM axioms

	Random sampling method		Adjacent sampling method	
	Mean (benchmark=2)	Wt. mean (benchmark=1)	Mean (benchmark=18)	Wt. mean (benchmark=16)
Mathematics (F)	14.5	10.9	40.1	28.1
Mathematics (H)	12.0	9.5	38.8	27.9
English	55.1	44.1	55.2	41.3

Table 4.6: Polytomous items with disordered thresholds

	No. polytomous items	No. with disordered thresholds
Mathematics (F)	27	22
Mathematics (H)	23	21
English	6	4

As in the cases of physics and economics, Rasch models (in this case polytomous models) were fitted to the data. These revealed disordered threshold parameter estimates for nearly all of the mathematics items, and for four out of the six English items (Table 4.6). Possible reasons for this, and implications for conclusions about the likelihood of the assessment constructs being interval-scaled, are discussed in the following section.

The two English items with well-ordered category thresholds were items 2 and 3: the two 8-mark items designed to assess the 'explain and comment on features of texts' assessment objective. The rubrics for these items require markers to assign the response to one of four levels: 'simple, limited' < 'some attempts' < 'clear, relevant' < 'perceptive, detailed' (with instructions to award zero marks if there is 'nothing to reward'). Some more detail about the attributes associated with a prototype or exemplar reponse in each level is included in the marking instructions. Two marks are allocated to each level, so

having determined a level (e.g. level 3, 'clear, relevant'), the marker then decides whether the response is a stronger or weaker example of what is required for the level, and awards the relevant mark (so either 5 or 6 marks, for a level 3 reponse).

This 'level of response' method of marking is applied to the higher tariff items as well, except that there are more 'fine gradations' within each level. So the 20-mark item has 5 marks available for each level, and the 24-mark item has six marks per level. The reason these items are higher tariff is because they are intended to have a greater weight in determining candidates' total scores.

The fact that the thresholds are well-ordered for the 8-mark level-of-response-marked items suggests that these are operating as intended in ordering candidates' responses. This is reflected in the dichotomised versions of these items, which display Guttman-pattern outcomes. Subjecting a subtest of the English paper consisting solely of these two 8-mark items (or 16 dichotomised items) to the ConjointChecks procedure revealed no violations of the single or double cancellation ACM axioms.

The polytomous mathematics items had tariffs of 2 or 3 marks in the majority of cases. The approach to marking is quite different from English. Generally speaking marks are awarded for method ('M', awarded for 'a correct method which could lead to a correct answer', according to the marking instructions), and for accuracy ('A', awarded 'when following on from a correct method. It is not necessary to always see the method. This can be implied'). Thus it is possible to have different profiles of method and accuracy marks for candidates classified as being in the same category within an item (e.g. profiles of M1M1A0 and M1M0A1 would both give a candidate 2/3 marks for a 3-mark item). As discussed further below, this feature of the way 'what good performance looks like' is conceptualised for GCSE mathematics may have contributed to the high levels of threshold-disorder that arise when the polytomous Rasch model is fitted to the data.

4.5 Discussion

4.5.1 Physics and economics assessments

The way scores are assigned to items in the high-stakes summative assessments considered in this study clearly generates (more-or-less fuzzy), partially-ordered data (the mathematical product of ordinal – totally-ordered – data for each item). Can this data reasonably be treated (that is to say, can the construct about which it is intended to give insight be treated) as (i) ordinal, and (ii) quantitative?

The answer to these questions would appear to be no, for the physics and economics examples. Even when pruned to obtain better fit to Rasch, the cancellation axioms that need to apply for the underlying property of ability to be quantitative, are not met. However, perhaps the most interesting result is for the pruned economics data. Although the assessment had not been designed with the aim of fitting a Rasch model, in fact it functioned reasonably well in those terms, and only one item was removed to obtain a good fit. The cancellation axioms were still violated in a proportion of the cases tested that was above the benchmark that would be expected simply due to measurement error – but not by as much as any in any of the other assessments considered in this study. One might tentatively conclude, therefore, that drawing conclusions from the economics assessment as if it provided quantitative proficiency measures on an interval scale is not unreasonable, and certainly less unreasonable than for the other subjects.

Overall, the results for the physics and economics tests illustrate the general point that the proposition: 'the data fits the Rasch model reasonably well' (within the normal pragmatic benchmarks for item infit and outfit, for instance, that would often be taken to establish a reasonable fit), does not necessarily entail that the supposed underlying construct is quantitative. On the other hand, if the property is quantitative, then,

because that entails the ACM axioms hold, a well-designed test should yield data that fits the Rasch model. That is, being quantitative is a sufficient, but not necessary, condition for the Rasch model to fit.

4.5.2 English and mathematics assessments

What if the data is a poor fit to the Rasch model? The valid conclusion to draw in this case is that either the underlying construct is not quantitative, or the tests were poorly-designed instruments for eliciting students' locations on a scale for an underlying quantitative construct (or both). Indeed the latter clause is is an underpinning tenet of Rasch measurement theory – as opposed to Rasch modelling – where a lack of fit to the model would express the fact that measurement is not possible in this case (because measurement is taken to require, by definition, location on a linear continuum).

We see this in both the GCSE English and mathematics examples. In English, the 8-mark items assessing AO2 yield clearly ordinal data, and a test that was just composed of these two ordinal items, if dichotomised, yields Guttman-pattern data. So the construct of 'AO2' does appear to be ordinal. The cancellation axioms are satisfied (as they must be for Guttman data when rows and columns of the (student, item) matrix are ordered): but this is not sufficient to prove that AO2 is quantitative (can be expressed on an interval scale). As noted by Zand Scholten (2011), Guttman data does not satisfy Archimedeanness (ACM axiom 5).

However, the (fuzzy) partially-ordered data generated by the whole examination paper in English cannot be approximated by an ordinal construct - and this is because of the lack of exchangeability of the notion of 'a mark' (a unit) between items. This problem is exacerbated by imposing fine-grained distinctions within the ordered levels of the levels-of-response mark scheme for the assessment, primarily for the purpose of *weighting* the

different items. Of course the whole notion of 'a percentage weighting' for part of an assessment (whether an 'intended' or an 'achieved' weighting: see Adams and Murphy (1982), Delap (1994)) already assumes an underlying quantitative structure. This point was recognised in part by Cresswell (1987), in a discussion that does not appear to have generated any further investigation in the English public examination context.

In mathematics, the conceptual process underlying the allocation of marks to items is quite different from English. Its focus, for most of the low-tariff, polytomous items, being on 'method' marks and 'accuracy marks'. Thus a mark of 1/2, for instance, could represent 'right/well-explained method, wrong answer', *or* 'right answer, wrong/poorly-explained method'. As seen in the threshold parameters from the polytomous Rasch model, there was generally not a progression in terms of most likely response to an item, conditional on an estimate of overall ability or quality-of-attainment, from 0 to 1 to 2 marks. These differences in how candidates respond in practice to the different kinds of demands made by the items may reflect Michell's (2012) suggestion that '[mathematics] abilities are attributes composed of ordered hierarchies of cognitive resources, the differences between which are heterogeneous'.

It may be that by changing or adapting the conceptual process of allocating marks for mathematics items (and, therefore, changing the notions of 'good' and 'betterness' for the assessment construct – in other words, changing the construct itself), it would be possible to develop a construct that does have quantitative structure, and hence is amenable to measurement on an interval scale. However, the degree of misfit to the polytomous Rasch model, and the results of the ConjointChecks procedure, suggest that as currently conceived, the construct of 'GCSE mathematics', or the property of 'GCSE mathematics ability', is not quantitatively-structured.

4.5.3 Substantive importance

These results are not simply matters of purely theoretical concern. The assessments considered here are high stakes. For instance GCSE qualifications in English and mathematics are often requirements for progression to employment, further or higher education.

Partly in recognition of this, the government has established national 'reference tests' for GCSE English language and mathematics, using GCSE-style items, that aim to measure changes in performance in GCSE English and mathematics over time, by comparison with a 2017 baseline (Ofqual (2019). This is done by estimating changes in the proportions of students at certain points on an ability scale (an interval scale), calculated by applying item response theory models to the data. The test administrators state that item response theory analysis 'creates a scale of latent trait on which test takers and test items can be placed. In educational assessment, the latent trait can be thought of as subject attainment' (Ofqual (2019), p. 13).

Thus the decisions taken on the basis of the reference test results, which can affect how awarding organisations are subsequently allowed to grade students in GCSE English and mathematics, assume that it is valid to treat the assesment constructs as having quantitative structure. This is another reason for critically examining this underpinning assumption for these assessments.

4.5.4 Limitations and areas for further research

There are clearly some conceptual issues that arise in using manifest (observed) variables – whose values are taken to be numerical – to proxy the hypothesised latent variables whose status as quantities is being tested. The justification for this manoeuvre is summarised by Kyngdon, 2011 as resting on the view that 'empirical structures can be based

on numerical entities provided they have been empririically obtained', which is the position taken by Luce and Narens (1994)[4]. However it should be noted as a clear limitation of the ConjointChecks approach.

Clearly, further investigation is needed in relation to these issues. These caveats notwithstanding, the evidence presented here does suggest empirically that the assumption of quantitative structure for educational attainment is problematic. To the extent that it is possible to transform response data for example by collapsing categories, and then treating polytomous items as approximated by sequences of dichotomised items, the results are suggestive that at least part of the reason for poor model fit is a lack of underlying quantitative structure in the measurand. But further research is needed on this.

Also more investigation is needed on the question of weighting different parts or aspects of assessment constructs, in the case that the assumption of quantitative structure is not justified. The results arising from the application of levels-of-response mark schemes may in fact be reasonably well handled in practice by simply using sum scores, *provided* the question of weighting is solved (see Fattore, 2016, for some discussion about weightings in the context of partially ordered data). Because of the Guttman pattern of responses, partial order methods (such as Guttman's own partial order scalogram analyis (Shye, 2009), or other approaches such as formal concept analysis (Bartl, Bělohlávek, and Scharaschkin, 2018) may not add much to existing methods. In other subjects where the criteria for classifying responses by level do not work in this way (as seen in the mathematics results here), order-theoretic methods may have more to offer. As ever in educational assessment, it is unlikely that a 'one size fits all' approach is appropriate.

[4] Luce and Narens (1994) were responding to Adams (1966) contention that 'IQs are not measurements because they do not establish a numerical representation of empirical operations and relations'. They claim that 'empirical structures can be based on numerical information just as long as those numbers arise from empirical means – which is the case for ability measurement. What is lacking from the RTM [representational theory of measurement] perspective are axiomatic theories for the various kinds of ability measurement. We do not see any in principle impediments to the development of such axiomatic theories, although in practice none have been devised and it does not appear to be easy to do.'

5 [Paper 2] Educational assessment without numbers

ABSTRACT

Psychometrics conceptualises a person's *proficiency* (or *ability*, or *competence*), in a cognitive or educational domain, as a latent numerical quantity. Yet both conceptual and empirical studies have shown that the assumption of quantitative structure for such phenomena is unlikely to be tenable. A reason why most applications of psychometrics nevertheless continue to treat them as if they were numerical quantities may be that quantification is thought to be necessary to enable *measurement*. This is indeed true if one regards the task of measurement as the location of a measurand at a point on the real number line (the viewpoint adopted by, for example, the representational theory of measurement, the realist theory of measurement as the discovery of ratios, and Rasch measurement theory). But this is not the only philosophically respectable way of defining the notion of measurement. This paper suggests that van Fraassen's more expansive view of measurement as, in general, *location in a logical space* (which could be the real continuum, as in metrological applications in the physical sciences, but could be a different mathematical structure), provides a more appropriate conceptual framework for psychometrics. Taking educational measurement as a case study, it explores

what that could look like in practice, drawing on fuzzy logic and mathematical order theory. It suggests that applying this approach to the assessment of intersubjectively constructed phenomena, such as a learner's proficiency in an inherently fuzzily-defined subject area, entails recognising the theory-dependent nature of valid representations of such phenomena, which need not be conceived of structurally as values of quantities. Finally, some connections are made between this 'qualitative mathematical' theorisation of educational assessment, and the application of techniques from machine learning and artificial intelligence in this area.

Keywords: Theory & philosophy of measurement, psychometrics, educational assessment, van Fraassen, qualitative mathematics, concept lattice, fuzzy logic

5.1 Introduction

The question of what it could mean to *measure* phenomena that form the basis of theory and debate in the human sciences, such as human attitudes, opinions, dispositions, or psychological or cognitive traits, has been a subject of critical enquiry since at least the mid eighteenth century (Michell, 1999a). For example, the question of whether such phenomena could be *quantified* was contested by Reid (1748 [1849]), even before a clearer definition of 'a quantity' had been put forward by Hölder in 1901.

This paper considers the question of measuring educational constructs, such as a learner's *ability*, or *proficiency*, or *competence* in a subject, field of study, or educational domain. Many educational tests and assessment procedures—some of them used to make high-stakes decisions about the test-takers—apparently produce, or claim to produce, numerical measurements of such properties, such that learners can be placed on a quantitative *scale* with respect to them. Psychometrics is the application of statistical methods to the

study of psychological and educational phenomena. It relies on the particular mathematical characteristics of quantitative structures (in practice, the real numbers and vector spaces over the reals) to perform calculations and procedures that are used as the warrants for substantive conclusions, such as 'how much' ability a student is estimated to have, or how to equate measurements of ability derived from different tests.

The paper argues that the reliance of psychometrics on quantitative structures is grounded in an assumption that *quantification* is necessary to allow *measurement*. It proposes, however, that psychological and educational measurement need not be reliant on numbers. It suggests that van Fraassen's (2008) account of measurement as a process whereby the measurand is located in an appropriate 'logical space' is well-suited to serve as a foundation for an account of the measurement of educational phenomena such as students' abilities or competencies in a subject domain—phenomena that are arguably inherently 'fuzzy' and multifaceted. Such a logical space *could* be the particular mathematical structure that uniquely characterises the real numbers (a complete ordered field, in mathematical terminology), but it need not be.

The structure of the paper is as follows. Section 5.2 briefly outlines the approach to measuring cognitive and educational constructs, by assuming quantitative structure, that became standard in psychometrics over the twentieth century. It summarises critiques of the quantity assumption, and argues that these critiques have sufficient conceptual and empirical weight to warrant a serious explanation of what an approach to psychological and educational measurement could look like if the assumption is set aside. Taking the example of summative educational assessment in particular, it suggests that in many cases construct validity may be better served by a more generalised view of measurement, of the kind proposed by van Fraassen (2008). Van Fraassen's approach is explained in more detail in section 5.3.

Section 5.4 makes the discussion more concrete by comparing quantitative and qualita-

tive measurement approaches for a toy example of an educational test. This is extended in section 5.5 to a consideration of the practicalities—in particular, the computational complexity—of applying qualitative mathematical (fuzzy order-theoretic) methods to the kinds of test response data that arise in real practice. And since traditional methods of analysis of educational assessment data are increasingly being supplemented, or even supplanted, by the application of techniques from natural language processing, machine learning, and artifical intelligence (AI), section 5.6 considers some of the connections between educational measurement and AI-enabled classification procedures. Finally, the concluding discussion in section 5.7 poses some questions for further research. It concludes that it is worth pursuing further conceptual and technical development of non-quantitative measurement approaches in psychometrics, especially since, with the rapid rise and application of AI (e.g. Davier, Mislevey, and Hao, 2021), there is a risk that psychometrics is simply replaced with data science—with the loss of substantive theoretical content concerning construct definition and the design of valid measurement procedures. A way forward is for psychometrics itself to develop into a discipline that rests on quantitative measurement when it is appropriate, but does not exclude a broader view.

5.2 Quantification in psychometrics

5.2.1 Abilities as latent quantities

Psychometrics normally conceptualises a learner's *ability* (or *proficiency*, or *competence*) in a domain as a latent numerical quantity, θ (Kline, 2000; Linden and Hambleton, 1997b). For each learner, a value of θ is calculated from the observed data arising from an assessment (e.g. item response data). The 'more θ' a learner has (the higher their value of θ), the 'better at' the assessment construct they are taken to be (modulo some

'measurement error'). That is to say, the relation of *betterness*, between learners, as to the different levels, states, or configurations of their abilities, is taken to be adequately captured by the relation of *order* ($\geq$) between numerical values. Moreover, to allow a value of θ actually to be derived for each learner, the set of all possible θ-values is normally supposed not only to be totally ordered, but quantitative and continuous[1]. Making these structural assumptions about the property of *ability* enables it to be treated as if it were a real number. Hence the whole array of statistical techniques whose mathematical validity depends on the metric and topological properties of the real numbers (such as factor analysis, item response theory, maximum likelihood estimation, etc.) can be applied to obtain numerical values that are taken to be *measurements of learners' abilities* in the cognitive or educational domain in question.

This paper will argue that one should not think of the 'betterness' relation between learners, as to their proficiency in a particular educational domain, as a total order relation (a ranking), in general, but rather as a partial order[2]. Sometimes the way in which the assessment construct is defined will allow learners to be ranked as to their proficiency with respect to that construct. In other cases, it may only be possible to infer, for some pairs of learners, that their proficiency states, or levels, are non-comparable (qualitatively different). This does not preclude the possibility of grouping learners together

[1] See Appendix C for definitions of *total order* and *quantity*. Informally, a totally ordered set X is one in which all the members can be ranked – there is an ordering $\geq$ such that either $x \geq y$ or $y \geq x$, for all x and y in X. A property is a quantity if its values are totally ordered and also additive – that is, they can be combined in a way that mirrors the properties of the addition of numbers. Additivity is required for a property's values to form an *interval scale* or a *ratio scale*, in the terminology of Stevens (1946b). A quantitative property is *continuous* if its possible values form a continuum with no 'gaps'.

[2] See Appendix C for a formal definition of *partial order*. In essence, when entities are partially ordered, there may exist pairs of entities that are not directly comparable, and the entities cannot necessarily be placed in a single linear sequence (a ranking) with respect to the feature of interest. In educational tests, each individual item (question or task) typically totally orders the respondents with respect to that item (for example 'those who got the question right' $\geq$ 'those who got the question wrong'; or 'those who scored 3 marks' $\geq$ 'those who scored 2 marks' $\geq$ 'those who scored 1 mark' $\geq$ 'those who scored 0 marks'). In general, however, the joint result (the product) of all of these total orders is an overall partial ordering of respondents, with some patterns of item responses not being directly comparable with others.

into 'coarser' ordinal classes (such as examination grades), such that one can infer that those who 'pass' are more proficient than those who 'fail', for instance. It just means that, within the 'pass' category, there may be some learners whose proficiencies, although both of at least a 'pass' level, may be different, and non-comparable. This argument is developed further in section 5.4 below.

There is a literature that critically examines the plausibility of assuming quantitative structure for phenomena such as ability (for example, Michell, 2006a; Michell, 2009a; Michell, 2012a; Michell, 2013; Heene, 2013b; Kyngdon, 2011; McGrane and Maul, 2020, and from a broader perspective, Uher, 2021; Uher, 2022a). One focus of this has been what Michell (2012a) calls the 'psychometricians' fallacy': the implicit leap that is often made, from maintaining that a property has a totally-ordered structure (that its possible values, states, or levels can be ranked, that is, placed on an *ordinal scale*, as described by Stevens, 1946b), to treating it as if it had quantitative structure (as if its values formed an *interval* or a *ratio* scale, in Stevens' typology).

In some cases it is possible to test empirically whether a property whose values are ordered is plausibly likely to have the further structure required for it to be quantitative. This is discussed in Section 5.2.2.2. Yet at an even more basic level, one might question why a construct such as *ability* with respect to a given cognitive or educational domain (specified in a more-or-less precise way), should even be regarded as a property that necessarily ought to have a totally ordered structure. Must it be a phenomenon that only occurs in such a way that any one person's ability-state is always linearly comparable with (larger than, the same as, or smaller than) any other person's state? Uher (2022b) makes an analogous point with respect to the use of rating scales to 'measure' the property of agreement.

If one considers the actual data upon which the inferences derived from educational testing procedures are based, then as Kane (2008, p.104) notes, 'we are likely to have, at

best, a partial ordering, unless we arbitrarily decide that some patterns [of item response] are better than others'. In practice, and as discussed further in Section 5.4, almost all psychometric approaches to working with such partially-ordered data do indeed involve making decisions about how to use the data to generate a total order (with each learner's score being their location with respect to this total order).

The question whether such decisions are indeed 'arbitrary' (and if not, which one is best or most appropriate) hinges, again, on how the measurand—each respondent's ability in the domain in question—is conceptualised. This issue is well described by Maul (2017a, p. 60), who notes that

> Any effort to construct a measure of an attribute will have trouble getting off the ground in the absence of a sufficiently well-formed definition of the target attribute, including an account of what it means for the attribute to vary (i.e., what meaning can be attached to claims about there being 'more' or 'less' of it, between and possibly within individuals) and how such variation is related to variation in the observed outcomes of the instrument (i.e. item response behaviour).

It is suggested in section 5.3.2 that questions of this kind form part of what van Fraassen (2008) refers to as the *data model* for the target attribute. It is rather rare for psychometrics textbooks to devote much attention to these theoretical or conceptual issues, however. Often (e.g., Raykov and Marcoulides, 2011) it is stated that psychological and educational measurement is concerned with appraising how individuals differ with regard to hypothesised, but not directly observable, attributes or traits, such as intelligence, anxiety, or extraversion. It is assumed that these traits are in fact quantities (for instance Kline (2000, p. 18) simply states that 'the vast majority of psychological tests measuring intelligence, ability, personality and motivation ... are interval scales'), and models are then introduced to relate them to observable data such as test or questionnaire responses

in such a way as to enable the numerical latent trait parameters to be estimated, together with measures of precision such as standard errors—all conditional on the adequacy and plausibility of the model that has been assumed. Of course if the model is not adequate as a structural theory of the phenomenon itself, then results may simply reflect artefacts of the model (e.g. consequences—sometimes rather trivial tautologies—that follow from the metric structure of the real numbers), rather than corresponding to valid inferences with respect to the theory of the phenomenon.

Why should a phenomenon such as a learner's proficiency or competence in a particular domain be assumed to have the structure of a total order (let alone a quantity)? The reason probably goes back to a belief fundamental to the early development of psychometrics, that quantitative structure is necessary to enable measurement. For example, Thurstone (1928) claimed that

> When the idea of measurement is applied to scholastic achievement, ... it is necessary to force the qualitative variations [in learners' performances] into a quantitative linear scale of some sort.

If 'the idea of measurement' entails *locating a measurand at a point on the real number line*, then 'forcing' observed qualitative variations to fit a quantitative structure is an understandable approach to adopt (even if it raises questions about validity). Indeed two common theoretical frameworks for psychological and educational measurement—the representational theory of measurement, and Rasch measurement theory—could be construed as concerned with ways to 'force' qualitative variation into quantitative form: the former by aiming to define conditions under which qualitative observations can be mapped into numerical structures; the latter by rejecting observations that do not fit an assumed quantitative model. These approaches are unpacked a little in the next section.

5.2.2 Theories of measurement

5.2.2.1 The representational theory of measurement

Tal (2020), in his survey of the philosophy of measurement in science, describes the representational theory of measurement (RTM) as 'the most influential mathematical theory of measurement to date'. Wolff (2020b), in a recent structuralist account of quantity and measurement, calls it 'arguably the most developed formal theory of measurement'. Michell (1990b) claimed that it is 'the orthodox theory of measurement within the philosophy of science'.

The canonical text on RTM (Krantz et al., 1971b, p. 9) takes *measurement* to mean 'the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful'.

RTM supposes that we are given an 'empirical relational structure' (itself an abstraction of certain features of an 'observed reality'). This structure consists of objects, relations between them, and possibly also ways of combining or composing them. For example in educational measurement contexts, we might take as objects students' responses to a writing task, and consider a binary relation $\succeq$ of *betterness* as being of interest (as in 'student X's piece of writing is a better response to the task than student Y's: $X \succeq Y$). Or we might be interested in how parts of a test or assessment combine (via a binary operation $\bullet$) to form an overall measure. For example, 'correctly answering questions 3 and 4 demonstrates a higher level of proficiency than correctly answering questions 1 and 2': $q_3 \bullet q_4 \succeq q_1 \bullet q_2$. We might then wish to investigate whether these aspects of students' responses to tasks—this empirical relational structure—can be mapped to a numerical ordering or scoring system, in such a way that the structure is preserved

(e.g. relative betterness between responses is mirrored by the relative magnitudes of the numbers assigned to those responses).

The idea is that if such homomorphisms can be shown to exist, then inferences in the numerical relational structure (normally taken to be the real numbers with the usual order relation $\geq$ and binary operations $+$ and $\cdot$) provide warrants for conclusions in the substantive domain of the empirical relational structure. If, further, we posit that differences in the observed outcomes of an educational assessment procedure, such as the administration of a test or examination, are *caused by* differences in the configurations, between learners, of their 'underlying proficiency', then establishing a homomorphism between the empirical relational structure and the real numbers (i.e. establishing that the outcomes can be 'placed on an interval (or ratio) scale') serves to justify the assumption of quantitative structure for this assumed underlying proficiency trait, and hence to enable the measurement of each test-taker's proficiency by locating them at the point on the real line that corresponds to their level of proficiency.

5.2.2.2 Qualitative relational structures and testing for quantity

The adequacy of RTM as a theory of measurement has been extensively critiqued (see, e.g., Michell, 1990b; Michell, 2021a; see also Luce and Narens, 1994), with commentaries noting that its abstract nature sidesteps the actual process of measuring anything, the construction of measuring instruments, and any discussion of measurement error. The merits of such critiques are not discussed further in this paper, because the position adopted here will be that of Heilmann (2015). Heilmann does not assess RTM as a candidate for a theory of measurement, but rather as a collection of mathematical theorems: theorems whose structure makes them useful for investigating problems of concept formation. He proposes (2015, p. 789) viewing theorems in RTM as

> providing us with mathematical structures which, if sustained by specific conceptual interpretations, can provide insights into the possibilities and limits of representing concepts numerically

He regards RTM as studying not mappings from an empirical relational structure to a numerical relational structure, but rather from a *qualitative relational structure* (QRS) to a numerical relational structure. Taken in that sense, he argues, RTM can provide tools for testing the extent to which abstract concepts (captured or described as qualitative relational structures) can be represented numerically[3].

Arguably, this is how RTM (including in particular the subset of RTM theorems that form the so-called theory of *conjoint measurement*: see Luce and Tukey, 1964a) does in fact tend to be used in the literature exploring the plausibility of assuming quantitative structure for educational, psychological, or social measurands.

For example, Michell (1990b) re-analysed data collected by Thurstone (1927c) regarding judgements as to the seriousness of various crimes. Thurstone claimed that his theory of *comparative judgement* (Thurstone, 1927a) enabled the construction of a *quantitative scale* for the measurement of seriousness of crime, by applying the theory to the outcomes of a collection of pairwise comparisons, in which subjects were repeatedly asked which of two crimes presented to them was the more serious. Michell carefully stated the assumptions of Thurstone's theory, and demonstrated by applying results from RTM that (1990, p.107) 'either seriousness of crimes is not a quantitative variable, or else some other part of Thurstone's theory of comparative judgement is false'.

Rooij (2011) applied theorems from RTM to explore whether properties of objects, that

[3] A further extension of Heilmann's position would be to consider mappings from a QRS to another QRS: in other words, to relax the restriction that the 'representing' structure should be numerical. Such a generalisation might permit both RTM and van Fraassen's approach to be located, from a formal mathematical perspective, within the general theory of structure known as category theory, but will not be pursued here.

manifest linguistically as adjectives with comparative degrees, can be represented numerically, what scale properties may hold for them, and hence whether inter-adjective comparisons (such as 'x is P-er than y is Q') can be meaningful. This is analogous to the vexed question, in educational assessment, of inter-subject comparison when it comes to setting and maintaining qualification standards (see, e.g., Newton et al., 2007; Coe, 2008).

Karabatsos (2001a) and Karabatsos (2018), Kyngdon (2011), Domingue (2014a), and Scharaschkin (2023) applied theorems from RTM to the question of testing whether psychometric attributes comply with requirements for quantitative structure, combining the RTM results with a stochastic approach to address expected 'measurement error' in most measurement scenarios with reasonable numbers of test-takers and test items. Domingue found that the results of a well-known test of reading showed that it was highly implausible that reading proficiency was a quantitatively-structured variable. Scharaschkin found that the results of a test of physics for school-leavers did not support the assumption of quantitative structure for a hypothesised 'physics proficiency' construct. On the other hand, he found that the results of a similar test of economics were approximately consistent with an assumption of quantitative structure.

None of these applications require assuming the validity or adequacy of RTM as a substantive theory of measurement—indeed, Michell (2021a) explicitly rejects it. Yet they do shed light on the extent to which qualitatively-structured data can be treated *as if* it were a manifestation of quantitatively-structured latent traits, and provide empirical evidence that it is not always valid to do so.

This is relevant to the practice of educational assessment and test construction because most practitioners and test developers probably do work within a pragmatic 'as if' framework, as summarised by Lord and Novick (1968, p. 358):

> Much of psychological theory is based on trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behaviour.

Some of the ways in which theories of cognition have been more directly incorporated into the use of quantitative latent variable modelling, and their relation to the ideas considered in this paper, are discussed further in section 5.5.4.

5.2.2.3 Rasch measurement theory

Psychometrics conducted in the Rasch measurement tradition (Andrich and Marais, 2019a) takes the view that measurement is only meaningful for quantitative phenomena. Thus if a putative measurement procedure such as an educational or psychological test yields results that are inconsistent with a underlying quantitative variable, then the procedure is not, in fact, *bona fide* measurement, and requires modification. In practice this means modifying tests by deleting or changing items until a sufficiently good fit to the Rasch model is obtained[4].

So rather than trying to find a model that fits the data that has been obtained from the administration of a test, the Rasch measurement approach is to try to make the data fit the model. Modifying the measurement instrument to achieve this may come at the cost of severely constraining the theory of (or, in the terminology of section 5.3.2, the relevant

[4]The Rasch model, also known as the 1-parameter item response model, postulates that the log-odds of a test-taker of ability θ correctly answering an item of difficulty δ is simply $\theta - \delta$ (in the case of a test consisting of a sequence of dichotomously-scored items). There are of course other item response models that postulate additional item parameters, but Rasch theorists hold that the 1-parameter model is theoretically more appropriate as a basis for enabling measurement because it enables, within a given collection of persons and items, so-called invariant comparisons of persons (as to their ability) and items (as to their difficulty): see Andrich and Marais (2019a, p. 80).

data model for) the substantive phenomenon or construct of interest. It might be that the construct cannot be sufficiently constrained or re-defined without significantly departing from its underpinning theory of value. In an educational assessment context, this would be the case if making such changes to the assessment instrument would compromise construct validity: the assessors' understanding of what constitute the key attributes of proficiency in the given domain, and how relatively better/worse/different states of proficiency would present with respect to these attributes. In such cases the choice would seem to be either to abandon the idea of measuring the construct at all, or to abandon the restriction of measurement to locating measurands within solely quantitative mathematical structures. This paper explores the latter option.

5.2.2.4 Measurements as ratios

Michell (1999a) traces the evolution of the concept of measurement in psychology since the publication of Fechner's *Elemente der Psychophysik* in 1860. He bemoans the movement away from the conceptualisation of measurement that had become standard in nineteenth century physics, namely (Michell, 1999, p.14) 'the discovery[5] or estimation of the ratio of the magnitude of a quantitative attribute to a unit (a unit being, in principle, any magnitude of the same quantitative attribute)'. In other words, as elementary physics texts still state, physical quantity = real number × unit, where the real number is the measurement of the physical quantity.

Michell notes (p.19) that 'according to the traditional understanding of measurement, only attributes which possess quantitative structure are measurable. This is because

[5]The development of quantum theory in the twentieth century problematised the classical epistemological viewpoint on measurement as 'discovery'. As Peres (1995, p. 14) observes, 'classical physics assumes that the property which is measured objectively *exists* prior to the interaction of the measuring apparatus [in educational measurement, the administered task or test together with its valuation procedure] with the observed system [the learners' performances]. Quantum physics, on the other hand, is incompatible with the proposition that measurements discover some unknown but pre-existing reality.'

only quantitative structure sustains ratios'. He argues that, this being the case, it is incumbent on psychometricians to investigate whether the phenomena they study do, in fact, have quantitative structure, before applying statistical models that assume it. Since in practice this is almost never done, his claim is that, for the most part, 'psychometrics is built upon a myth' (Michell, 2012a). Once again, the choice appears to be to accept the constraints of the 'traditional understanding of measurement', or to explore whether psychometrics could benefit from engagement with a more expansive conceptualisation of what it means to measure something. The next section considers such a viewpoint.

5.3 van Fraassen's account of measurement

5.3.1 Basic principles and relevance to psychometrics

Bas van Fraassen's *Scientific Representation: Paradoxes of Perspective* (2008) is an empiricist structuralist account of measurement and representation in science. This stance eschews debate about the ontological status of the phenomena or reality that scientifc theories describe, and concerns itself rather with elucidation of what van Fraassen argues is the key aim of developing and testing such theories, namely their empirical adequacy. He claims (2008, p.2) that 'measuring, just as well as theorising, is representing ... measuring *locates* the target in a theoretically constructed logical space'. To be more precise (p. 164),

> measurement is an operation that locates an item (already classified in the domain of a given theory) in a logical space (provided by the theory to represent a range of possible states or characteristics of such items).

A key point here is the theory-relatedness of measurement procedures. Echoing Maul's (2017) requirements, quoted in section 5.2.1, for a 'well-formed definition of the target

attribute' as fundamental to psychometric measurement, van Fraassen suggests (p. 166) that 'once a stable theory has been achieved, the distinction between what is and is not genuine measurement will be answered *relative to that theory*'.

It is argued in section 5.4 that a candidate theory for the phenomena (proficiency or competence in a domain) that form the subject matter of educational measurement, is a description of what consitutes betterness between learners' possible states or configurations of proficiency in a given domain. 'Betterness'—which, as noted in section 5.2, may be a more general order relation than a simple ranking—has to be defined in terms of criteria that may, in general, be manifested with *fuzzy degrees of truth* in the responses of learners to tasks that have been designed to provide information about their proficiency in the domain in question.

van Fraassen considers several measuring procedures in classical and quantum physics (pp. 157-172 and 312-316), and concludes (p. 172) that they are all 'cases of grading, in a generalised sense: they serve to classify items as in a certain respect greater, less, or equal. But ... this does not establish that the scale must be the real number continuum, nor even that the order is linear. The range may be an algebra, a lattice, or even more rudimentary, a poset'. In fact, section 5.4 below considers the case of lattices as logical spaces for educational measurement procedures[6].

It is worth exploring how van Fraassen's approach could be applied to educational measurement for at least two reasons. Firstly because, as discussed in section 5.2.2.2, the mathematically necessary conditions for a learner's proficiency in a given educational domain to have the structure of a quantity often do not hold; and it is not possible to massage the assessment instrument to make them hold without loss of construct validity.

[6] Algebras, lattices, and posets (short for partially-ordered sets) are types of mathematical structures. In particular, a lattice is a partially-ordered set (see Appendix C for a definition) in which each pair of elements has a least upper bound and a greatest lower bound.

In such cases, it would arguably be inappropriate to theorise the construct as quantitative, and hence its measurement as location *on the real line*, rather than in some other, theory-relevant, logical space.

Secondly, the approach of thinking about educational assessment constructs in terms of fuzzy criteria of value (what will count as creditworthy, or indicative of good/bad performance, in relation to what particular domain content) is what *actually happens in practice*, when subject domain experts develop and administer at least one kind of high-volume, high-stakes, educational assessment procedure, namely the public examinations taken by school pupils aged 16 and 18 in the UK. This brings us to a consideration of what van Fraassen calls *data models*.

5.3.2 Data and surface models

Measurements arise from the results of procedures designed to gather information about a phenomenon of interest. As noted in section 5.2.2.2, these entail selective attention to specific features that are deemed to be relevant. That is to say, measuring a phenomenon involves collecting data structured in a specific way. Van Fraassen calls such a structure a *data model* for the measurand in question. He notes (2008, p.253) that

> A data model is relevant for a given phenomenon, not because of any abstract structural features of the model, but because it was constructed on the basis of results gathered in a certain way, selected by specific criteria of relevance, on certain occasions, in a practical experimental or observational setting designed for that purpose.

In educational measurement we have gathered in a certain way (via an assessment procedure such as a test), selected by specific criteria of relevance (construct-relevant criteria: Pollitt and Ahmed, 2008) on certain occasions (at a particular point or points in time),

in a practical setting designed for that purpose (e.g. the rules of administration and physical requirements for conducting an examination).

In the case where the test consists of a sequence of dichotomously-scored items $I := \{i_1, ..., i_n\}$ administered to a collection $L := \{l_1, ..., l_m\}$ of learners, we can think of this measurement setup as a map $V : L \times I \to \{0, 1\}$ that assigns to each instance of a learner encountering an item the valuation 1 if they answer it correctly, and 0 if they answer it incorrectly. Equivalently, we can think of the information collected by the assessment procedure as organised in an $m \times n$ matrix whose (m, n) entry is $V(l_m, i_n)$. There is, however, more structure entailed by the 'betterness' ordering within each item (namely that '1' is better than '0') than immediately stands out from simply viewing the data as a table. As discussed in section 5.4.2, the totality of the results-plus-valuation-system can be viewed as a lattice (the so-called *concept lattice* for the data table)—and it is suggested in section 5.4 that such lattices (generalised to incorporate fuzzy valuations if necessary) form the natural data model for the phenomena that educational measurement procedures, such as tests and examinations, aim to measure.

Van Fraassen describes (2008, p.253) constructing a data model as 'precisely the selective relevant depiction of the phenomena *by the user of the theory* required for the possibility of representation of the phenomenon.' In the context of educational tesing, the proficiencies being studied are proficiencies or competencies *with respect to* a specified domain (such as 'high school chemistry', or 'A level French'). What 'good performance' or 'good demonstrated attainment' looks like in these domains (and hence what would count as evidence of better or worse levels, or states, or configurations, of learners' *proficiencies*) is always subject to a prevailing understanding or agreement as to what potential aspects of the domain are chosen as relevant for discrimination between learners' performances as to their quality. In other words, the criteria for creditworthiness of candidates' responses to tasks in an assessment can be regarded as the selective relevant depiction of

the phenomenon of interest, by those members of the competent authority (the 'users of the theory') who design, administer, and grade the tests. For that reason, concept lattices derived from the outcome data from the tests, that encode the relationship between learners and the assessment criteria, are appropriate data models.

In practice, van Fraassen (2008, p.167) notes that data models may be 'abstracted into a mathematically idealised form' before empirical or experimental results are used to explore theories or explanations, or for substantive purposes. He gives the example of a data model consisting of relative frequencies, which is 'smoothed' such that frequency counts are replaced with probabilities. An idealised or simplified version of a data model is called a *surface model* for the phenomenon in question. Surface models are considered further in section 5.5.

5.4 Theories of constructs: comparing item response theory and fuzzy concept analysis

5.4.1 A small example

Table 5.1 shows results from an assessment that generates data on each of three items (or attributes) $\{i_1, i_2, i_3\}$ for six learners $\{l_1, ..., l_6\}$. Here 0 means 'not demonstrated', $\frac{1}{2}$ means 'partially demonstrated', and 1 (or $\frac{2}{2}$) means 'fully demonstrated'.

A traditional psychometric approach to analysing this kind of data would be to treat each learner's results from the assessment as a vector in $\mathbb{R}^3$, and each learner's proficiency measure as a quantity (a point in $\mathbb{R}$). For example, we could treat the label for each item response category as a number, and add them to get a total score for each learner. This orders learners, with respect to proficiency, equivalently to fitting a Rasch

Table 5.1: Example: data from a test

$\setminus$	i_1	i_2	i_3
l_1	0	$\frac{1}{2}$	$\frac{1}{2}$
l_2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
l_3	1	1	$\frac{1}{2}$
l_4	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
l_5	0	$\frac{1}{2}$	0
l_6	$\frac{1}{2}$	1	1

model (a 1-parameter item-response model), since total score is a sufficient statistic for estimating proficiency in this model. Or we could do a principal components analysis and take the projection of each learner's item-response vector onto the component that accounts for the most variance as their proficiency measure (this is equivalent to fitting a 2-parameter item-response model: see Cho, 2023). Doing so for the data in Table 5.1 yields three components of which the first accounts for 72% of the variance in outcomes, with the other two accounting for 19% and 9% respectively. We could therefore take the loading (projection) of each learner's results onto the first component as their score on an 'underlying' quantitative variable that represents the assessment construct reasonably well. Figure 5.1 shows how learners' proficiency measures differ depending on the approach taken.

However, in view of the problems associated with assuming quantitative structure for proficiency discussed in section 5.2.1 (tantamount, in section 5.3.2's terms, to replacing the data model with a radically different surface model), let us consider a non-quantitative approach. If we take each learner's test response not as a vector of numbers, but rather a vector of ordered labels, then the observed data can be characterised as a collection of partially-ordered nodes: a network of 'betterness' relations between nodes. In this data

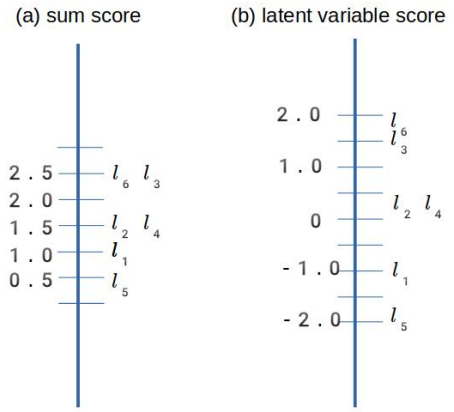


Figure 5.1: IRT-derived proficiency measures

model, shown in Figure 5.2, each node is a *type of performance* on the assessment.

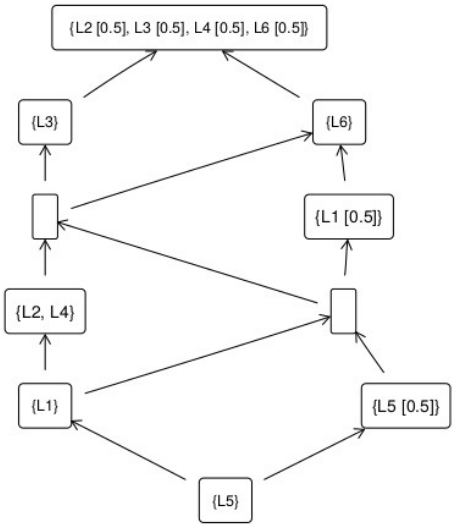


Figure 5.2: Fuzzy concept lattice for assessment data

Each type of performance is defined by a collection of *attributes*, that *characterise* it; or (dually) by a collection of *learners*, who *demonstrate* it. The boxes in Figure 5.2 are the different types of performances on the test. The best performance is at the top of the

diagram, and the worst performance at the bottom Attributes, and learners, may belong to nodes to a *fuzzy degree*. Thus learner 5 belongs to (demonstrates) the lowest type of performance completely (to degree 1). Learners 2, 3, 4, and 6 all demonstrate the highest type of performance to degree 0.5.

Better types of performance are characterised by showing *more* attributes (and, dually, are demonstrated by *fewer* learners) than worse types of performance. An arrow from a box A to a box B means that B is a better performance than A (and by extension better than any performance C such that there is a connected path from C to A). If there is no path between two types of performance, then they are not comparable. Locating a learner (measuring their proficiency), with respect to this data model for the construct which the three-item test aims to assess, then means finding the 'highest' node that they belong to in the network. This intuitive description is made more precise in the following section.

5.4.2 Formal concept analysis and proficiency measurement

Formal concept analysis (Ganter and Wille, 1999b; Carpineto and Romano, 2004) is an important development of mathematical order theory that has been applied extensively to fields such as linguistics, political science, information sciences, medicine, and genetics. A recent application (Bradley, Gastaldi, and Terilla, 2024) is to elucidating the mathematical representation of structure in large language models such as ChatGPT, discussed briefly below in section 5.6. It can be thought of as a way of making explicit the information structure that is implicit in a matrix—such as that in Table 5.1—which relates objects to attributes (or learners to test items). It provides methods to extract the concepts and implications that can be deduced from such data, and introduces a logic to reason and infer new knowledge.

Consider first the case of measuring proficiency in a domain by administering an n-item test to m learners, where each item is dichotomously scored, i.e., for each learner l and item i, it is either the case that l answered i correctly, or that l did not answer i correctly. Given a subset of learners $L_1 := \{l_1, ..., l_k\}$, let $I_1 := \{i_1, ..., i_j\}$ be precisely those items that all learners in L_1 got correct. Then the pair (L_1, I_1) is an instance of a *formal concept* present in the data. L_1 is called the *extent* of the concept, and I_1 is called its *intent*. We can equally well start with a subset $I_2 := \{i_1, ..., i_p\}$ of items, and then form the concept (L_2, I_2), where L_2 is precisely the set of learners who got all items in I_2 correct.

The collection of all formal concepts extracted from a matrix or data table simply restates the information present by virtue of the way the data is structured due to the choice of attributes (test item responses, in this example), and the ordered valuations chosen for attributes (just the two categories $1 \geq 0$ in this case). However, it makes this structure more apparent (and graphically representable, as in Figure 5.1) because concepts are (partially) *ordered* via the set-theoretic notion of inclusion. A concept (L_1, I_1) is *more general* than a concept (L_2, I_2) if $L_1 \supseteq L_2$ (or equivalently, if $I_1 \subseteq I_2$). The most general concept is the one that has the largest extent (and smallest intent). In test performance terms, the most general concept corresponds to the bottom, or worst, performance: because every other performance has a larger intent (entails more correct items). Similarly, the least general concept (with the smallest extent and largest intent) corresponds to the top, or best, level of performance[7].

We can think of formal concepts as different ways of performing on the test (i.e. different

[7] Normally concept lattices are drawn as so-called Hasse diagrams with the least general concept at the bottom, and the most general concept at the top. An arrow is drawn upwards from concept A to concept B if B is more general than A. In the educational assessment context, we naturally regard the best performance as the *top* concept, which means we need to reverse the usual ordering (in mathematical terms, we use the *dual* lattice). This is done throughout this paper, for example in Figure 5.2, where the worst level of proficiency (exhibited, to degree 0.5, by learner l_5) is at the bottom of the diagram, and the best level (exhibited by learners l_2, l_3, l_4, and l_6, also to degree 0.5) is at the top.

ways of exhibiting proficiency in the subject domain). Each type of performance—or exhibition of proficiency—can be described *extensively*, by showing the learners who demonstrated it. Or it can be described *intensively*, by showing the item-profiles that characterised it. These two modes of presentation correspond to different ways of training 'measuring instruments' (traditionally, human judges; more recently machine-learning methods such as neural nets) to recognise what good/bad performance (high/low proficiency) looks like. One can either give *examples* of a certain kind of performance, until an assessor can correctly classify new instances, or one can give *descriptions* of that kind of performance (in this case, the relevant profile of item responses), to enable new instances to be classified (measured) correctly[8].

For a small educational measurement procedure of this kind (small in terms of the number of items/tasks/relevant attributes on which data is collected, as well as small in terms of the number of subjecs to which it is administered), the qualitative equivalent of a quantitative score is a learner's location in the concept lattice: the highest concept, in the partial order, to whose extent they belong. This level of proficiency is described, not as a numerical 'amount' (location on a line), but rather by the intent of the relevant concept: the actual items they mastered (or, more generally, the construct-relevant attributes their performance demonstrated). For larger (more realistically sized) assessments, the concept-lattice data-model becomes too granular, as shown in section 5.5, and we develop a notion of 'prototypical' kinds of performances at a manageable number of levels, such that each learner's level, or state, of proficiearency can be described approximately in terms of its qualitatively closest prototype.

Before moving on to that discussion, it is necessary to consider the question of the fuzziness of the criteria that structure data models in many educational measurement procedures.

[8] As Weyl (1952, p. 8) noted, 'For measurement the distinction is essential between the "giving" of an object through individual exhibition on the one side, in conceptual ways on the other'.

5.4.3 Truth degrees and fuzzy concepts

5.4.3.1 Assessment results as truth degrees

Table 5.1 illustrates a situation that often obtains in educational assessment. Learners are given tasks, such as questions on a test, and they may be successful in engaging with them *to a certain degree*. The outcome of a learner's interaction with an item is not necessarily captured by the crisp dichotomy of {correct, incorrect}.

The usual way of dealing with this in psychometric models is to model response categories for polytomous items as a sequence of threshold points on a latent quantitative continuum. A learner's response is in a higher category if it results from their proficiency-state being higher than, but not otherwise different from, a learner whose response is in a lower category. Differences in proficiency must be conceived of as differences in degree, not in kind. Yet as Michell (2012a, p. 265) notes, in the context of mathematics tests, 'the differences between cognitive resources needed to solve easy and moderately difficult items will not be the same as the differences between reources needed to solve moderately difficult and very difficult mathematics items. This observation suggests that abilities are composed of ordered hierarchies of cognitive resources, the differences between which are heterogeneous.'

An alternative approach is to start by the viewing the dichotomous situation as providing information about learners' performances in the form of *propositions* of the form 'learner l answered item i correctly'[9]. This proposition is true just in case the (l, i) entry in the data table arising from the assessment is 1. So we can think of the entries in the table as truth values (with 0 meaning false and 1 meanining true).

[9] As Michell (2009a) observes, 'Tabulated numbers are shorthand for a set of propositions that tell where the numbers came from. Furthermore, deductions from a data set are inferences from these propositions.'

It has long been recognised that, in situations in which there is inherent fuzziness, vagueness, or semantic uncertainty in concepts, bivalent logics, in which the only possible truth values for a proposition are {`false`, `true`} can be unduly restrictive (see e.g. Goguen, 1969b; Goertz, 2006b; Bělohlávek, Dauben, and Klir, 2017). *Fuzzy logic* (Hajek, 1998b; Bělohlávek, Dauben, and Klir, 2017) allows propositions to have truth values drawn from ordered sets of *truth degrees*, that can be more extensive than {`false`, `true`}.

Thus we can view the example in Table 5.1 as providing information about propositions with three truth-degrees, that we could label $\{0, \frac{1}{2}, 1\}$, or {`false`, `partially-true`, `true`}. For example, it is false that learner l_1 demonstrated attribute i_1 (or we could say, she demonstrated it to degree 0), and it is partially-true that she demonstrated attribute i_2 (she demonstrated it to degree $\frac{1}{2}$).

When the outcomes of educational measurement procedures are not completely and crisply dichotomous with respect to all the construct-relevant attributes about which information is collected, the concept lattice for the resulting matrix of fuzzy truth values is itself fuzzy. Objects and attributes belong to concepts with degrees of truth, rather than crisply. In the concept lattice in Figure 5.2, the label '0.5' after a learner-identifier means that learner belongs to the concept (i.e. has demonstrated that type or level of performance) to degree $\frac{1}{2}$).

Although a discussion of the concept of 'measurement error' in psychological testing and educational assessment would take us beyond the scope of this paper, it may be worth clarifying, for the avoidance of doubt, that the application of fuzzy logic in this context is not simply an alternative to using probability theory. Probability is a tool that can be used to study (epistemic) *uncertainty* (the lack of precision that arises from incomplete or poor information), whereas fuzzy logic is a tool that can be used to study (ontological) *vagueness* (the inherent fuzziness, or necessary inexactness, of concepts like 'proficiency' in a certain domain). Erwin Schrödinger, when considering what the

development of quantum mechanics meant for the measurement of physical phenomena, distinguished these two facets when he noted (Trimmer, 1980, p.328) that 'There is a difference between a shaky or out-of-focus photograph and a snapshot of clouds and fog banks'.

The statement 'Mary has a fairly good understanding of physics' is vague but certain, whereas 'Mary will pass the physics test tomorrow' is precise but uncertain. Working with propositions such as the former (i.e. deploying what Goguen (1969b) calls a 'logic of inexact concepts') is core to educational assessment, because of the contestable and intersubjective nature of educational constructs, discussed further in section 5.7.2.

5.4.3.2 Truth degrees and quantities

Buntins, Buntins, and Eggert (2016a) apply fuzzy logic to psychological tests in a somewhat different way to that proposed here. They take the view that scores obtained from a test should not 'refer to latent variables but to the truth value of the expression "person j has construct i", where a *construct* is defined by a collection of relevant *attributes*, each of which may be *possessed* by a test-taker to a certain degree, and each of which may be *relevant* for the construct to a certain degree. Modelling truth degrees as real-valued quantities in the interval [0,1], they present an algorithm for aggregating them across attributes to arrive at an overall score for each learner: the truth value of the proposition 'this learner has the construct'. They are careful to distinguish the semantic vagueness of a construct definition (recognised in the use of fuzzy truth values) from the idea of 'measurement error'.

Buntins et al. claim that this approach 'neither relies on latent variables nor on the concept of [quantitative] measurement'. However they do state it is arguable that 'although there is no measurement theory involved in the ... formalism, the application to

actual test behavior does presume item answers to be assessed on an interval scale level', because 'test answers have to be real numbers between 0 and 1, reflecting the subjective truth-values of the corresponding attributes for the tested person ... However, these only refer to the item level and do not extend to theories about latent variables.'

In fact truth degrees do *not* have to be real numbers between 0 and 1. What is required is that they have a way of being compared with each other—that is, an order structure (which could be a partial order)—and way of being combined with each other. In general these requirements are met by taking them to have the mathematical structure of a so-called complete residuated lattice (Hajek, 1998b). Further work on conceptualising truth degrees—and especially what that means for empirically eliciting them—is important, as touched on in section 5.7, but beyond the scope of this paper.

Buntins et al. see their approach 'not as opposed to psychometric theory but tr[ying] to complement it with an alternative way to conceptualize psychological tests'. By contrast, the approach presented in this paper is suggested not as an alternative to, but an extension of, psychometric theory: one in which quantitative measurement forms an important, but special, case of a more general measurement framework.

5.4.3.3 Fuzzy relational systems

In summary, the argument in this section is that in general, educational assessment procedures that aim to measure constructs such as proficiency, ability, or competence in a fuzzily-defined domain, generate *fuzzy relational systems*: matrices of truth-values for propositions of the form 'learner l has demonstrated construct-relevant attribute i'. As data models, these are equivalent to fuzzy concept lattices: partially-ordered hierarchies, or networks, of types of performance on the assessment, that are discriminable with

respect to these construct-relevant attributes. The next section considers whether these data models can provide insight for realistically-sized assessments.

5.5 Practicalities of educational assessment with non-quantitative data models

5.5.1 Granularity of data models

An issue with data models of the kind discussed in the previous section is that their combinatorial complexity increases geometrically with the numbers of learners and construct-relevant attributes of performance (or test items) involved. Figures 5.3 and 5.4, for instance, show the concept lattices for subsets of outcomes of a physics test[10] with increasing numbers of learners and attributes. Clearly the information here is too granular to be useful, and we need to simplify or 'smooth' it in some way.

For quantitative data models, where learners' test responses as thought of as vectors in n-dimensional Euclidean space, the analogous granuarity-reduction is often performed using latent variable models that aim to find a k-dimensional subspace with $k < n$ (often a one-dimensional subspace, i.e. a line) that is oriented in such a way as most closely to approximate the direction of most of the variation between the positions of these points (possibly subject to some other constraints as well, for certain factor-analytic models: see Bartholemew et al., 2008). Each learner's latent-variable score is then the projection of the vector that represents their test performance onto this subspace. Calculating these scores entails factorising the (transpose of the) matrix Z of normalised test scores. If there are m learners and n test items, then the $n \times m$ item-by-learner matrix Z^T is

[10] part of paper 1 of the AQA A level physics examination taken in 2018. Unusually for an A level assessment, the items here are all dichotomous (multiple-choice questions). The lattices would be even larger if the items admitted fuzzy valuations.

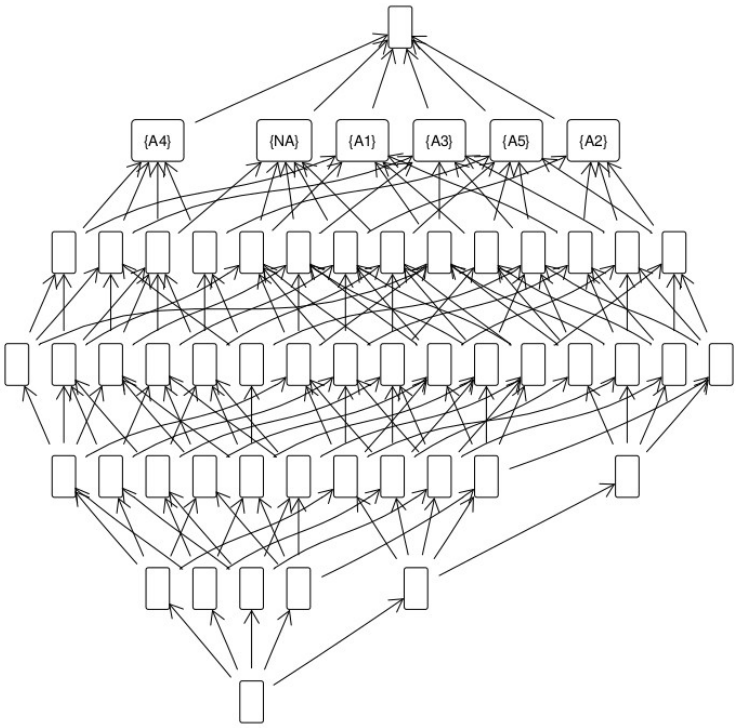


Figure 5.3: Concept lattice for a 5-item test with 100 leaners

factorised into the product of a $n \times k$ item-by-factor matrix L and a $k \times m$ factor-by-student matrix F, plus some error: $Z^T \approx LF$. Then using standard results in linear algebra, it can be shown (e.g. Reyment and Jöreskog, 1993) that the factors are the eigenvectors of the covariance matrix ZZ^T.

5.5.2 Factorising qualitative matrices

Bělohlávek (2012a) studied the question of factorising a matrix of fuzzy truth values. Now the matrix product is no longer defined in terms of operations on quantities, but rather in terms of operations on truth values.[11] Let M be an $m \times n$ matrix arising from

[11] The product of two real-valued matrices A and B is defined by setting its (i, j) entry $(AB)_{ij}$ to the inner product of row i of A with column j of B: i.e. $(AB)_{ij} := \sum_{p=1}^{k} A_{ip} B_{pk}$. When the matrix entries are truth values, they are elements of a type of lattice that is equipped with an operation $\otimes$

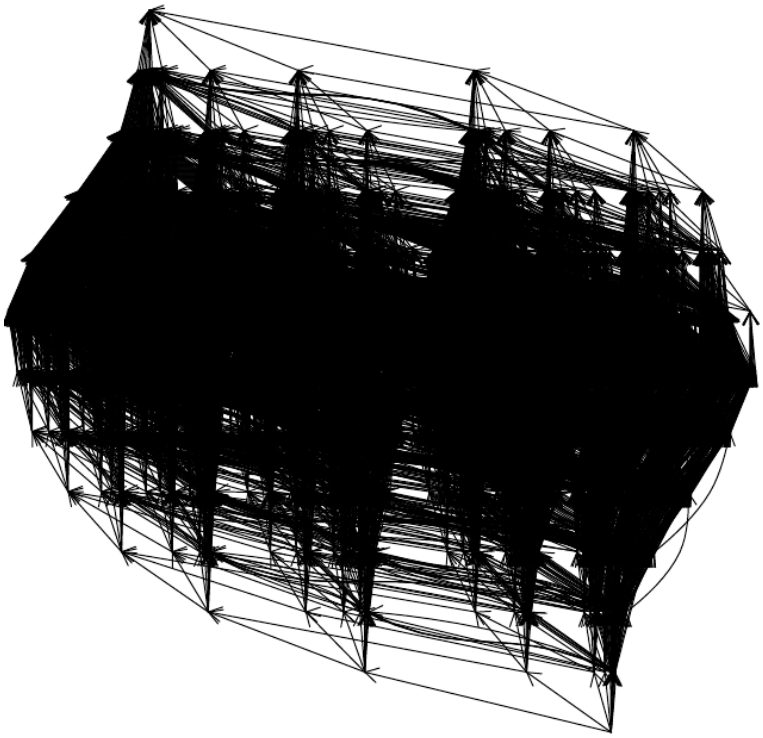


Figure 5.4: Concept lattice for a 12 item test with 200 learners

an educational measurement procedure conceptualised as in section 5.4.3, so that M_{ij} is the degree to which learner i displays attribute j. By analogy with the quantitative case, consider an approximate factorisation of M into a $m \times k$ learner-by-factor matrix A and a $k \times n$ factor-by-attribute matrix B, i.e., $M \approx A \circ B$. The key theorem in this case, due Bělohlávek (2012a), is that *the factors are particular formal concepts* from the concept lattice for M. That is, 'picking out key concepts' (particular types of learners' responses to the assessment) is equivalent to 'logically factorising' the matrix of truth-degrees that is the outcome of the measurement procedure.

The factors are the (extents and intents) of specific concepts in the concept lattice for M. The intuition is that, with $M_{ij} = A_{ip} \circ B_{pj}$:

to combine values. In this case the matrix product $A \circ B$ is defined as $(A \circ B)_{ij} := \bigvee_{p=1}^{k} A_{ip} \otimes B_{pk}$, where $\bigvee$ is the supremum over the indicated set (see Appendix C).

- A_{ip} is the degree to which learner i is an example of (in the extent of) factor p;
- B_{pj} is the degree to which attribute j is one of the manifestations of (in the intent of) factor p;
- $M = A \circ B$ means: learner i displays attribute j if and only if there is a factor (formal concept) p such that i is an example of p (or p applies to i); and j is one of the particular manifestations of p.

Thus, the qualitative analogue of projecting a Euclidean space onto a lower-dimensional subspace consists in picking out certain points in a partially ordered set. Specific formal concepts are selected, similarly to the way in which specific vectors—the eigenvectors of the covariance matrix—are selected when learners are scored on quantitative latent variables. The analogues of scores on a latent variable are the degrees to which learners' performances 'display' or 'participate in' or 'reflect' these specific concepts, which may be thought of as *prototype* or *standards of performance* on the construct. They have the advantage, over hypthesised latent variables whose values are abstracted from observed data, that they are directly expressible in terms of the construct-relevant attributes—that is, in terms of the features of learner's responses to assessment tasks that are taken to be important in a 'theory' of 'what (good) performance means', for the educational construct in question. They can be described both by means of their extent (the collection of actual learners' performances exemplifying the concept/standard in question), and by means of their intent (the collection of [fuzzy] attributes that characterises the standard in question).

5.5.3 Measures and meanings: comparing quantitative and qualitative approaches

Bartl, Bělohlávek, and Scharaschkin (2018) examined this qualitative factor analytic approach to educational assessment data, with the aims of exploring its applicability in practice, and its application to the study of the construct validity of an examination: the degree to which students' responses, assessed as being at a particular level, matched the intentions of the assessment designers in terms of the qualitative performance standard intended to broadly characterise responses at that level. This is the kind of question that is difficult to study using traditional quantitative methods.

The technical issues involved (for example how to determine the coverage and number of factors that broadly explain the data—analogous to a scree plot in quantitative principal components analysis) will not be rehearsed here. See Bartl, Bělohlávek, and Scharaschkin (2018) for computational details. For a deeper theoretical treatment of the relationship between eigenvectors (of quantitative covariance matrices) and formal concepts (of qualitative matrices of truth values), see Bradley (2020). The key point is that this approach allows drawing out key features associated with responses assigned to a particular level, by the assessment procedure, and an appraisal of the degree to which each learner's performance on the examination embodies or matches those features. Indeed, it 'explained' the data (in terms of proportion of data covered or variance explained) as well as standard principal components analysis, but generated factors exemplifying attributes of performance that seemed to be more easily interpretable.

Figure 5.5 shows an example of this, for the educational measurement data studied by Bartl, Bělohlávek, and Scharaschkin (2018), in which learners were assessed on 14 fuzzy attributes $\{y_1, ..., y_{14}\}$, each of which reflected an aspect of the construct, in this case proficiency in the specific subject of 'A level Government and Politics'. Each of the

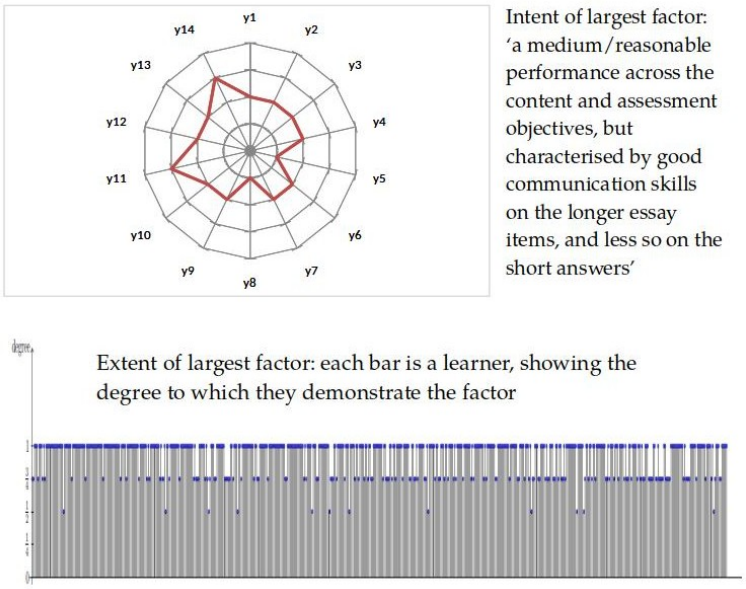


Figure 5.5: Factor representation for fuzzy data

attributes corresponds to demonstrating specific types of knowledge and understanding, in accordance with the examiners' agreed understanding of what better/worse proficiency means in this domain. Hence the intent of any given concept can be interpreted by users of the assessment as a description of broadly what that level of proficiency means (and likewise the extent of the concept can be interpreted as an indication of the degree to which each learner has demonstrated that level of proficiency).

The question of the interpretability or explainability of the results of educational measurement procedures—whether those results are numerical scores, or broader grades or levels—is particularly important for high-stakes assessments such as those that underwrite school-leaving qualifications. For learners, clarity about *why* their response to an assessment merited their being characterised as demonstrating a certain level of proficiency is arguably required for reasons of natural justice. For teachers, understanding qualitatively what their students did well, and what they would have to do better to demonstrate more proficiency in a subject domain, is clearly valuable as an input into

their future pedagogical practice. Bartl, Bělohlávek, and Scharaschkin (2018, p. 204) concluded that their approach to qualitative factor analysis yielded 'naturally intepretable factors from data which are easy to understand', but that more research is needed both on technical implementation and on the views of learners and teachers.

5.5.4 Other order-theoretic approaches to educational assessment

In the 1940s Louis Guttman began to develop an approach to psychological measurement (e.g. Guttman, 1944a) that led him to think of it as a structural theory (Guttman, 1971a), rather than as a process of quantifying amounts of latent traits, and to the development of *facet theory* and *partial order scalogram analysis* (Shye and Elizur, 1994). In the 1980s, Jean-Paul Doignon and Jean-Claude Falmagne developed *knowledge space theory* (Doignon and Falmagne, 1999), later evolved into a theory of learning spaces, in which assessment constructs are represented as partially-ordered sets.

Applications of facet theory and knowledge space theory (including related approaches such as Tatsuoka (2009)'s *rules space* and Leighton and Gierl (2007)'s *cognitive diagnostic models*) normally assume or overlay quantitative latent variable models, to account for 'underlying' proficiencies or competencies that determine a learner's progression through such partially-ordered outcome spaces.

However, from the mid 1990s onwards, there has been a strand of research investigating how to extend knowledge space theory to incorporate a focus on skills and competence, leading to the development of *competence-based knowledge space theory* (see e.g. Stefanutti and Chiusole, 2017). Here, a learner's proficiency or competence is itself conceptualised as a partially-ordered space, rather than a quantity. Ganter and Glodeanu (2014) and Ganter et al. (2017) suggested that formal concept analysis could be applied to study competence-based knowledge space theory, and this is now starting to be done.

For example, Huang et al. (2023) consider how to transform maps from competence-states to 'knowledge-states' (types of demonstrated performances) into formal contexts, and hence to represent them as concept lattices. Each node in the lattice then embodies a knowledge-state and a competence-state as its extent and its intent respectively. This is clearly analogous to the approach set out in section 5.4 above.

A very clear application of these methods is to formative, adaptive, assessment and learning systems, where, for instance, they provide an alternative to traditional IRT-based adaptive tests that is more grounded in a theory of learning.

To date there has been less attention to examining summative assessment, and what is often called 'educational measurement', from this perspective. Yet, as argued above, application of non-quantitative approaches needs to be investigated here too, since the pragmatic 'as if' approach to routine application of latent variable models is not always justifiable.

5.6 Connections to artifical intelligence

A final reason why it is imperative to pursue research in this area is the rapidly growing appication of machine-learning methods, and generative artificial intelligence in particular, in educational contexts. For example, Li et al. (2023) report on using the large language model ChatGPT to score students' responses to (essay style) examinations, and to provide rationales for the scores awarded.

Because the outputs of generative AI applications using large language models are no more than statistically plausible sequences of words, albeit expressed in well-formed natural language, their validity, fairness and reliability is hard to establish theoretically. That is because they are produced using so-called *subsymbolic* approaches to AI (see e.g.

Sudmann et al., 2023), such as deep neural nets, rather than *symbolic* methods that aim to use forms of explicit logical inference to arrive at results: analogously to reasoning about a learner's response to a task with reference to criteria for betterness that define the kind of proficiency one intends to measure by administering the task.

An interesting angle opened up by the qualitative measurement approach described above is the possibility of combining formal concept analysis with neural networks to enhance the explainability of, for example, scores derived from applying a classifier based on a large language model to learners' performances on an examination.

Some initial work in this area has been done by Hirth and Hanika (2022) and Marquer (2020), among others. This kind of analysis could complement quantitative approaches to explaining marks or scores awarded to learners' responses, such as dimension-reduction of the high-dimensional vector space that the language model uses to represent lingustic artefacts—such as learners' responses to assessment tasks—as numerical vectors. In fact, Bradley, Gastaldi, and Terilla (2024) have recently shown that there is a relationship between quantitative techniques based on linear algebra, such as latent semantic analysis, and formal concept analysis, such that the latter can be seen as a more general form of the former. They have applied formal concept analysis to elucidating how semantics appears to arise from syntax, and to study the stucture of semantics, when large language models are used to produce outputs from qualitative data.

Clearly, the practice of educational (and psychological) measurement is changing as technology changes. Tasks can be administered digitally; the widespread availability of devices with reasonable processing power means the possibilities for task design are much more open than they were a decade ago, and they will continue to evolve. The data that is gathered about learners, given their responses to these tasks, can be more unstructured than category-labels or scores: it may be text, audio, or video, and/or representations of such data for example in a vector-space language model. To the extent that human

assessors form part of measurement procedures, for example to apply scoring rubrics, they may be partially or wholly replaced by AI.

What remains fundamental, however, is the need to base these measurement procedures in a theory of what defines or constitutes better or worse proficiency, in the domain of interest, and hence what substantive and semantic content is entailed in statements such as 'this learner got a score of 137', or 'this learner has 1.07 logits of proficiency'; or 'this learner has demonstrated three of the four prototypical aspects of proficiency that define a "grade B" standard', or whatever — what it means to locate them, via a measurement, at a certain position in a (quantitative or other) space.

5.7 Discussion

5.7.1 Qualitative educational assessment is possible in principle, and includes quantitative measurement as a special case

This paper has argued that it is not warranted to assume the phenomena studied in psychometrics, and in educational measurement in particular, are necessarily appropriately conceptualised as quantities. In cases where an assumption of quantitative structure *is* appropriate, then measuring an instance of such a phenomenon means locating it at a point on the real continuum. In cases where the assumption is not appropriate, the idea of measurement becomes, more generally, locating the measurand in a suitable logical space, that is defined in a way that is relevant for the phenomenon.

When the measurand is quantitative and the logical space is the real numbers, the usual methods of psychometric analysis for estimating latent parameters can be deployed. But, *contra* Thurstone (1928), the paper has argued that it is not necessary to "force" theoretically well-supported constructs into a more reductive quantitative form if that is

not appropriate. Hence the argument of this paper is not that psychometrics should be replaced, but that its repertoire of measurement approaches should be widened to cope with measurands that are intrinsically non-quantitative in nature.

The paper suggests that the outcomes of educational measurement procedures can be thought of, in general, as fuzzy relational systems; and that fuzzy formal concept analysis is an appropriate tool to describe data models for the measurands they aim to locate. These models instantiate the 'betterness' relation for the measurand: they model the notion of 'what good performance looks like'. Such an account or understanding is prior to, and necessary for, an understanding or agreement as to 'what being (more or less) proficient' means, in an educational domain. It forms the theory of the construct (one might say, the theory of *value* for the construct, and hence a foundation for evaluation of construct *validity*).

5.7.2 Educational constructs are contestable, intersubjective, temporally-located phenomena

These theories of constructs such as proficiency or competence in a domain are necessarily contestable, intersubjectively constructed, and liable to change over time. Intersubjectivity (Chandler and Munday, 2011) refers to the mutual construction of relationships through shared subjectivity. Things and their meanings are intersubjective, within a given community, to the extent that the members of the community share common understandings of them. Thus the community that constitutes the competent authority for defining an educational construct decides what particular knowledge, skills, and understanding it will encompass, and what will count as better or worse configurations of these aspects as possible ways of being proficient in the domain in question. Thus, for instance, the job of someone marking responses to an examination that is designed to measure that construct is to apply the mutually constructed and agreed standard consistently to each

response she marks (irrespective of whether she personally agrees that it is the 'right' standard).

We do not have to think of data models that encode these intersubjective constructions as (more or less accurate) representations of some objective or underlying 'true' account of the measurand in question. As van Fraassen (2008, p. 260) notes, 'in a context in which a given model is *someone's* representation of a phenomenon, there is **for that person** no difference between the question *whether a theory fits that representation* and the question *whether that theory fits the phenomenon.*'

5.7.3 More research is needed on using partial orders in practice, on linking different assessments of the same construct, and on fuzzy valuations

Section 5.4 argued that in general the data models for measurands such as proficiency in an educational domain are partial orders. This perhaps goes against a relatively strongly ingrained concept of educational assessment as synonymous with *ranking* (e.g. Holmes, Black, and Morin, 2017). Yet in many cases, once a theory of (betterness for) a construct has been settled, rankings are neither necessary nor needed. Two learners' proficiency values may simply be qualitatively different (non-comparable). For instance in Figure 5.2, this is the case for learners 3 and 6. But both learners 3 and 6 have performed better than learner 1. So if learner 1's performance was sufficient to merit a 'pass' grade, let us say (or was picked out as a 'pass' grade prototype), then we know that learners 3 and 6 are also sufficiently proficient to be awarded a pass, even though it is not meaningful to say that their actual demonstrated proficiencies were the same, or that either one is more or less proficient than the other. More work is needed on the scope for using visualisations such as concept lattices to help educational assessment designers and teachers engage

with and interrogate the outcomes of educational measurement procedures (see, for a start, Bedek and Albert, 2015a).

A common application of quantitative latent variable models is to *equating* or *linking* different forms of tests of learners' proficiency in a certain domain. Typically, equating studies are designed to answer questions like 'what score on form X of a test is equivalent to (represents the same level of proficiency as) a given score on form Y of the test?'. In practical applications in many educational contexts however, such as grading students' responses to school-leaving examinations (Newton et al., 2007), one is not so much interested in constructing a monotone map from scores on X to scores on Y, as in ensuring that the levels or kinds of proficiency demonstrated by students graded, say, A, on this year's examination, are 'equivalent', or 'of a comparable standard' to the type of proficiency demonstrated by students graded A on last year's examination.

An area for further research is how to implement such comparability studies in the fuzzy-relational approach to educational assessment proposed in this paper. For example one could take the students graded A on each of the two forms of an assessment, and examine the intents of the formal concepts that form their largest factors (cover an appreciable proportion of the data, in the terms of Bartl, Bělohlávek, and Scharaschkin, 2018). Are these sufficiently similar to count as equivalent demonstrations of proficiency, and what criteria should be applied to appraise similarity?

A deeper question is how the truth degrees that summarise each learner's demonstration of each construct-relevant attribute are determined. In some cases this is straightforward in practice (e.g. for dichotomously-classified test items such as multiple-choice questions); but when judges are needed as part of the measurement procedure, different judges may give different truth values, so what counts as a reasonable or acceptable value? A full account of this aspect of qualitative valuation may need to draw on *rough fuzzy logic*

(Dubois and Prade, 1990; Bazan, Skowron, and Swiniarski, 2006), itself an active area of research in machine learning. Certainly more research is needed here.

Having said that, there is strong support for connecting fuzzy relational structures to cognitive theories of concept formation, when exploring the question of how experts—and these days, AIs—learn to categorise (value) responses to tasks, given some prototypical exemplars: see for example Bělohlávek and Klir (2011).

The outcomes of educational measurement procedures are ultimately underpinned by value judgements about exactly what to assess and how to assess it. As Wiliam (2017, p. 312) puts it: 'whereas those focusing on psychological assessment tend to ask, "Is this correct?", those designing educational assessement have to ask, "Is this good?"'. So questions about how to use mathematical methods in these contexts, in a way that leverages their power, but is not unduly reductive, will no doubt always be debated. It is hoped this paper makes a helpful contribution to that debate.

6 [Paper 3] Beyond latent variable models for non-quantitative constructs: applying fuzzy relational methods to the analysis of examination standards

ABSTRACT

What it means to be proficient in an educational domain—such as 'English writing', or 'basic numeracy', or 'A level French'—is generally a contestable, multifaceted, necessarily fuzzy concept, that is constructed intersubjectively by a certain expert group, at a certain point in time. This suggests conceptualising the property of *a learner's proficiency* in an educational subject-area as having the mathematical structure, not of a *quantity*, but of a particular kind of directed *network*, in which learners may belong to different profiency-states to fuzzy degrees, and the states themselves are partially ordered, from a bottom (no proficiency in the domain) to a top (mastery of the domain). This paper demonstrates how a methodology for analysing such structures can be seen as an extension of traditional latent variable methods for quantitative data. It then uses this methodology to outline

a new approach to appraising examination grading standards over time, that directly addresses the question: 'is the *kind of attainment* that students have to demonstrate to be awarded a particular grade *qualitatively equivalent* between different examination versions?'. The approach is illustrated using data from high-stakes examinations in English language and mathematics, and its benefits and limitations are discussed.

6.1 Introduction

6.1.1 Aims of this paper

This paper has two aims. Firstly, to compare a methodological approach for the analysis of the data generated in educational assessments, treated as fuzzy relational systems, with more traditional approaches that use such data to estimate values of quantitative latent variables. The paper illustrates how both qualitative and quantitative representations of learners' proficiencies, in the domain being assessed, can be obtained from methods of matrix factorisation. In the quantitative case, this leads to proficiency scores as points on a continuum; whereas qualitative proficiencies are not scores, but specific kinds or 'prototypes' of performances.

The second aim is to apply the qualitative mathematical approach to appraise whether, when students' examination grades are determined using the methods currently applied for certain high-staked examinations (GCSE examinations in England), the performance standards associated with particular ('key') grades have been maintained over time (i.e. between different versions of the examination in a subject).

The substantive question of *maintaining performance standards* in the English public examination system has been studied extensively, and methods that use various statistical data and forms of expert judgement have been developed to address it (see, e.g.,

Newton et al., 2007). Some of these underpin the current regulatory framework for public examinations in England that is stipulated by the standards regulator Ofqual. But to date there has been no application of *qualitative mathematical* methods in this arena to analyse directly the question: 'has the grade X standard in a subject been set such that grade X candidates' proficiency is qualitatively equivalent between different versions of examinations in the subject?'. This paper therefore makes an original contribution to the field.

6.1.2 Outline

The conceptual framework that underpins the method of studying examination standards explored in this paper is outlined in section 6.2. It builds on the the approach taken by Bartl, Bělohlávek and Scharaschkin (2018), where the data analysed is interpreted not as providing numerical scores, but rather truth-values of propositions about learners. In order to locate this approach in a context that is more familiar for psychometricians, section 6.2 shows that the fundamental building blocks of a qualitatively-structured conception of proficiency, namely the so-called *formal concepts* derived from the examination data, are analogous to the fundamental building blocks of quantitative latent variable models (such as item response models), namely the *eigenvectors of the covariance matrix* for item scores.

The formal concepts that account for most of the data play a similar role to the eigenvectors with the largest eigenvalues in quantitative models. They can be thought of as fundamental configurations, types, or ways of demonstrating proficiency with respect to the construct being assessed in the examination. Thus by comparing how, or whether, these proficiency-factors change between different administrations of an examination, we can explore the extent to which a grading procedure has maintained performance standards.

Section 6.3 then describes the problem of maintaining performance standards over time for high-stakes assessment procedures such as public examinations in the UK. It summarises how the problem is currently handled in practice, in order to award GCSE and A level qualifications (assessed by externally set and graded examinations generally taken by learners aged 16 and 18 respectively). It suggests that current practice is an uneasy alliance of judgemental notions of 'the standard of a grade', on the one hand, and statistical methods of linking or equating test forms between years, on the other.

For GCSE examinations the 'key' grades, where experts are expected to form a view as to whether students' work meets the qualitative performance standard expected of the grade, are grades 1, 4, and 7. At these grades 'judgement' and 'statistics' have to be reconciled to come to a conclusion on performance standards. Substantively the decisions at grades 4 and 7 have the largest impact on examination outcomes.

Hence this study focuses on grade 4 and grade 7 performances on a single paper (paper 1 of GCSE English language, and paper 1 of the higher tier of GCSE mathematics), and uses the methodology described in section 6.2 to compare them between years with respect to construct-relevant attributes. The aim is to investigate whether the grades assigned to candidates based on the actual practice of summing marks and thresholding result in students performances at grades 4 and 7 being similar or different between 2017 and 2018.

Section 6.4 describes the structure of the assessments studied, and the datasets used. The results of the analyses are then given in section 6.5.

Finally section 6.6 discusses the substantive conclusions that can be drawn from this study, and some of the wider questions that this qualitative mathematical approach raises for designing, and drawing inferences from, educational assessment procedures

more generally. It concludes with some suggestions for areas in which further research would be valuable.

6.2 Conceptual overview: calculating qualitative proficiency factors (fuzzy formal concepts)

6.2.1 Assessments as fuzzy relational systems

The data generated by an assessment procedure such as an examination is usually collated as a table in which each row represents a candidate, and each column represents an item, so that the entry in row i, column j, is candidate i's score (mark) on item j. Michell (2009b, p.45) notes that any such table of numbers is 'shorthand for a set of propositions that tell where the numbers came from. Furthermore, deductions from a data set are inferences from these propositions'. Thus in the case of tests made up of dichotomous items, an entry of 1 in position (i, j) is shorthand for the proposition 'candidate i demonstrated attribute j', where attribute j is the component of knowledge, skills, or understanding tested by item j[1]. Conversely an entry of 0 in position (i, j) is shorthand for this proposition being false, rather than true. If the test consists of only two items, then a candidate's having a total score of 2 allows us to infer that the candidate demonstrated attribute 1 *and* attribute 2. From a score of 1 we can infer the candidate demonstrated attribute 1 *or* attribute 2; and a score of 0 means they demonstrated *neither* attribute 1 *nor* attribute 2.

[1] Of course the attributes selected to provide the information necessary to measure learners with respect to the assessment construct must, in some sense, cohere. They must capture an appropriate range of different aspects of, or viewpoints on, the phenomenon of interest. This is necessary for the assessment to be 'about' some kind of unified construct, rather than a mere unconnected collection of properties. The constructs assessed in public examinations in England are examples of what Cartwright and Runhardt (2014) call '*Ballung* concepts': concepts with a fuzzy and context dependent scope. Section 6.4 discusses how construct-relevant attributes are defined for the assessments of proficiency with respect to the two specific constructs studied in this paper.

Hence a candidate's item-scores on a test made up of dichotomous items can be viewed as the *truth values* ($1 = \texttt{true}$; $0 = \texttt{false}$) of propositions about the candidate's demonstrated proficiency with respect to specific attributes relevant to the construct being assessed by the test. Combining item scores by the numerical operation of addition is equivalent to combining these truth values via the logical operations of conjunction and disjunction. For example, considering now a three-item test, a total score of 2 is equivalent to the proposition $(a_1 \wedge a_2) \vee (a_1 \wedge a_3) \vee (a_2 \wedge a_3)$, where '$a_j$' means 'demonstrated attribute j', '$\wedge$' means 'and', and '$\vee$' means 'or'.

This way of thinking about assessment data can be generalised to consider propositions whose possible truth-values may be more extensive than $\{\texttt{false}, \texttt{true}\}$. *Fuzzy logic*, (or *multivalent logic*[2]) allows propositions to have truth values drawn from ordered sets of *truth degrees*, such as $\{\texttt{false}, \texttt{partly-true}, \texttt{true}\}$, or $\{\texttt{false}, \texttt{somewhat-false}, \texttt{somewhat-true}, \texttt{true}\}$ (Hajek, 1998a; Belohlavek, Dauben, and Klir, 2017). This is often appropriate when there is inherent fuzziness, vagueness, or uncertainty in the concepts being discussed or analysed (see e.g. Goguen, 1969a; Goertz, 2006a; Belohlavek, Dauben, and Klir, 2017).

Scharaschkin (2024) suggests that fuzzy logic lends itself to the analysis of assessment procedures, because in general, learners are given tasks, such as questions on a test, and they may be successful in engaging with them *to a certain degree*. The outcome of a learner's interaction with an item is not necessarily captured by the crisp dichotomy of {correct, incorrect}. However, treating polytomous response categories as if they were simply a sequence of intervals imposed on an underlying quantitative latent continuum, as is typical in psychometric modelling, is problematic if 'levels of proficiency' are conceived of as configurations of cognitive resources that may differ not only in degree, but also in kind (Michell, 2012b). Scharaschkin therefore proposes modelling educational assessment

[2]Multivalent logics admit more than two degrees of truth. The term *fuzzy logic* is often used to refer to logics in which the range of truth-values is the real unit interval [0,1]. However in this paper the term is used in its more general sense, as described in Appendix D.

outcomes as *fuzzy relations between learners and construct-relevant attributes*. It may only be partially true that a learner demonstrates an attribute (itself assessed by an item or items on a test). If we think of the information generated by an assessment procedure in this way, then we have to deal with tables or matrices of *truth-degrees*, rather than matrices of *numbers*.

It is perfectly possible to calculate with such 'logical' matrices, just as it is with the more familiar 'numerical' matrices. The difference is that in the former case, calculations mean logical operations taking place within a truth-degree structure, rather than arithmetic operations taking place within a numerical structure.

Appendix D gives some more detail on the technicalities. There are several different approaches to modelling truth-degrees in practice, but in all the examples used in the rest of this paper, we will work with a (totally-ordered) set of degrees that will be labelled $\{0, 1/4, 1/2, 3/4, 1\}$, where 0 means '(completely) false' and 1 means '(completely) true'. These possible truth values are ordered as per the standard order relation on numbers. Each attribute for the assessement constructs considered in sections 6.4 and 6.5 is therefore regarded as a property that may be manifested with any of these degrees of truth by any particular learner, when faced with a task or tasks that aims to elicit information about that aspect of the learner's proficiency in the domain in question.

What are these contruct-relevant attributes? They are the aspects of the assessment construct that define it. The assessment constructs considered in this paper are (proficiency in) GCSE English language, and GCSE mathematics. Now, for the constructs assessed in public examinations in England, as noted in Scharaschkin (2017), the relevant attributes are specified by the so-called *assessment objectives* and *subject criteria* that are determined by the government for the examinations in question. The subject criteria define the specific content about which learners are expected to have knowledge,

and which they are expected to employ in demonstrating the skills described in the assessment objectives.

This is made concrete in section 6.4, for the examples of GCSE English language and GCSE mathematics: the two assessments considered in this paper. These two subjects were chosen for analysis because—apart from their intrinsic importance given nearly all students are expected to attempt to gain these qualifications—they are very different in the way they are specified. In English language there is *no* specific subject content. There are only assessment objectives (nine in total; five tested on the examination paper discussed below). In mathematics there are *six* areas of specific subject content, each of which is elaborated in considerable granular detail), and only three assessment objectives.

6.2.2 Formal concepts and concept lattices

Consider again the case of a test, taken by m learners, that consists of a sequence of n dichotomous items $\{a_1, \ldots, a_n\}$. Here we assume each item a_j is assessing a separate construct-relevant attribute. Then we can, of course, represent the outcomes of the test as an $m \times n$ matrix of 0s and 1s. But we can also represent these outcomes as a *partially-ordered set*, or a *network* (directed graph) of *types of performances*, that is to say, ways of demonstrating proficiency in the domain assessed by the test, as follows.

Given a subset of learners $L_1 := \{l_1, ..., l_k\}$, let $A_1 := \{a_1, ..., a_j\}$ be precisely those items that all learners in L_1 got correct. Then the pair (L_1, A_1) is an instance of a *formal concept* present in the data. L_1 is called the *extent* of the concept, and A_1 is called its *intent*. We can equally well start with a subset $A_2 := \{a_1, ..., a_p\}$ of items, and then form the concept (L_2, A_2), where L_2 is precisely the set of learners who got all items in A_2 correct.

Formal concepts are (partially) *ordered* via the set-theoretic notion of inclusion. A concept (L_1, A_1) is *more general* than a concept (L_2, A_2) if it is exemplified by more learners ($L_1 \supseteq L_2$), or (equivalently), if it is described by fewer attributes ($A_1 \subseteq A_2$). The most general concept is the one that has the largest extent (and smallest intent). In test performance terms, the most general concept corresponds to the bottom, or worst, performance: because every other performance has a larger intent (i.e. entails more correct items). Similarly, the least general concept (with the smallest extent and largest intent) corresponds to the top, or best, level of performance. The partially ordered collection of formal concepts is called the *concept lattice* for the matrix in question[3].

We can think of formal concepts as different ways of performing on the test (i.e. different *ways of demonstrating proficiency* in the subject domain). Each type of performance—or demonstration of proficiency—can be described *extensively*, by showing the (actual responses of the) learners who *demonstrate* it, or it can be described *intensively*, by showing the item-profiles that *characterise* it.

If we generalise to the case where the assessment procedure collects information, not on crisp (dichotomous) items, but on fuzzy items, each formal concept becomes a pair of fuzzy, rather than crisp, sets. Its intent is a collection of attributes that describe it with partial degrees of truth, and its extent is the collection of all learners who exemplify it to a degree.

To make this concrete, Figure 6.1 shows the lattice, or network, of fuzzy concepts for a dataset consisting of 20 candidates who took AQA GCSE English paper 1 in 2018. As

[3]Normally concept lattices are represented graphically as networks called *Hasse diagrams* (Figure 6.1 is an example), with the least general concept at the bottom, and the most general concept at the top. An arrow is drawn upwards from concept A to concept B if B is more general than A. In the educational assessment context, however, we naturally regard the best performance as the *top* concept, which means we need to reverse the usual ordering (in mathematical terms, we use the *dual* lattice). This is done throughout this paper, so for instance in Figure 6.1, the worst level of proficiency exhibited by the learners in this example is at the bottom of the diagram, and the best level is at the top.

discussed in section 6.4.1, there are six attributes of English language proficiency, here labelled y_1 to y_6, assessed in this examination.

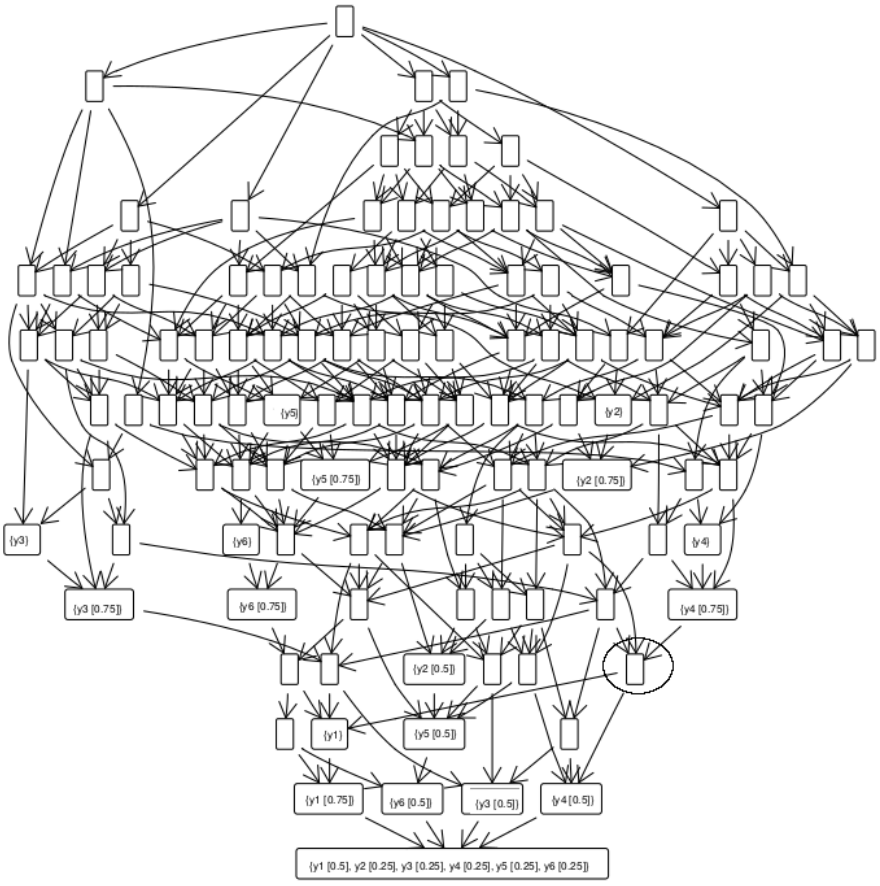


Figure 6.1: Concept lattice for 20 learners on GCSE English language paper 1

Figure 6.1 can be read in the following way. The boxes are fuzzy formal concepts (possible ways of demonstrating proficiency in the assessment construct). The bottom box is the worst level of proficiency displayed by the 20 students in this sample (the first attribute y_1—'identify and interpret'— displayed to degree 1/2, and the other five attributes all displayed to degree 1/4. All 20 students in this sample participated in this concept (proficiency level) completely (to degree 1). That is, their performances all fully exemplified at least this level of proficiency.

Then each box above that in the diagram represents a type of performance such that the attributes are displayed to the degrees of the highest-degree attributes connected in any continuous path below that node. For instance, the attributes of the circled node are: $y_1(1)$; $y_2(1/4)$; $y_3(1/4)$; $y_4(1/2)$; $y_5(1/4)$; $y_6(1/4)$.

The top box has all attributes displayed to the full degree. No student in this group of 20 participated in this concept fully (i.e. none of these 20 students fully demonstrated all the construct-relevant attributes of performance on this assessment). In fact the degrees of participation in this specific concept (i.e. the degree to which their performances exemplified the proficiency level of 'completely demonstrates these attributes'), for students 1 to 20, were respectively 1/4, 1/4, 1/4, 1/4, 1/4, 1/4, 1/2, 1/2, 1/4, 1/4, 1/4, 1/4, 1/4, 1/2, 1/2, 1/4, 1/4, 1/4, 1/2, 1/2.

This example shows the types of proficiency demonstrated (fuzzily) by only 20 students. As the number of students increases (and the typical total entry for this paper is around 500,000 in each summer examination series), the concept lattice grows rapidly, and and the full lattice quickly becomes too dense to be useful.

Thus a form of data reduction is needed, and this is the subject of the following section. Essentially it provides a way of choosing certain formal concepts (whose extents provide, as it were, prototypical *exemplars*, and whose intents provide stereotypical *descriptors*) that capture the main features of the performances of students who were awarded, say, a grade 4 on the examination. It could be seen as a qualitative analogue to the estimation of a small number of quantitative latent factors, for assessment constructs that can be reasonably taken to be quantitatively structured.

6.2.3 Factor analytic methods for fuzzy data

Latent variable models for quantitative data, such as candidates' item scores on a test, can often be framed in terms of factorising the matrix of scores, as set out in Appendix E. For instance, given an $m \times n$ matrix M of m candidates' scores on the n items that make up a test, we may wish to investigate whether we can find an approximate factorisation $M \approx LF$, where L is a $m \times k$ matrix of *loadings*, and F is a $k \times n$ matrix of *factors* (with $k < n$). The idea is that the *manifest data* (e.g. test outcomes) are expressed (approximately) as linear combinations of a small number of *latent factors*. For instance, in the unidimensional case, if the n-dimensional vectors of candidates' item scores tend to lie approximately around and along a line, we could regard each vector's projection onto that line as representing the profiency value or score of the associated candidate.

When the entries in the data matrix M are not numbers, but truth-values, it is likewise possible to examine matrix factorisations $M = L \circ F$, with the difference that (because we are no longer working within the field of real numbers) the matrix product operation $\circ$ is defined in terms of logical operations in the truth-value structure, rather than in terms of sums and products of numbers. The key theorem is that the *factors are particular formal concepts* from the network of concepts associated with the data matrix M (Bělohlávek, 2012b) .

That is, 'logically factorising' the matrix of truth-degrees representing the outcomes of an educational assessment procedure is equivalent to 'picking out key concepts'. The factors are the (extents and intents of) specific formal concepts, that is, *specific ways of demonstrating proficiency* in the domain in question.

Specifically, let M_{ij} = degree to which student i demonstrates attribute j. Factorise M so that $M = L \circ F$. Then:

- L_{il} = degree to which student i **exemplifies** (is in the extent of) **factor** l.

- F_{lj} = degree to which attribute j is one of the **descriptors of** (is in the intent of) **factor** l.

Columns $1, 2, \ldots, k$ of L are the **extents** of factors $1, 2, \ldots, k$. Rows $1, 2, \ldots, k$ of F are the **intents** of factors $1, 2, \ldots, k$.

How do we choose which formal concept will constitute each factor? Analogously to the quantitative case, we want to minimise the loss of information that occurs by reducing a dataset with n attributes to one that has only k ($< n$) factors. So if $k = 1$, we want to choose a formal concept F_1 such that the product of the $m \times 1$ matrix L and the $1 \times n$ matrix F, constructed as in the preceding paragraph, most closely reproduces the matrix M. If $k = 2$ we want to choose F_1 and F_2 such that the product of the $m \times 2$ and $2 \times n$ matrices best approximates M, and so forth. In general, we have a problem analogous to the quantitative problem of approximating a matrix of item scores on a test by another matrix of lower rank. Appendix E summarises the solution to this problem proposed by Eckart and Young (1936), and shows that it amounts to choosing the dominant eigenvectors (the eigenvectors with the largest eigenvalues) of the item covariance matrix.

Further, it is argued in Appendix E that, in common psychometric models such as the Rasch and 2-parameter logistic models, it is the *dominant eigenvectors* (of the item covariance matrix or an item comparison matrix) that are the *quantitative latent factors*. The analogy for the non-quantitative approach adopted in this paper is that the *dominant formal concepts* (of the truth-degree matrix) are the *qualitative latent factors*.

By 'dominant', in the quantitative case, we mean 'associated with the largest eigenvalues'. In the qualitative case we mean 'associated with the largest coverage values', where the

coverage value of a factor is the extent to which adding it to an existing (possibly empty) list of factors increases the accuracy of the approximation of M that is obtained from the product $L \circ F$. Thus, a collection of factors perfectly covers the data M if the product $L \circ F$ reproduces the matrix M. It covers 50% of the data if it reproduces 50% of the entries of M, etc. We can define 'reproducing' entries to mean either matching them exactly, or (since the entries themselves are fuzzy truth values) matching them approximately, such that the original and reproduced entries are equivalent in a fuzzy-logical sense[4]. Both metrics are reported for the analyses in section 6.5.

When dealing with quantitative models, heuristically, we look at scree plots to assess the extent to which one, or a small number, of quantitative factors are dominant (i.e. to judge the 'dimensionality' of the data). If the ratio of the dominant eigenvalue to the sum of the eigenvalues, expressed as a percentage, is $x\%$, we say that the associated factor 'explains' $x\%$ of the variance in outcomes. The equivalent plot we can use for qualitative data is a coverage plot, such as Figure 6.2a below, showing how the coverage rate increases as more factors are selected by the factor-selection algorithm. We likewise can say that $x\%$ of the data is 'explained' by the chosen factors, if they cover $x\%$ of the data matrix.

The algorithm that is used in this paper to select factors proceeds in a 'greedy' manner[5]. It first selects the factor F_1 that explains (covers) the largest part of the data M. It then selects a second factor F_2 that, in combination with F_1, explains the largest part of the data. This process continues with a third factor, until the data is fully explained, or a given stopping criterion (e.g. $k = n$) is reached. Some limitations of this algorithm are discussed in section 6.6.

[4] the proportion of entries that are similar, between $L \circ F$ and M, is calculated in this study as $\frac{\Sigma_{i,j} M_{ij} \leftrightarrow (L \circ F)_{ij}}{mn}$, where $\leftrightarrow$ is the Łukasiewicz biresiduum operation: see Appendix D

[5] The algorithm is described in Bělohlávek and Vychodil (2016). It was coded in R for the analyses in this paper

Before discussing the results of applying this fuzzy-matrix factorisation approach to actual examination data, the following section sets out the context and the substantive question of interest, namely the maintenance of standards over time.

6.3 The problem of maintaining qualification standards over time

6.3.1 Performance standards, attainment-referencing, and the meaning of grades

Nearly all 16-year-old school pupils in England (typically around 700,000 annually) sit externally set and marked examinations whose results determine the grades they are awarded for qualifications known as GCSEs. Around 300,000 18-year-olds enter for the award of qualifications known as A levels—the main determinant of entry to university—which are similarly externally assessed by exam boards[6].

The qualifications exam boards award (the certificates they issue) are effectively *promises*, underwritten by these assessment procedures, that testify to the fact that a student has demonstrated a certain level of proficiency in a subject (for instance, that Mary has demonstrated grade A level proficiency in A-level chemistry, say). *Maintaining performance standards* requires that these promises retain the same meaning over time. A candidate awarded a grade A in chemistry in 2023 has demonstrated equivalent chemistry proficiency (in terms of the kinds of things they know, understand, and can do) to a candidate awarded a grade A in 2024.

[6]GCSE formally stands for General Certificate of Secondary Education, a term introduced in 1988 to replace the previously existing Certificate of Secondary Education, and the 'Ordinary level' (O level) of the General Certificate of Education, the latter also being available at 'Advanced level' (A level). These days the terms 'GCSE' and 'A level' are effectively simply used as nouns in their own right

There is a strong sense, in the UK context, that grades do carry substantive meaning of this kind: that being classed as, for instance, 'grade 4' in a GCSE (a grade often colloquially referred to as 'a pass') means qualitatively more than simply having scored a certain number of marks. Teachers may refer to 'a (typical) grade 4 student', for instance, or consider a student to be 'on track' to being a grade 4 student, because they have demonstrated certain facets or aspects of proficiency to sufficient degrees, based on a mental model of what a 'grade 4' level, or style, of demonstrated proficiency tends to 'look like'.

Newton (2020, p.4), in an official government document published by the qualifications regulator Ofqual, references this notion of meaning in stating that

> During normal times, we tend to view the maintenance of standards primarily through the lens of **meaning**; that is, in terms of how grades need to be **interpreted**. We strive to ensure that equivalent grades, across successive versions of the same subject exam, can be **interpreted** in the same way; that is, in terms of attainment. We say that exam standards have been maintained when equivalent grade boundary marks across adjacent exams correspond to equivalent levels of attainment. We call this principle Attainment-Referencing

What, though, is a 'level' of attainment? Newton goes on to say (p.6) that

> We work on the assumption that each mark on the mark scale (from zero to the total mark for the exam) can be described in terms of a specific level of attainment. This is the level of attainment that it takes to score that many marks on the exam. ... For the purpose of setting standards and awarding grades, we assume that all candidates who are awarded the same number of marks share the same level of attainment. Hence, the level of attainment that

> is common to candidates at the grade boundary mark constitutes the exam standard for that grade.

Here 'attainment' is defined as 'what it takes to score x marks'. This 'what it takes' could be interpreted as a property of examinees' responses, namely the qualitative features or attributes of a candidate's response—for example the essays they wrote in answer to the examination questions—such that applying the scoring rubric(s) and mark aggregation rules to them yields a mark of x. Alternatively, it could be interpreted as an hypothesised property of the examinees themselves: their proficiencies in the subject matter, such that examinees with certain levels or configurations of proficiency have what it takes to (produce performances that) score x marks. In other words, it appears to refer to either *hypothesised*, or *demonstrated*, *proficiency* in the examination subject in question.

The assumption that there is a one-to-one correspondence between possible configurations or levels of proficiency and possible total examination marks is tantamount to assuming that the former is structured as a totally ordered collection (a ranking) of possible states (one for each total mark). However, as demonstrated by, for example, Scharaschkin and Baird (2000), experts do not always agree that different pieces of work that have been awarded the same total score are qualitatively equivalent in terms of the level of performance they demonstrate. Moreover they typically differ in different ways, with the first better in some respects, and the second in other respects. This is consistent with understanding proficiency as a partially-ordered, rather than a totally-ordered, structure (Scharaschkin, 2024).

The term 'weak criterion referencing' (adopted by Baird, Cresswell, and Newton, 2000) might therefore be a more suggestive way of describing the general principle that *maintaining examination standards over time* requires awarding grades to students in such a way that *qualitatively equivalent* demonstrations of proficiency (with respect to relevant attributes for the construct being assessed) *are awarded the same grade*, irrespective

of which version of the examination a candidate takes. The fuzzy nature of construct-relevant criteria for appraising students' proficiency means that two candidates on the same mark may have different demonstrated proficiency-states in terms of the extent to which they exhibit construct-relevant criteria of creditworthiness. Proficiency states, unlike mark categories, may have fuzzy boundaries. That does not mean, however, that it is impossible to use them as the basis of inferences about performance standards.

The current approach to maintaining GCSE and A level standards in England, though, as stipulated by the regulator Ofqual, is to calculate 'statistically recommended' grade boundaries for each examination using the so-called *comparable outcomes* approach summarised in the next section. Then experts are asked to scrutinise sample examination responses on marks in a short interval around the statistically recommended boundaries[7]. In performing this scrutiny they can look at archive examples of work on the relevant grade boundary mark in a previous, reference, version of the assessment. Their aim is to ensure that the boundary mark set in the current version of the assessment best replicates the performance standard demonstrated in the archive examples, having made allowance for the fact that students have to respond to different questions and perform different tasks in the two versions.

There is a considerable body of research examining how experts make such comparability judgements (see Baird, 2007, for a summary). In short, evidence suggests that the cognitive demands imposed by the this task are such that experts find it difficult to compensate adequately for changes in (the difficulty of) question papers when comparing responses; to base decisions on exemplars of the standard they are presented with, rather than their own mental models of the standard; to make sufficient allowance for the 'compensation' afforded by summing item marks that permits weak performance in one

[7] Although grade boundaries apply at the whole-examination level, the expert decision making is carried out separately for each of the papers–typically two or three–that make up the whole examination. Generally experts are asked to look at examples of candidates' work in a five-mark range centred on the statistically recommended boundary for each grade boundary decision that is taken.

area to be offset by stronger performance in another; and to make consistent qualitative distinctions between candidates' work on adjacent marks. Baird (2007, p. 142) states 'for weak criterion referencing to be acceptable, we must be able to trust that the qualitative judgements made by examiners are reliable and fair'. This is true to the extent that the only source of evidence on candidates' relative qualitative performances between examination versions is examiner judgement. This paper explores, as an alternative, evidence derived in a principled way from the performances themselves, based on markers' judgements of the extent to which they exhibited the construct-relevant attributes captured in the marking rubrics.

6.3.2 The comparable outcomes approach

Ofqual (2017b) describes the principle of comparable outcomes as requiring that 'roughly the same proportions of students achiev[e] each grade as in previous years, providing the cohort of students is similar to previous years'. Clearly this principle hinges on what is meant by 'similar'. In practice, this is normally defined in terms of the cohorts' distributions of scores on a prior-attainment based reference measure. For GCSE examinations, each student's prior attainment is the average of their normalised scores on English and mathematics tests taken at the end of primary school. For A level examinations, each student's prior attainment is an average of points accrued based on their GCSE results.

This prior attainment measure is often referred to as students' *ability* (e.g. Jadhav, 24 March 2017), and changes in the distribution of the measure between the entrants for different versions of an examination are taken as indicating that the cohorts are likely to differ with respect to the extent to which they will demonstrate the attributes of subject-proficiency being tested in the examination. Predicted grade distributions—and hence suggested grade boundary marks—for the current version of the examination are derived by assuming the same outcomes, conditional on prior-attainment, as for the

reference version. As Bramley and Vidal Rodeiro (2014) note, the comparable outcomes methodology is conceptually similar to the statistical test equating method known as frequency estimation equipercentile equating for non-equivalent groups with an anchor test (Angoff, 1971; Braun and Holland, 1982).

The comparable outcomes approach aims to maintain standards using a quantitative conception of proficiency, i.e. standards are maintained if cut-scores are set such that the minimum *quantity of proficiency* required to obtain a grade X is equal, year on year. The statistically recommended grade boundaries are then intended to be a starting point for discussion by examiners in making their final decision on boundary marks that best maintain performance standards[8]. But in practice 'the statistics' dominate and very largely determine these decisions, partly because of the concerns about potential detrimental impacts of unassisted judgement, addressed in section 6.3.1, and partly because there is no standard framework that can be used to justify the prioritisation of the judgemental evidence against the statistical evidence, with the possible exception of the method of comparative judgement.

6.3.3 Comparative judgement and performance standards

One approach to studying performance standards in terms of expert judgement is to use the method of comparative judgement (Thurstone, 1927a; Bramley, 2007; Jones and Davies, 2023). Experts may be asked to make pairwise comparisons of students' responses (essays, mathematical proofs, sketches, or whatever they might be) from current and reference versions of an examination, in order to calculate numerical scores that are taken

[8] in the 'steady state' case: we are not considering here methods for mitigating the impact of exogenous shocks such as the covid-19 pandemic during 2020-2022.

to represent the quality of each of the judged responses[9]. By regressing marks on 'quality scores' derived in this way, the cut score for a given grade on the current examination can be set to match, as closely as possible, the 'amount of quality' associated with the cut score for that grade on the reference examination.

Many of the factors known to influence experts' judgements of performance quality mentioned in section 6.3.1, such as the difficulty of making sufficient allowance for changes in the tasks when appraising the quality of students' responses to them, are still present when comparative judgement is used. Moreover, as noted by Kelly, Richardson, and Isaacs (2022), claims made for the validity of comparative judgement grounded in assertions about the underpinning psychological processes involved in comparative versus 'absolute' judgements are incomplete and inconsistent. There are, therefore, significant caveats attaching to its use as a tool for studying examination standards over time.

Most significantly in the context of this paper, comparative judgement requires assuming that the *quality* of performances, such as responses to essay questions, is a *quantitative* property, in the same way that other psychometric models are based on the assumption that proficiency is a quantitative property. The approach outlined below, by contrast, avoids this reduction of quality to quantity.

[9] Using the Rasch model implementation of Thurstone's original 'law' of comparative judgement (Andrich, 1978a), this amounts to postulating that $\text{logit}(q_i \geq q_j) = q_i - q_j$, where q_i and q_j are the quality-values for responses i and j. Assuming this model to be true, the observed data can then be used to generate a quality score for each response, for example by maximum likelihood estimation.

6.4 Methodology

6.4.1 Overview of the assessments studied

The data for this study came from GCSE English language and mathematics examinations taken by students in 2017 and 2018. These examinations are taken by nearly all 16 year olds in England. They are offered by four assessment providers (exam boards): this study uses data from the examinations provided by AQA. GCSE qualifications are graded on an ordinal scale with nine categories, from grade 1 (lowest) to grade 9 (highest).

The English language examination consists of two papers of one hour 45 minutes each. The mathematics examination consists of three papers of one and half hours each. Mathematics is offered in two versions (tiers): a less demanding (foundation tier) version targeted at candidates expecting to get up to a grade 5, and a more demanding (higher tier) version targeted at candidates expecting to get up to a grade 9. The higher tier papers cover more content than the foundation tier.

Trained examiners, who have been 'standardised' with respect to the criteria for assigning credit to responses that are summarised in the mark scheme (rubric) for the item, *mark* each candidate's response to each task, that is, assign it to one of a number of ordered categories labelled by integers. In GCSE mathematics assessments the tariff (number of possible evaluation categories) for an item is normally low (for example, 1, 2, 3, or 4 marks), and the mark scheme for the item gives rules for assigning partial credit to responses. Tariffs for GCSE English language items are generally high (for example, between 8 and 24 marks), and the mark scheme indicates typical features or attributes of a response that would tend to place it in one of a small number (typically 4) 'levels', with a final mark then being awarded based on the examiner's judgement of whether the response is, for example, a good, average, or borderline example of attainment at that

level. Sections 6.4.2 and 6.4.3 explain how item marks were used to generate truth-degrees for the extent to which each student's examination response demonstrated construct-relevant attributes for the two subjects investigated.

In order to assign each student to a grade, 'boundary marks' (threshold scores) are determined using predicted grade distributions based on students' prior attainment (in this case their scores on English and mathematics tests taken at the end of primary school), supplemented with expert judgement of examples of students' responses around the suggested thresholds, as discussed in section 6.3. The 'key' grades, where experts are expected to form a view as to whether students' work meets the qualitative performance standard expected of the grade, are grades 1, 4, and 7.

Boundary marks are applied at the level of the assessment as a whole (i.e. to students' total scores across all papers), but subject grade boundaries at the key grades are the aggregate of paper-level grade boundaries which are what examiners actually set when scrutinising work. So while a 'grade 4' student may not actually have a 'grade 4' on each paper separately, teachers often think in terms of the 'grade 4 standard' as being applicable to each paper ('this is the kind of performance I would expect from a grade 4 student on this paper'). Hence this study focuses on grade 4 and grade 7 performances on a single paper (paper 1 of GCSE English language, and paper 1 of the higher tier of GCSE mathematics), and uses the methodology described in section 6.2 to compare them between years with respect to construct-relevant attributes. The aim is to investigate whether the grades assigned to candidates based on the actual practice of summing marks and thresholding result in students performances at grades 4 and 7 being similar or different between 2017 and 2018.

Table 6.1 gives the total number of responses overall, and for each of grade 4 and grade 7, for both subjects and years. There are proportinonally more grade 7 responses in

Table 6.1: Numbers of responses for English and mathematics papers

		Total	Grade 4	Grade 7
2017	English P1	406,884	69,414	40,729
	Mathematics P1(H)	42,969	2,511	11,655
2018	English P1	461,262	77,930	47,384
	Mathematics P1(H)	41,984	2,355	10,280

mathematics than English, because the mathematics higher tier paper is targeted at grades 5-9, whereas the English paper is intended for all candidates.

6.4.2 English assessment attributes

The English language assessment is designed to test five assessment objectives, summarised in Table 6.2. The examination requires candidates to engage with unseen texts, and to write extended prose responses. There are two questions (2 and 3) testing the 'explain and comment' assessment objective, and responses to question 5 are marked separately with respect to both the 'communicate clearly' and 'use language accurately' objectives.

Question 1 asks candidates to list four things that can be inferred from a source text. The remaining questions require students to write connected prose responses, and are marked using so-called *levels-of-response* mark schemes. Examiners are asked to assign a response to one of five categories, based on the given response's degree of match to descriptors of what would typically be expected at each of the levels. The bottom level, 0, is reserved for work that has displays no creditworthy features at all. The top level is for work that displays all, or very nearly all, the features that demonstrate mastery of the given assessment objective in respect of the task.

Table 6.2: English language assessment: attributes

Assessment objective	Label	Q. no.	Tariff
Identify and interpret information in texts	ident	1	4
Explain and comment on use of language in texts	expl1	2	8
Explain and comment on use of language in texts	expl2	3	8
Evaluate texts critically	eval	4	20
Communicate clearly, effectively, and imaginatively	comm	5	24
Use language accurately	acc	5	16

Levels are converted to scores by means of an additional step whereby examiners must decide, effectively, whether a given response is a relatively better or worse example of a level X response, where the granularity of this judgement is determined by the mark tariff for the question. Thus questions 2 and 3 have a total of 8 available marks, so 2 marks are allocated to each level (1-2 for level 1, 3-4 for level 2, up to 7-8 for level 4). Having decided that a response merits a level 2, say, the examiner marking this question would then allocate it either 3 or 4 marks depending on whether it was a 'weaker' or 'stronger' example of a level 2. For questions marked out of 16, 20, and 24, the same approach is used, but with 4, 5, and 6 marks, respectively, available per level.

The reason for having these variable mark tariffs is to meet requirements about the 'weighting' of the different assessment objectives. Since candidates' grades are determined based on summing their marks, the assessment objectives that are supposed to describe the most important aspects of overall proficiency are assigned more marks. In this way the weightings of the different objectives are expressed as percentages of the total marks available. This notion of quantitative weighting is discussed further in section 6.6.

To compare performances qualitatively over time, we work directly with levels, rather than marks. Thus the response to each question is converted to a truth-value for the the proposition 'the response displayed attribute X', using the mapping: level $0 \mapsto 0$; level $1 \mapsto 1/4$; level $2 \mapsto 1/2$; level $3 \mapsto 3/4$; and level $4 \mapsto 1$. The attributes themselves are, in this case, taken to be the assessment objectives in Table 6.2, with the elaboration that the 'explain and comment' objective is tested in two different ways, labelled expl1 and expl2 in Table 6.2. The labels given in the table are thus labels for the attributes that are tested in the two different versions of this assessment (the 2017 paper and the 2018 paper).

6.4.3 Mathematics assessment attributes

The assessment construct for GCSE mathematics is defined in a manner quite a different from English. For mathematics, there are six areas of subject content: number, algebra, geometry, ratio and proportion, probability, and statistics. Within each of these areas, the official government specification sets out detailed granular requirements, e.g. 'substitute numerical values into formulae and equations', or 'identify and interpret gradients and intercepts of linear functions graphically and algebraically' (Department for Education, 2013). There are three assessment objectives, namely:

- AO1: Use and apply standard techniques;
- AO2: Reason, interpret, and communicate mathematically; and
- AO3: Solve problems within mathematics and other contexts.

Each item on the mathematics paper therefore is designed to test these assessment objectives as applied to content from the specification. In practice, however, it is hard to separate both AO1 (which is concerned with recalling and knowing terminology and

notation) and AO2 (which is concerned with calculating and deducing) from the specific content to which they are applied. This can be seen even in the two examples given about of sub-content from the algebra content area, which are both phrased in terms of knowing about, and making deductions within, specific contexts. By contrast, items testing AO3 may require the student to combine knowledge from more than one content area, and/or to translate a substantive problem into a relevant mathematical form. For instance an algebra question may ask the student to solve a pair of simultaneous equations, whereas a problem-solving question may require the student to realise that a problem about the costs of certain items can be solved by setting up a pair of simultaneous equations using information from the context.

Clearly the boundaries between precisely what different items test may be somewhat fuzzy. The problem-solving example above also tests algebra, and because of the hierarchical structure of mathematics content, an item testing, for example, a student's ability to solve a pair of simultaneous equations is also implicitly, at least, testing content in the number domain (e.g. to recognise that $5x + 7x = 12x$ they need to know that 5+7=12). However there is a primary intent, for each assessment item (normally items with tariffs between 1 and 4 marks) as to which content area it is testing, or whether it is intended primarily as a test of AO3. So for the purposes of describing the attributes of proficiency in GCSE mathematics (at least the part of GCSE mathematics tested in this paper), it seems reasonable to code each item as testing either knowing about and making deductions in one of the six content areas, or as testing problem solving (AO3) in general. Thus we obtain seven construct-relevant attributes, as set out in Table 6.3.

Each item in the 2017 and 2018 paper was coded to the relevant attribute, and then a score for each candidate with respect to that attribute was obtained as the sum of their marks on the relevant items. These attribute scores were then mapped to truth-values, such that 0 marks mapped to 0 (attribute not demonstrated); the maximum mark

Table 6.3: Mathematics assessment: attributes

Attribute	Label	No. items 2017	No. marks 2017	No. items 2018	No. marks 2018
Number	num	6	10	8	16
Algebra	alg	11	23	8	16
Geometry	geom	7	17	5	13
Ratio and proportion	ratio	4	5	4	8
Probability	prob	2	4	4	6
Statistics	stats	3	9	4	8
Problem solving	AO3	3	12	5	13

mapped to 1 (attribute fully demonstrated); and then the remaining marks were divided equally between the categories $1/4, 1/2, 3/4$, with any remainder marks first allocated to the middle category ($1/2$), and then to $1/4, 3/4$ respectively. For instance for an attribute with 8 available marks, the degree of demonstration of the attribute is rendered as: $0 \mapsto 0$; 1,2 $\mapsto 1/4$; 3,4,5 $\mapsto 1/2$; 6,7 $\mapsto 3/4$; 8 $\mapsto 1$.

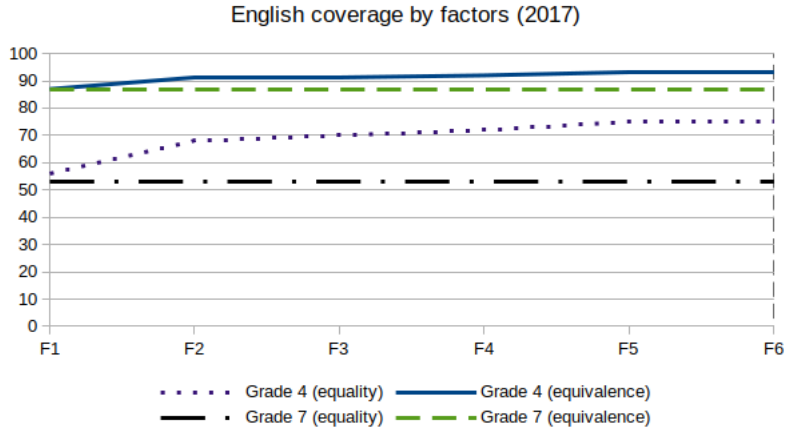
6.5 Results

6.5.1 English

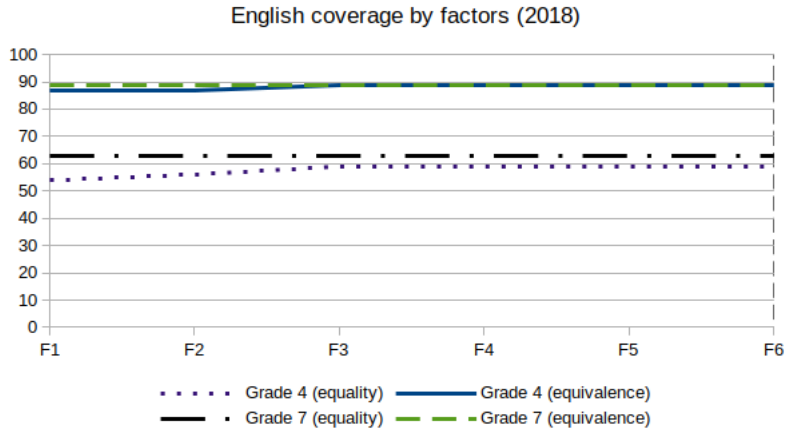
The performances of students graded 4 and 7 in each of 2017 and 2018 were expressed as matrices of truth-degrees, as explained in section 6.4.2. The factorisation method discussed in section 6.2 was then applied to extract the two formal concepts (factors) that accounted for most of the data. Figure 6.2 shows the proportion of data covered, or

explained, according to the two measures (equality and equivalence) described in section 6.2.3. Figures 6.3 and 6.4 show the intents of these concepts, i.e. the attributes associated with 'prototypical' grade 4 or 7 performances each year.

For both English and mathematics, the data coverage rates tend to flatten out after the second factor. This point is considered further in section 6.6.3. And as can be seen from Figures 6.2a and 6.2b, the second factor does not add very much explanatory power, so focussing solely on the first as a fuzzy descriptor of prototypical performance at each grade seems a reasonable approximation.

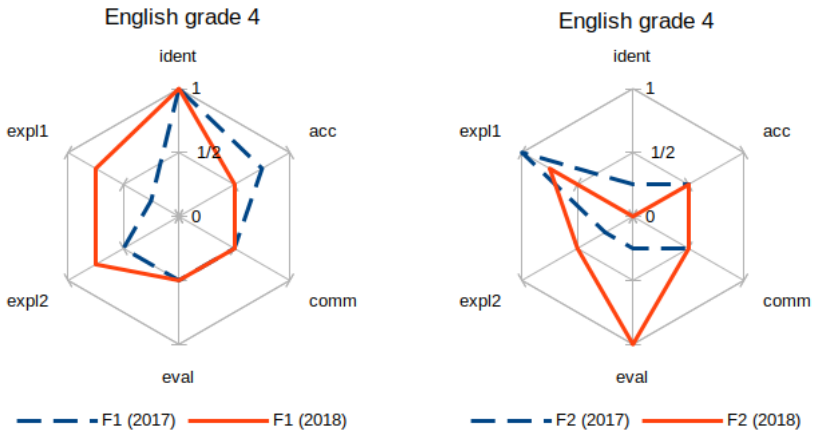


(a) 2017



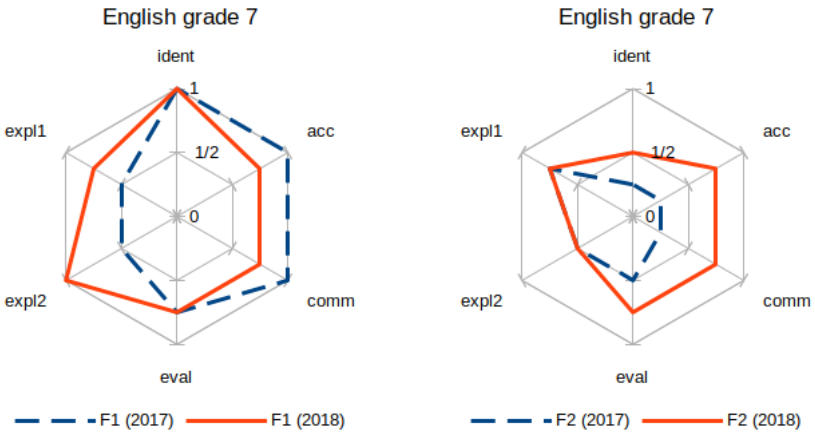
(b) 2018

Figure 6.2: English data coverage by factors



(a) First factor (b) Second factor

Figure 6.3: Proficiency factors for English grade 4



(a) First factor (b) Second factor

Figure 6.4: Proficiency factors for English grade 7

The dashed lines in Figures 6.3 and 6.4 show the degree to which each attribute was manifested by students' performances in 2017, and the solid lines show the results for 2018. Together they show how the 'fuzzy outline' or 'shape' of 'prototypical' performances at each grade change between 2017 and 2018. One could think of these shapes as describing the most representative ways of demonstrating English language proficiency, with respect to the assessment criteria, for students awarded a grade 4, and for students awarded a grade 7. Of course any individual student will display such a prototypical performance

to a greater or lesser degree, and this information is captured in the *extents* of the factors (not shown here). A grade 4 student whose performance participated in the first factor for 2017 to degree 1 would provide a perfect exemplar of this kind of performance. As would be expected, the prototypical grade 7 performances tend to demonstrate each attribute to a higher degree than the grade 4 performances.

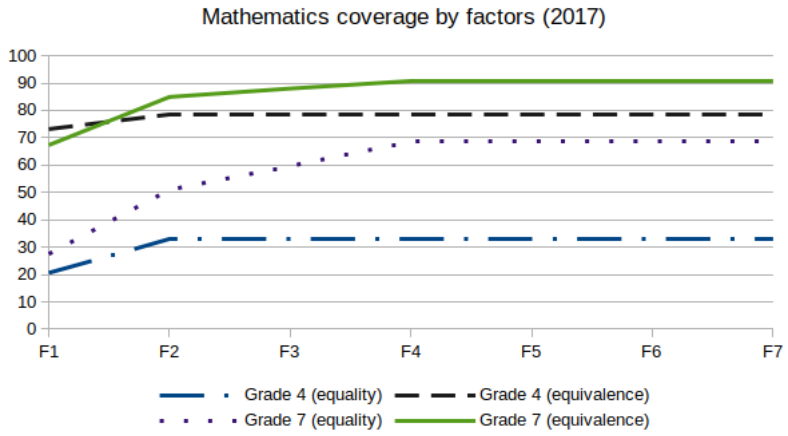
In order to conclude that performance standards were maintained, for this paper, between 2017 and 2018, we would expect to see similar shapes for prototypical performance, at grades 4 and 7, for 2017 and 2018. If the 2018 outline lies outside the 2017 outline, that would indicate an overall improvement in standards, and *vice versa* if the 2018 outline tends to lie inside the 2017 outline.

Figure 6.3a suggests that, with the exception of the 'use language accurately' (acc) attribute, the grade 4 performance standard was actually somewhat better in 2018 than in 2017. Likewise Figure 6.4a indicates an improvement in performance at grade 7 in 2018 on all attributes except 'use language accurately' and 'communicate clearly, effectively and imaginatively' (comm) – although if the second factor is taken into account as well here (Figure 6.4b), those two attributes are also demonstrated to a higher degree in 2018 than in 2017.

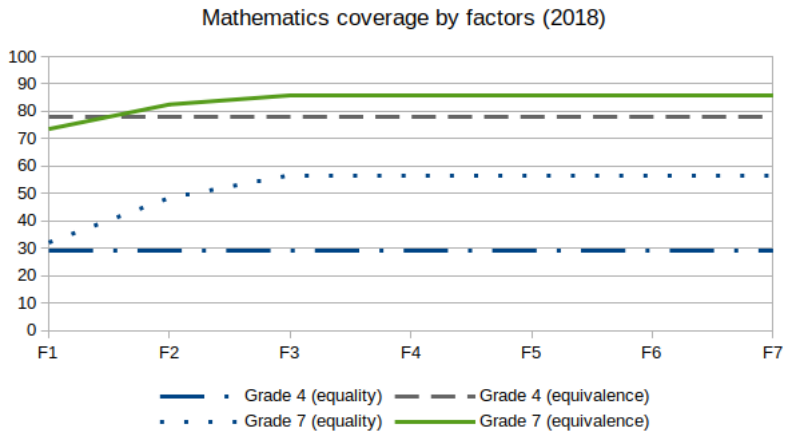
Overall, therefore, it seems not unreasonable to conclude that the approach to grading on this paper has broadly maintained standards, in terms of how a typical grade 4 or grade 7 student demonstrates the aspects of proficiency with respect to which the assessment construct is defined. In fact there has been a slight improvement in performance in several respects. As discussed further in section 6.6.1, there are reasons to expect this because of the operation of the so-called 'sawtooth' effect between 2017 and 2018.

6.5.2 Mathematics

Figure 6.5 shows how coverage rates for the mathematics data matrix increase as more factors are extracted. As for English, rates flatten off after extracting two or three factors. However, it is notable that the explanatory power of the first and second dominant factors is lower for mathematics than for English. The first two factors for mathematics at each of the two key grades are shown in Figures 6.6 and 6.7.



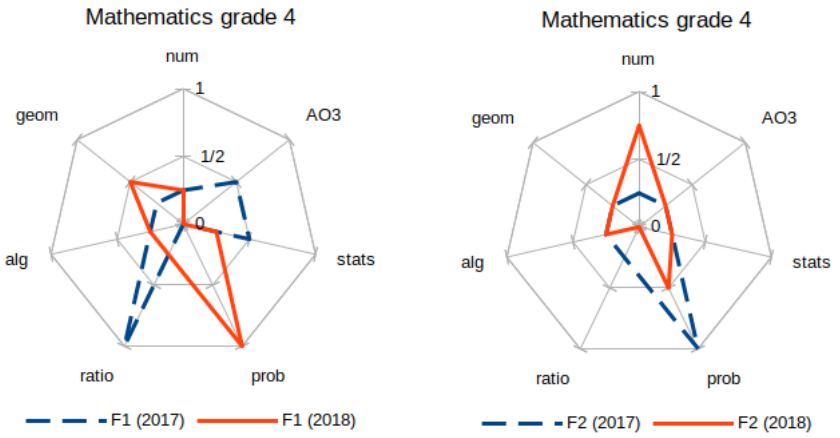
(a) 2017



(b) 2018

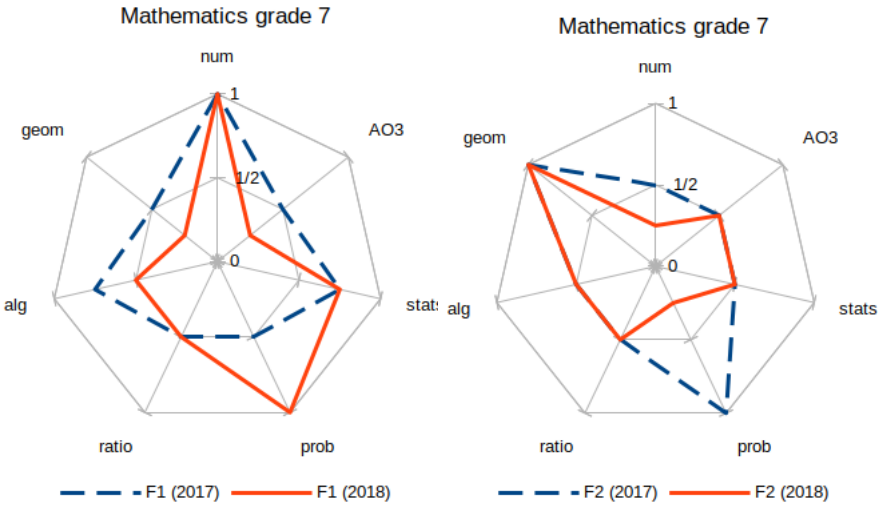
Figure 6.5: Mathematics data coverage by factors

Also by contrast with the results for English, prototypical performances for mathematics



(a) First factor (b) Second factor

Figure 6.6: Proficiency factors for mathematics grade 4



(a) First factor (b) Second factor

Figure 6.7: Proficiency factors for mathematics grade 7

appear qualitatively rather different between years, at both grade 4 and grade 7. Performances of the students awarded grade 4 tended to be relatively stronger with respect to ratio and problem solving in 2017 and with respect to geometry and probability in 2018. At grade 7, the outline for 2018 tends to lie inside the outline for 2017 (with the exception of the probability attribute), suggesting a slightly weaker overall performance in 2018. This is the opposite of what is seen in English, and, as discussed in section 6.6,

somewhat contrary to what might be expected.

6.5.3 Comparison with quantitative principal components analysis

The differences in mathematics response prototypes at grade 4 in particular calls into question the extent to which it is justified to treat mathematics proficiency (with respect to the specific construct of GCSE mathematics, at least) as a homogeneous property, such that 'gaining more marks' is always equivalent to 'demonstrating more of the property'. For example, an item with a maximum mark of 2 may be scored such that candidates get 1 mark for selecting the correct method to perform a calculation, and 1 mark for performing the calculation correctly using the method chosen (whether or not that method was correct). A candidate who scores 1 mark on the question may therefore be demonstrating that they understand the conceptual basis of the problem (they can select the method, if if then make a computational error on following through); or they may be demonstrating that they can do a sum, even if they don't know why a doing a particular sum answers a more substantive question. Arguably, therefore, the step from 0 to 1 marks on this question may demonstrate a difference in kind, just as much as a difference in degree, of mathematical proficiency.

If proficiency is thought of as a quantitative property, then a proxy for homogeneity is the dimensionality of the observed data. A standard approach for examining this would be to perform a principal components analysis. Doing so for the (normalised) English item mark data suggests that the first principal component accounts for 52% of the variance in outcomes, and adding a second component increases this to 68%. For mathematics, on the other hand, the first component accounts for only 29% of the variance, and including the second and third components as well increases coverage to 42% and 55% respectively. So in order to account for more than half of the variance in outcomes, a single latent factor suffices for English, but three separate dimensions are needed for mathematics.

Of course if the datasets are restricted to those candidates whose total marks were in the interval that defines a grade 7 or a grade 4, the variance in outcomes that is accounted for by principal components reduces, to around 40% (for one factor) or 60% (two factors) in the case of English, and 15% (one factor) or 30% (two factors) for mathematics. These are lower than the coverage rates for the qualitative factors calculated in section 6.5.

6.6 Discussion

6.6.1 Summary of findings

This paper has described a new method for studying examination standards over time, and demonstrated its application to high-stakes examinations in English and mathematics.

The results for English broadly suggest, for the paper studied, an overal slight improvement, from 2017 to 2018, in the degree to which the examination responses of candidates graded 4 or 7 tended to demonstrate construct-relevant attributes of proficiency in the subject. This is consistent with what would be predicted from the literature on the so-called sawtooth effect on performance, when a new version of a curriculum-related assessment is introduced (see e.g. Cuff, Meadows, and Black, 2018). The specifications for both GCSE English and mathematics were reformed by the government for first teaching in 2015, and first examination in 2017. Thus the results presented here are for the first and second summer examination series for these subjects.

The expectation is that candidates sitting a new version of an examination for the first time will tend not to perform as well as they would had they been able to sit the examination in a subsequent year, when teachers were more experienced with the specification, and more exemplar material from previous examinations was available. Indeed this is a

key rationale for using comparable outcomes to grade candidates when there is a specification change. Because these candidates may be competing against others who sat the equivalent examinations in subsequent years for jobs or places in further or higher education, there is an ethical argument (Cresswell, 2003) for keeping overall outcomes stable (conditional on the 'comparability of the cohort' control that is chosen, such as prior attainment): even though this may entail accepting a lower performance standard for a given grade than will be acceptable in subsequent years.

Although the results for English are consistent with this expectation, the results for mathematics are less clear cut. As explained in section 6.3.2, grade boundaries are set at grades 4 and 7 based on statistical recommendations derived from applying the comparable outcomes methodology, supplemented with examiners' judgements of qualitative performance standards. In both English and mathematics grade boundaries increased between 2017 and 2018 at the subject level (i.e. aggregated across the two or three papers for the full examination), as would intuitively be expected due to sawtooth. For English the increase at both grade 4 and grade 7 was 4 marks out of a subject total of 160 marks, i.e. an increase of 2.5 percentage points at each grade. For mathematics the increase in boundary marks was 5.4 percentage points at grade 7, and 0.4 percentage points at grade 4.

For the specific paper examined in this study, the boundary marks set increased by 5 percentage points at each of grades 4 and 7 for English. As noted, this appears to have led to candidates being graded such that performance standards were maintained, or in fact slightly improved, between years. For the mathematics paper, the grade 7 boundary increased by 1.25 percentage points from 2017 to 2018, but the grade 4 boundary decreased by 2.5 percentage points. The rationale in the awarding meeting was that the 2018 paper was harder, for students around the grade 4 level of proficiency, than the 2017 paper. Of course the purpose of setting grade boundaries is to allow for changes in

the nature of the tasks and/or approaches to scoring responses to those tasks between examination versions, such that performance standards are maintained. However, in this case (consistent with the research, mentioned in section 6.3, that experts find it hard to make allowance for changes in tasks when judging the quality of responses), we might tentatively conclude that the level of increase at grade 7 was a little generous to candidates, given that the prototypical grade 7 performance in 2018 tended to demonstrate most construct-relevant attributes to a lesser degree than the equivalent performance in 2017.

6.6.2 Broader implications

As noted in section 6.5.3, the differences in response prototypes at grade 4 in particular calls into question the extent to which it is justified to treat mathematics proficiency as a homogeneous property. Treating proficiency in the English assessment as a quantity whose value, for each candidate, is derived from the sum of their item marks, appeared to be a reasonable approximation, for the purpose of assigning students to grades in a way that preserves the qualitative (performance-related) meaning of grades between examination versions. For mathematics, however, treating the information obtained from the assessment as if sum-scores adequately discriminated between learners as to their proficiencies is more problematic (given the specific construction of GCSE mathematics in terms of subject criteria and assessment objectives).

The qualitative factor analyses described in section 6.5 show how grading candidates based on cut-scores has broadly been effective for maintaining performance standards in English. In mathematics, they shows in what respects performances at the key grades differ, in respect of how they have demonstrated the construct-relevant attributes. In that respect they provide more insight that purely quantitative approaches that describe

differences between versions purely as numerical differences (e.g. that 'grade 4 candidates in 2018 had 0.07 logits more ability than grade 4 candidates in 2017', for instance).

These substantive conclusions are clearly warranted only to the extent that the methdology developed in this paper is sound. Some methodological limitations, and suggestions for further exploration, are now considered

6.6.3 Limitations

This study examined the question of comparing standards at the level of an examination component (paper), because that is the relevant locus of comparison with the existing qualitative inputs to the grade awarding process for public examinations. As explained in section 6.4.1, expert examiners make their judgements about performance standards at this level. However, the analysis would need to be extended to aggregate subject level to allow full comparisons over time to be made.

Regarding the factor extraction methodology itself, Bartl, Belohlavek, and Scharaschkin (2018) noted that a feature of the algorithm used in this study is that it may tend to favour the extraction of relatively 'flat' factors (factors whose intents have somewhat uniform attribute profiles) over those that seem more intuitively to match the distribution of truth-degrees in the data matrix generated by the assessment procedure. Bartl and Bělohlávek (2024) have recently explored this question further, and suggest that the reason is actually fundamentally traceable to the way the cardinality (size) of a fuzzy set has traditionally been defined. They propose a new definition of cardinality that generalises the standard so-called sigma-count definition. One way in which the method described in this paper could be extended would be to modify the selection algorithm to incorporate this more general definition, and to explore the extent to which that may result in qualitatively different dominant factors.

6.6.4 Areas for future research

The truth-value approach seems to align more closely to the level-of-response approach to marking used in English than to the rules-based approach in mathematics. It may be that markers have more scope to compensate for differences in the specific items used to assess an attribute when they can decide what level to match the response to, than when they have to follow specific rules about how to assign marks. The rules might work quite poorly as ways of assessing a candidate's fuzzy match to an attribute if the items are not well designed (e.g. if the experts think that two questions are equally difficult—in the sense that any given candidate would get the same result on both of them when the marking rules are applied—and that turns out not to be the case). Moreover there is an element of arbitrariness in the mappings of marks to truth-values described in section 6.4.3, especially when dealing with slightly different numbers of marks per attribute. Exploring the application of fuzzy-relational methods to subjects like mathematics probably entails revisiting the way in which attributes are marked or valued.

Another question that merits further exploration is the weighting of assessment objectives. The idea of weighting an assessment objective is to give it relatively more importance or emphasis as a factor that determines the grade that a student is awarded. So in the GCSE English paper considered in this study, the assessment objective 'communicate clearly, effectively, and imaginatively' is weighted, or emphasised, more than the objective 'use language accurately'. This is expressed in the specification for the assessment by stipulating that the former assessment objective is weighted at 30% (of the available marks), and the latter at 20%.

Of course the notion of 'a percentage weighting' for an attribute or part of an assessment construct presupposes an underlying quantitative structure for the construct. For

examinations marked using rubrics of the kind discussed in section 6.4.1, it requires that either 'a mark' is an exchangeable unit of measurement between items, whose marginal value is constant, irrespective of for what aspect of a response it was awarded; or at least that the relative values of marks for different items or subsets of the examination can be expressed as ratios on the assumption of an 'underlying' interval scale (see Cresswell, 1987, for a development of this idea).

Whether or not such an assumption is tenable, however, it seems highly unlikely that, in the context of awarding, examiners judging the qualitative value of perfomances make these precise quantitative distinctions. For instance the author has heard an awarding committee, in referring to assessment objectives, state that 'remember that for this paper, the assessment objectives are weighted AO3 > AO2 > AO1', consonant with an ordinal, not a quantitative, understanding of relative importance.

While assessment attributes have not been distinguished with respect to weight in the analysis in this paper, there are least two areas in which further investigation would be valuable. The first, building on this ordinal conception of weights, would be to attempt to construct separate 'intended' and 'actual' concept lattices for the construct, analogous to an approach which is now being explored in the application of formal concept analysis to competence-based knowledge-space theory: see Huang et al. (2023). This also opens up a possible further extension towards moving away from marking altogether, as a means-to-an-end for grading, in cases where representing the construct of proficiency (of the kind relevant to the domain in question) is not well served by an approximate numerical model.

An alternative forward would be to examine the potential to use so-called 'truth-stressing hedges' in setting up the fuzzy relations between candidates and assessment attributes. As noted in Krupka (2008), hedges are logical operators that can be used as parameters to select 'important attributes' and 'important objects'.

Finally, as noted by Scharaschkin (2024), a rather deep question that warrants further examination is how truth values are determined. In the case of tests consisting of dichotomous items this is staightforward, but when judges are required as part of the measurement procedure (to apply the scoring rubrics), different judges may give different truth values. Scharaschkin suggested that drawing on *rough fuzzy logic* (Bazan, Skowron, and Swiniarski, 2006) to help appraise what would count as a reasonable or acceptable truth value in these circumstances might be helpful.

There are also a numbers of issues that should be explored at the technical, operational level, e.g. different options for realising logical operations than the Łukasiewicz operations described in Appendix D.

6.6.5 Conclusion

This paper shows that a non-quantitative approach to modelling learners' proficiencies is possible and worth exploring further. Its application to the problem of maintaining examination standards over time could provide an alternative source of evidence to complement existing approaches. More generally, representing manifest proficiency levels as partially ordered concepts, rather than locations on a totally-ordered continuum, may be an alternative to quantitative measurement approaches when the construct of interest is not well described as a quantity (or vector of quantities).

7 Discussion

This study has considered the question of measuring learners' proficiencies in educational domains. It has considered the question of if, and when, it is reasonable to consider the phenomenon of proficiency as being a quantity that can be measured using an assessment to produce a numerical score which summarises each learner's proficiency in the domain in question. It has suggested that thinking of proficiencies as quantities is not always reasonable, and has proposed an alternative approach to educational measurement in such cases. It applied that approach to the substantive question of appraising performance standards in high-stakes examinations. The discussion below summarises some key themes that run through the three papers that make up this integrated dissertation, and suggests areas for further development and research.

7.1 Proficiency in an educational domain, assessment constructs, and quantitative structure

Chapter 2 showed that summative educational assessment procedures, such as examinations used to underwrite the award of qualifications attesting to a learner's level of proficiency, are designed to elicit information about *assessment constructs*, which, in the UK at least, have been conceptualised in two different ways. On the one hand, approaches

to educational assessment that derive from psychological testing and measurement tend to regard constructs as *latent cognitive or mental properties* of the people being assessed. On the other hand, curriculum-related educational assessment is often concerned with constructs thought of as *specifications* of the particular combination of knowledge, skills, and understanding that characterise (good) attainment in a domain.

Assessing learners with respect to a construct such as *GCSE English language*, under the latter conceptualisation of constructs, means using the data generated by an assessment procedure to classify them, such that learners who demonstrated similar kinds of knowledge, skills, and understanding (according to the specification of the domain) are placed in similar categories. The categories are usually ordered, and labelled by numbers or letters that show the ordering. The fewer the number of steps between categories, the more similar they are in terms of their associated qualities of attainment. So for a totally-ordered collection of categories labelled 9, 8, 7, ..., 2, 1 (as the results of GCSE examinations are), a learner classed as '6' has demonstrated better attainment than one classed as '5', but is more similar to that learner than to one demonstrating much poorer performance who was classed as '2'.

Under the latent trait conceptualisation of constructs, assessing learners with respect to GCSE English language means using the data generated by an assessment procedure to estimate values of a hypothesised proficiency-at-English property, such that learners with similar levels of proficiency are given similar values.

The way in which these two ways of thinking about constructs in educational assessment are linked is though the mathematical *structure* that is hypothesised for this phenomenon of proficiency. At any given point in time, a learner may 'have' a particular configuration of cognitive and mental resources, such that she demonstrates (more or less fuzzily) a particular collection of construct-relevant attributes, as elicited by means of an educational assessment procedure. (Chapter 2 defines *attributes* as the specific aspects, features, or

characteristics of learners' responses to tasks about which information is collected in an assessment, and discusses the question of the coherence of separate attributes that make up a construct.) Thus, as developed further in Chapters 5 amd 6, the results of summative educational assessments are in fact *propositions about learners*, that declare which *possible proficiency state* (which of the states that are *discriminable* by the procedure) they manifested when assessed. The *structure* of the phenomenon of *proficiency*, for the domain in question, is then a description of the relations between these possible states. Chapter 5 suggests that this is what van Fraassen, in his account of measurement, would call the *data model* for the phenomenon.

Psychometrics normally takes the possible states of proficiency as being points on a linear continuum. Proficiency is assumed to be a phenomenon that has *quantitative structure*, that is, that can validly be represented mathematically as a mapping from learners to a structure (a complete ordered field) that is isomorphic to the real numbers (or to vectors of real numbers, in the multidimensional case). Making this assumption, that each learner's proficiency can be modelled as a point on the real continuum, allows psychometricians to apply statistical models and techniques that rely on the mathematical properties of the real number system, such as factor analysis, item response theory, and maximum likelihood estimation, to obtain numerical values that are taken to be *measurements* of learners' proficiencies in the educational domain in question.

The normal practice in high-stakes examinations in the UK is to use assessment data (i.e., the collection of item-scores for each candidate, in the case of a test or examination) to classify learners by means of summing scores across items. As noted in Chapter 2, this so-called 'compensatory' approach to the assessment of constructs, whereby students are classified as equally proficient if they have the same total mark, even if they display different profiles of strengths and weaknesses with respect to attributes that define the construct, yields the same level of proficiency for each student as would be generated if

a specific psychometric model—the Rasch model—is assumed, because total score is a sufficient statistic for the estimation of the proficiency parameter in that model.

But taking either a compensatory approach or fitting a Rasch model requires the assumption of quantitative structure. For the Rasch approach, this is because the parameters to be estimated are assumed to be real numbers; for compensatory classification because it requires that attributes with ordered values (item responses) can be manipulated as if they were numbers, with the intervals between any two categories (each such interval constituting a mark) having a constant meaning as an equal amount of an underlying quantity.

The assumption of quantitative structure can, in some circumstances, be tested empirically. Chapter 4 used (a stochastic version of) the theory of *conjoint measurement* (itself a special case of the *representational theory of measurement*, discussed further below) to do this for some high-stakes summative assessments in England. The results suggested that requirements for quantitative structure were generally not met, although the extent of deviation varied by subject. It was concluded that a key area for further research was to compare alternative approaches to measurement, that do not assume the assessment construct is quantitative, with existing approaches that are generally (if uncritically) accepted by teachers and students, noting that the trade-offs between validity and pragmatic useability are likely to vary depending on the assessment domain.

7.2 Theories of measurement

Chapter 5 addressed the issue of measuring non-quantitative constructs. It reviewed the three theoretical frameworks that have received the most attention in the literature relating to the conceptual foundations of educational measurement, namely the representational theory of measurement, Rasch measurement theory, and the realist theory of

measurement as the discovery of ratios. It noted that, irrespective of their claims to be adequate as theories of measurement more generally, these approaches can at a minimum provide useful ways of exploring questions that are relevant to engaging with the key concerns, in educational assessment, of *classifying* students consistently and coherently with respect to the ways in which they display proficiency in a domain, and (the equivalent question, seen from another viewpoint) of *discriminating between* students consistently and coherently with respect to the ways in which they demonstrate proficiency.

For instance, by regarding the representational theory of measurement as studying mappings from qualitative relational structures to a numerical relational structure, one can use results from the theory (such as the particular set of theorems that constitute the theory of *conjoint measurement*) to test the extent to which abstract concepts, captured or defined as qualitative relational structures, can be represented numerically. This is especially helpful in educational assessment because, as argued in Chapter 5, educational assessment procedures can, in general, be modelled as qualitative relational structures. They are in fact, it is suggested, *fuzzy relational systems*, that generate matrices of truth-values for propositions of the form 'learner l has demonstrated construct-relevant attribute a'. As the results of Chapter 4 showed, sometimes it is reasonable to approximate these structures by numerical quantities. This was the case for the economics test evaluated in that chapter, but not for the physics test. The latter could not reasonably be taken as providing estimates of the values of a quantitatively-structured latent proficiency variable, even when the test items were pruned in order to force a better fit to a Rasch model.

Rasch measurement theory itself is clearly a helpful and practical approach in cases where the data model for the assessment construct can be taken as quantitative. Indeed, Rasch theorists would argue that in cases where assessment results do not fit the quantitative model, the construct itself needs to be changed, because as specified it is not susceptible

to quantitative measurement: it does not reliably allow each learner to be located, in respect of their proficiency in the subject being assessed, at a point on an equal-interval scale.

It may be the case, however, that re-defining or constraining the construct such that a better model fit is obtained compromises construct validity: the assessors' understanding of what the key attributes of proficiency actually are, for the domain in question, and how relatively better, worse, or just different, states of proficiency would present, with respect to those attributes. In such cases the choice would seem to be either to abandon the idea of measuring the construct at all, or to abandon the restriction of the idea of 'measurement' to locating measurands within solely quantitative mathematical structures. Chapter 5 examined what this latter option could look like for educational assessment, building on van Fraassen's (2008) conception of measurement as, in general, *location in a logical space*.

7.3 Educational assessments as fuzzy relational systems

It is argued in Chapter 5 that a fundamental aspect of educational assessment is a notion of *betterness* between learners as to the different levels, states, or configurations of their abilities. But in general one should not think of this betterness relation as being a total order relation (a ranking)—even though the special case of a total order may sometimes be appropriate. In many cases, however, a partial order is more realistic. Given the way the assessment construct is specified, it may only be possible to infer, for some pairs of learners, that their proficiency states, or levels, are non-comparable (qualitatively different), not that one is better than the other. This does not preclude the possibility of grouping learners together into 'coarser' totally-ordered classes (such as examination grades), such that it can be inferred that those who 'pass' are more proficient than those

who 'fail', for instance. It just means that, within the 'pass' category, there may be some learners whose proficiencies, although both of at least a 'pass' level, may be different and non-comparable.

Such partially-ordered structures arise directly from the data matrices produced by tests and examinations, in which the (i, j)-entry records the level of learner i's response to item j. In fact, as shown in Chapter 5, associated with every assessment procedure that produces such tabulated data as its outcomes is a finite (although often very large) number of possible ways of demonstrating proficiency between which it can discriminate. Each of the proficiency-states that can be discriminated by the measurement procedure (the assessment) can be described *extensively*, by showing the specific performances that exemplify it, or *intensively*, by listing the specific attributes that characterise it. The mathematical duality between *grouping-together* (showing extensively), and *discriminating-between* (describing intensively) is considered further in the remarks on areas for further research below.

In the terminology of mathematical order theory, these proficiency-states are precisely the *formal concepts* associated with the assessment. They form a partial order (in fact, a lattice, called the *concept lattice*) that can be represented as a network, or directed graph, with a 'bottom' element (no proficiency in the domain), and a 'top' element (mastery of the domain). Any two states a and b in the network are connected ($a \to b$) if b is equal or better than a in the partial order; they are not connected if b is simply different from, but not directly comparable with, a. There are generally multiple paths between the bottom and the top of the network (from novice to mastery, as it were), where a path is a sequence of connected proficiency-states. Chapter 2 showed that for the special case of assessments that yield biordered (Guttman) data, the observable intermediate levels of proficiency, between 'no evidence of attainment' and 'mastery', are totally ordered: in this idealised case the concept lattice is simply a ranking. However, the more outcomes

deviate from perfect biorders, the more complex and larger the concept lattice becomes.

It is necessary in educational assessment to deal with inexact, intersubjectively defined concepts, because the very definition of 'what counts' as important in the curriculum, as well as what 'good' or 'proficient' attainment looks like in respect of what counts, are ultimately value judgements. Moreover the procedures for ascribing value to learners' responses to assessment tasks (for example, marking essays according the a rubric as discussed in Chapter 2) often depend on understandings of qualities of performances that are constructed by, and shared between, expert human assessors. Chapter 5 proposes that this consideration may be incorporated into the analysis of assessments as relational systems by extending the notion of formal concepts to *fuzzy formal concepts*. This entails thinking of the outcomes of assessments such as examinations as tables or matrices of *truth degrees*, rather than matrices of numbers.

Specifically, if the (i, j)-entry in such a matrix is v, that means that the proposition 'learner i demonstrated attribute j' is true to degree v. In the case where attributes are Boolean or 'crisp' (assessed by dichotomously scored items), this reduces to an item score of 1 meaning it is true that 'learner i demonstrated attribute j': i.e., they got the answer to the item eliciting information about attribute j correct. An item score of 0 means that the proposition is not true (they responded incorrectly to the item). More generally, it may be true to a certain degree, between (completely) false and (completely) true, that a learner demonstrates an attribute in an assessment. Chapter 6 showed how item response data derived from applying scoring rubrics to examinations in English and mathematics could be viewed as providing truth-values of propositions about candidates having demonstrated construct-relevant attributes.

Chapter 5 noted the importance of distinguishing between epistemic *uncertainty* (the lack of precision that arises from incomplete or poor information), and ontological *vagueness* (the inherent fuzziness, or necessary inexactness, of concepts like 'proficiency' in a certain

domain). Probability theory is a tool that can be used to study the former, whereas fuzzy logic is a tool that can be used to study the latter. As touched on further below, quantitative approaches to educational assessment tend to conflate these two aspects of inexactness as 'error' attaching to the numerical estimates of 'how much' proficiency any given student has.

While truth-values (or degrees of truth) are often represented using numerical labels for convenience (as was done in Chapter 6), it is important to remember that they are not quantities but elements of an ordered structure (a so-called complete residuated lattice, as summarised in Appendix D), and cannot be manipulated (added, multiplied, etc.) as if they were numbers. Rather, they are combined using the logical operations defined on the truth-value structure.

When the outcomes of assessments are not crisply dichotomous, the concept lattice is itself fuzzy. Learners and attributes belong to concepts with degrees of truth, rather than crisply.

7.4 Extending educational measurement to non-quantitative constructs: theory and practice

Chapter 5 argued that applying this lens of order theory and fuzzy logic to educational assessment provides a basis for an approach to educational measurement that extends the purely quantitative view of measurement, from location of a measurand at a point in a particular *numerical structure* (the real continuum), to location in a *logical space*: the view of measurement taken by Bas van Fraassen. By doing so it aimed to move the debate on from simply critiquing approaches, such as psychometrics, on the grounds that (as shown in Chapter 4) the constructs they study are often not appropriately

regarded as quantities (in Joel Michell's rather harsh words (Michell, 2012, p. 255), that 'psychometrics is built on a myth'). It noted that quantitative approaches may indeed be appropriate and pragmatically useful in many cases: but that if, indeed, the complex phenomena studied in educational assessment really are different in nature from pure quantities, then a conception of measurement that goes beyond quantification is worth considering.

van Fraassen (2008, p. 172) noted that measuring procedures in classical and quantum physics are all 'cases of grading, in a generalized sense: they serve to classify certain items as in a certain respect greater, less, or equal. But ... this does not establish that the scale must be the real number continuum, nor even that the order is linear. The range may be an algebra, a lattice, or even more rudimentary, a poset.'

This dissertation has suggested that fuzzy concept lattices, as described above, are appropriate mathematical objects to use for the representation of (i.e. measurement of, in this wider sense of not-necessarily-quantitative measurement) learners' proficiencies in educational domains. Chapter 5 emphasised that, in the context of educational testing, the proficiencies being studied are proficiencies or competencies *with respect to* a specified domain, such as 'high school Chemistry' or 'A level French'. What 'good performance'—and hence what would count as evidence of better or worse levels, or states, or configurations of learners' proficiencies—is always subject to a prevailing understanding or agreeement as to which potential aspects of the domain are chosen as relevant for discrimination between learners' performances as to their quality. The criteria for creditworthiness of candidates' responses to tasks in an assessment can be regarded as the selective relevant depiction of the phenomenon of interest—domain-specific proficiency—by those members of the competent authority who design, administer, and grade the tests. The concept lattice for a test locates each test-taker, and each discriminable type of response to the test, in a partially-ordered structure, with degrees of fuzziness if necessary.

This view shifts the notion of measuring proficiency, from finding out or estimating an *amount of* something (a quantity), to summarising a type, or *kind of*, something (expressed either—extensively—as prototypical *exemplars*, or—intensively—as stereotypical *descriptors*).

A practical issue with using concept lattices to represent the information collected in an assessment is that they rapidly become very large, once any more than a small number of learners and attributes are involved. Chapter 6 investigated a method of reducing the complexity of (all or a part of) a concept lattice by *factorising* the matrix of fuzzy truth-values it represents, into a product $L \circ F$ of *loadings* and *factors*. Here, expressing a matrix as a product of factors means using the logical operations by which truth values may be combined to form the product, rather than (but analogously to) using the usual numerical operations of multiplication and addition of matrix entries.

Chapter 6 showed that this process could be compared to the problem, fundamental to many quantitative data reduction methods, of approximating a matrix of real numbers by one of lower rank. Eckart and Young's solution to this problem was published in the first volume of *Psychometrika* in 1936. It is argued in Chapter 6 that Eckart and Young's theorem suggests that, when working within an assumption of quantitative structure, the key mathematical construct derived from the observed data that provides the mechanism for generating *measurements* of students' proficiencies is the *eigenspace decomposition*—the eigenvectors and their relative dominance—of the symmetric matrix of item covariances (or comparisons, in the case of Rasch measurement).

In quantitative analyses of test data, the covariance matrix Σ provides the information used to measure students' proficiencies. The eigenspaces of Σ can be thought of as different fundamental proficiency states, such that each students' measured proficiency is a combination of these states. The associated eigenvalues indicate the relative importance of each state, so that if the first eigenvalue is large by comparison with the others, one

can regard proficiency measures as being captured to a good degree of accuracy by each student's projection onto this space. (As noted below, this way of thinking about quantitative measurement is actually analogous to the mathematical formalism used to represent the measurement of a physical observable in quantum mechanics.)

Chapter 6 argues that a natural extension of this underpinning quantitative foundation, to the case where constructs are thought of as composed of attributes whose truth values may be fuzzy, replaces the eigenspace decomposition of Σ with a decomposition of the matrix of truth values M in terms of (fuzzy) formal concepts. The analogue of dominant eigenspaces is prototypical formal concepts.

In the unidimensional case, a learner's quantitative proficiency is estimated as a *score* (a position on the dominant eigenline). A learner's qualitative proficiency is estimated as a a *degree of membership* of a *particular type of performance* that is described in terms of construct-relevant attributes.

Chapter 6 applied this approach to a problem in assessment that is at the heart of the public examination system in the UK, namely the maintenance of qualification standards over time. To award a qualification is to make a promise about the standard of proficiency demonstrated by the qualification holder. The possible standards that can be attested for GCSE and A level qualifications in England are the grades reported for each. Teachers and general users expect grades (especially key grades such as grade A for A levels, and grade 4 for GCSEs) to have the same meaning irrespective of which particular version of an examination a student took: that is, that students who demonstrate that they know, understand, and can do similar things in response to the assessment tasks will be awarded the same grade.

Part of the warrant for making such promises is therefore expert judgement of whether the kinds of responses associated with key grades for different examination versions are

qualitatively equivalent. But concerns about the robustness of these judgements means that in practice grading decisions are dominated by statistical evidence, derived from sources other than the examinations themselves. Chapter 6 proposes that prototypical performances (specific formal concepts), obtained from the concept lattices for students at key grades, can be used as a source of evidence for evaluating whether performance standards (i.e. the extent to which students, classified as being at a certain grade, tend to demonstrate particular construct-relevant attributes) has changed over time.

The results suggested that grading candidates using the officially specified approach to setting cut-scores appeared to be broadly effective for maintaining performance standards at key grades in GCSE English between 2017 and 2018. The results were less clear cut for mathematics, but the analyses indicated the respects in which students' performances differed at key grades between 2017 and 2018 in a way that is hard to elicit using a purely quantitative approach.

Clearly there is more to do to test this methodology, as well as looking at further substantive applications, which brings us to the question of areas for further investigation and exploration suggested by this study.

7.5 Areas for further research

7.5.1 Specific topics relating to the papers presented

Chapter 4 considered the question of using a stochastic version of conjoint measurement theory to examine the extent to which it was reasonable to view some educational assessments as providing information about latent variables that could reasonably be supposed to be (approximately) quantitatively structured. The methods developed to date have focused on tests made up of dichotomous items. While it would be useful to explore a

larger collection of real assessment data from dichotomously-scored tests, a clear area for further research is the extension of the approach to tests with polytomous items. Some preliminary steps were sketched in Chapter 4, but more investigation is needed to develop a rigorous approach to testing the conjoint measurement axioms in the polytomous case.

The practical importance of such work is perhaps the level of assurance it gives to practitioners such as examination boards or test providers as to when the pragmatic use of the standard repertoire of quantitative techniques in test development and analysis is reasonable.

It was suggested in Chapter 4 that when item responses are categorised using levels-of-response (pattern-matching style) mark schemes (scoring rubrics), then classifying test-takers by means of their total scores may in fact be a good practical alternative to using more elaborate order-theoretic approaches to aggregation, provided that the issue of the quantitatively-expressed intended *weighting* of different parts of the test can be satisfactorily addressed. Chapter 6 came to a similar conclusion based on the analysis of the qualitative performance standards exemplified by students in GCSE English examinations. This question of weighting—its conceptual meaning (is it the relative importance of something? Is it the relative amount of something?), and its practical implementation—needs further consideration. Chapter 4 noted that the literature on the application of partial-order methods in other contexts, such as the analysis of multidimensional deprivation in international development, may provide some useful suggestions. Chapter 6 also recommended two possible avenues for exploration. One approach could be to ivestigate the way the notion of 'importance' can be captured in fuzzy logic, using logical operators known as hedges. A more comprehensive approach would be to attempt to construct separate 'intended' and 'actual' concept-lattices for the assessment construct. This has recently started to be explored in the application of formal concept analysis to

competence-based knowledge space theory (Huang et al., 2023).

As noted in Chapter 5, various different order-theoretic approaches to educational assessment that have been investigated since Guttman's work on scalogram analysis in the 1940s are starting to become more unified with respect to their mathematical underpinnings. Structures such as concept lattices and knowledge spaces can now be studied as variations within the overall framework of order theory (or more generally, category theory), just as different factor analytic and item response theory models can be regarded as different types of quantitative latent variable models. There is certainly more to be done to elucidate further the connections between non-quantitative approaches.

An area that perhaps provides the most significant challenge to the research reported in Chapters 5 and 6 is the question of 'where do the truth values come from?'. When judges are needed as part of the assessment or measurement procedure, different judges may ascribe different truth values to the proposition 'this candidate has demonstrated attribute/criterion x', so what should count as a reasonable or acceptable value? If a student produces some text in response to a prompt designed to elicit information about their proficiency with respect to the attribute 'writes persuasively', it presumably doesn't come equipped with an inherent, objective valuation waiting to be discovered (more or less accurately) by a judge. Rather, it only demonstrates the attribute of persuasive writing to the extent that different observers perceive it to. Just as the official approach to reviewing marking of public examinations in England no longer recognises a concept of 'true mark' (see §2.5.4.1), so truth values derived from an assessment procedure should probably be regarded as themselves somewhat fuzzy or vague.

Chapter 5 suggests that methods of *rough fuzzy logic* may help to provide a conceptual framework for this. In any case, there is a substantial literature on the mathematics of what might be termed 'higher-order' fuzziness (it is a topic of interest in machine learning, for instance), and there is scope for considerably more exploration of these

ideas in educational assessment. A further area for development is the relation of these ideas to traditional conceptions of measurement error as randomly-distributed noise or uncertainty added to a true value. For instance, might the 'definitional uncertainty' aspect of Black and Newton (2016)'s suggested taxonomy of reasons for disagreement between raters in applying marking rubrics be modelled as fuzziness, while the 'procedural/attentional error' component could be treated as error in this more usual sense?

The application of fuzzy relational methods to analysing examination standards in Chapter 6 can be extended in many ways. Exploring examiners' views on the usefulness and credibility of fuzzy prototypes of grading standards is probably best carried out at the individual paper (component) level within an examination (the kind of analysis that was done in the paper), because that is the locus for their qualitative judgements that feed into the standard setting process. But it would be illuminating to consider whole-subject level representations of the prototypical standards for each grade, which could be done by coding all the items across all papers. Although judges do not look at the complete work of candidates across all components of the examination, it is at the subject level that the grading standard officially resides, as it were.

Subject-level investigations might also benefit from a more constant allocation of marks to attributes in subjects such as mathematics, which might alleviate one of the challenges with the method in the form described in Chapter 6, namely converting marks to truth values. In fact more work across more subjects is needed to investigate whether the idea of using item marks as proxies for truth values really works properly for subjects that use rules-based rather than pattern-matching style marking rubrics. As discussed in Chapter 2, the latter correspond to the prototype theory of concept formation, whereas the former correspond to the classical theory. The fuzzy logical approach to concept analysis fits more naturally with a prototypical approach. If, for instance, mathematics subject experts really do come to an intersubjective understanding of what relatively

better or worse proficiency means quite differently from the way English experts do, then further development of the method (for example by taking a more atomistic approach to what count as the fundamental construct-relevant attributes) may be necessary.

Finally, from a technical methodological perspective, more detailed work is needed on the possible impact of changes to the factorisation algorithm itself and to the particular choice of truth-value composition function, beyond the Łukasiewicz version used for the analyses reported.

7.5.2 More general questions for exploration

Formative and adaptive applications The case studies explored in this dissertation were summative assessments, but the general approach of modelling proficiency in a domain as a partial order (directed graph or network of paths) between a bottom (novice) state and a top (mastery) state seems attractive as an approach to linking learning and assesment. Whether or not one regards the states connected, via the network, to a given state s as a kind of Vygotskian zone of proximal development for learners assessed as being at proficiency state s, there are clearly applications to formative assessment (and indeed the main commercial application of knowledge state theory, the ALEKS system mentioned in Chapter 2, aims at being an integrated learning and assessment system). Item response theory-based adaptive tests may offer learners, depending on their estimated state (amount of) proficiency after the administration of a sequence of items, harder or easier items notionally located at points on a single line between 'easy' and 'hard' (equivalently—dually—between 'not proficient' and 'proficient' states). Adaptive tests based on representations of domain proficiency that are partially ordered, rather than isomorphic to the real continuum, may be more effective in routing learners to the aspects of the assessment construct they would most benefit from engaging with, to

achieve mastery. One could also envisage such tests using artifical intelligence to provide a classification, score, or truth value, to learners' responses to an item testing any given attribute: in which case the items need not be restricted to 'correct/incorrect'-style questions.

Vocational and technical assessments Although the papers collected in this dissertation focused on the assessment in academic educational contexts, the approaches investigated seem suited in principle to application to vocational and technical assessment as well. Sometimes the assessment of technical or vocational *competence* is seen as fundamentally different from the assessment of academic *proficiency*, because it is felt to be more 'purely' criterion-referenced in nature. But as shown in the preceding papers, academic assessment can also be seen as concerned with the extent to which learners are able to demonstrate criteria (construct-relevant attributes).

In England, the regulator of awarding organisations refers to a 'CASLO' model of assessment for vocational qualifications: assessment to Confirm the Aquisition of Specified Learning Outcomes. A learning outcome (an attribute) is intended to be a particular aspect of competence, and the idea is that a qualification attests to the learner having demonstrated all the specified outcomes—suppose there are n of them—for the domain in question. That is, the qualification is awarded just in case the conjunction of the propositions 'the learner has demonstrated the acquistion of learning outcome i' is true, for all $i \in \{1, 2, \ldots, n\}$. In practice, though, are these propositions really crisp demarcations between 'demonstrated' and 'did not demonstrate'? It would be interesting to explore this paradigm using fuzzy relational methods.

Measurement formalisms in physics and the human sciences Chapter 6 considered the connection between the quantitative measurement of a latent variable using a test,

and the eigenspace decomposition of a symmetric matrix (the item covariance or comparison matrix) for the test. Such matrices can also be viewed as linear operators on the real vector space of all test items. Before the administration of the test, we don't know how the items will perform. We can represent the states of the items, at this point in time, as a collection of independent (orthogonal) vectors of length n (if we have n items in total), using so-called 'one-hot encoding'. So item 1 is represented as the vector $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, item 2 as $\begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$, item n as $\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$. After administering the measurement procedure (the test), with covariance matrix Σ, the state of item y is transformed to Σy: a vector that we can think of as positioning or 'embedding' the item in a space in a way that captures its similarity to the other items. We then measure each item (e.g. calculate its difficulty) by projecting its state-vector onto the eigenline of Σ that has the largest eigenvalue. (We can do the same thing, dually, with students. Instead of using the operator $\Sigma = Z^T Z$, where Z is the matrix of item z-scores, we use the symmetric operator ZZ^T. These two operators have the same eigenvalues.)

It is suggestive that, in quantum mechanics, *observables* (physical quantities, such as position, momentum, electron spin, etc.) are represented by linear operators that are generalisations of symmetric matrices (so-called self-adjoint operators on Hilbert spaces. Symmetric matrices on real vector spaces are a special case of these). And *measurement* of observables means finding the eigenvalues of these operators (Susskind and Friedman, 2014; Isham, 1995).

There is much to be explored here, in particular, concerning how this conception of measurement in quantum mechanics is, as noted by Ellerman (2022), fundamentally about *discriminating between* objects (describing them intensively), its duality with *grouping together* objects (showing them extensively), as well as the relationship of the quantum formalism to formal concept analysis when the set of possible values for states is ex-

tended from numbers to truth-values (Bradley, 2020). Doing so could shed further light on the possibilities presented by considering educational assessment procedures as fuzzy relational systems.

Also, thinking about *representing* as *embedding in a vector space*—whether the embedding is based on item-scores or on other attributes that describe an artefact in a multidimensional 'feature space'—brings us finally to the question of language models and the use of machine learning and artificial intelligence in educational assessment.

Artificial intelligence in assessment . At the time of writing this is a very rapidly growing field, with the results of many investigations as yet unpublished. *Large language models* such as GPT4, Mistral, or Claude, are now being applied to marking, scoring, and providing feedback to students taking assessments.

Consider the case where the artefacts produced by candidates in response to tasks are written text. Applying a mark-scheme M to a candidate's response X to obtain a categorisation of the quality of X, or level of proficiency displayed by X, is in some sense a matter of investigating the extent to which X matches M, when the attributes of M and X are expressed in some common way. Large language models express the attributes of linguistic artefacts such as M and X as collections of vectors. Each word in the text is represented by its *embedding* into a vector space over the real numbers, obtained, for example, from algorithms such as word2vec (Mikolov et al., 2013) or GloVE (Pennington, Socher, and Manning, 2014). The positions of the vectors in the space tend to reflect semantic similarities between words.

Linguistically, the process of applying M to X can be expressed as a text, of length n words, such as 'Given the mark scheme *abcd*, mark the response *wxyz*'. A language

model represents this text as an $n \times m$ matrix T, in which each row is a word[1], and each column is a dimension of the vector space used for the word-embedding (i.e., T_{ij} is the score of word i on dimension j).

The building blocks of the *attention mechanism* used in neural networks such as GPT4 are three matrices that are derived from T, known as the *query*, *key*, and *value* matrices (Q, K, and V) respectively (Vaswani et al., 2017). Inuitively, Q and K together correspond to the mark scheme M, and V corresponds to the response X. Q is used to flag 'what are the important parts of X?', and K asks 'where are these aspects located in X?'. Each of these matrices has shape $n \times h$, where h is the number of 'heads' of attention: the number of different foci, or ways of paying attention to the text.

These matrices are created from learners' text responses by multiplying T by $m \times h$-matrices of weights (e.g. $Q = TW_Q$) that are learned by training the neural network on some examples of marked responses. Then QK^T is an $n \times n$ matrix whose entries encode 'the importance of Q with respect to all the words in the text', or 'how the words in the response relate to each other given Q'. Finally, a collection of *attention scores* can be created as $f\left((QK^T)V\right)$, where f is a function such as softmax, providing a score for each of the n words in the text on the h foci of attention. In generative applications of LLMs, these scores can be used to suggest the next word to output.

Clearly, h bears a resemblance to the number of construct-relevant assessement criteria. In the case $h = 1$, there is a single focus of attention, analogously to there being a single latent proficiency variable in a unidimensional psychometric model. In this case the attention mechanism produces a vector of n scores for each candidate. An area for further research is the relationship between these scores, and other ways of measuring

[1] Some more recent language models in fact encode sequences of more than one word into a single embedding vector

(in either the traditional psychometric, or the more expansive fuzzy-relational sense) the proficiencies of the students who produced the text responses.

7.6 Conclusion

This discussion brings us back to the point of departure for this dissertation, namely educational assessment as a process in which 'evidence of performance or attainment is elicited, interpreted, and acted on in some way', in William and Black's formulation (1996, p. 540).

A large part of the practice of assessment involves using data that is quantitative by nature (or assumed to be quantitatively structured) to generate scores, interpreted as measures of students' proficiencies with respect to some educational domain of interest.

The papers presented in this dissertation suggest that it some cases it may be reasonable to regard the construct of proficiency as a quantity or vector of quantities, and in such cases there is a rich array of methods in test theory and psychometrics that can help support valid inferences from such scores. In other cases, however, where the assumption of quantitative structure is problematic, the use of methods based on other types of mathematical structure should be considered. In particular, representing manifest proficiency levels as partially-ordered fuzzy concepts, rather than locations on a totally-ordered continuum, can help shed light on questions such as whether, and how, students' demonstrated attainment in particular curriculum domains is changing over time.

Messick (1989, p.13) defined validity as 'an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequancy and appropriateness of inferences and actions based on test scores or other modes of assessment'. If we stop thinking about proficiency as being an amount of something, and instead

regard it as a configuration of cognitive resources, or a state of mental characteristics, then the inferences of interest are often not really 'how much?', but 'what kind of?'. Propositions about learners that reflect the extent to which they have demonstrated aspects (construct-relevant attributes) of what it means to be proficient in a certain area (according to a particular consensus, at a particular time) might, for some courses of action, usefully underwrite scores, classifications, or grades. For other purposes, using the information obtained in an assessment procedure to express students' similarity to certain prototypical standards, or to generate an output in natural language from an appropriately applied natural language model, may be better supported, either empirically or conceptually. The more open we are to thinking about the mathematical analysis of assessment information in ways that need not always be predicated on the notion of scoring, the richer the possibilities for beneficial action.

References

Adams, Ernest W. (1966). "On the nature and purpose of measurement". In: *Synthese* 16.2, pp. 125–169. ISSN: 0039-7857, 1573-0964. DOI: 10.1007/BF00485355.

Adams, R.M. and R. Murphy (1982). "The achieved weights of examination components". In: *Educational Studies* 8.1, pp. 15–22. ISSN: 0305-5698, 1465-3400. DOI: 10.1080/0305569820080102.

Andrich, D. (1978a). "Relationships between the Thurstone and Rasch approaches to item scaling". In: *Applied Psychological Measurement* 2.3, pp. 451–462. DOI: 10.1177/014662167800200319.

Andrich, D. and I. Marais (2019a). *A Course in Rasch Measurement Theory*. Springer.

Andrich, David (1978b). "A rating formulation for ordered response categories". In: *Psychometrika* 43.4, pp. 561–574.

– (1985). "An elaboration of Guttman scaling with Rasch models for measurement". In: *Sociological Methodology* 15, pp. 33–80. ISSN: 00811750. DOI: 10.2307/270846.

– (1988). *Rasch models for measurement*. London: SAGE Publications.

– (2013). "An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any 'threshold disorder controversy'". In: *Educational and Psychological Measurement* 73.1, pp. 78–124. ISSN: 0013-1644, 1552-3888. DOI: 10.1177/0013164412450877.

Andrich, David (2016). "Inference of independent dichotomous responses in the polytomous Rasch model". In: *Rasch Measurement Transactions* 30.1, pp. 1566–1569.

Andrich, David and Ida Marais (2019b). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Singapore: Springer Singapore. ISBN: 9789811374951 9789811374968. DOI: 10.1007/978-981-13-7496-8.

Angoff, W.H. (1971). "Scales, norms and equivalent scores". In: *Educational measurement*. Ed. by R.L. Thorndike. Washington, DC: American Council on Education, pp. 508–600.

Axler, S. (2023). *Linear algebra done right*. Cham: Springer. DOI: 10.1007/978-3-031-41026-0.

Badinger, H. and W.H. Reuter (2015). "Measurement of fiscal rules: introducing the application of partially ordered set (POSET) theory". In: *Journal of Macroeconomics* 43, pp. 108–123.

Baird, J. (2007). "Alternative conceptions of comparability". In: *Techniques for monitoring the comparability of examination standards*. Ed. by P. Newton et al. London: Qualifications and Curriculum Authority, pp. 124–165.

Baird, Jo-Anne, M. J. Cresswell, and Paul E. Newton (2000). "Would the real gold standard please step forward?" In: *Research Papers in Education* 15.2, pp. 213–229.

Baird, Jo-Anne et al. (2017). "Assessment and learning: fields apart?" In: *Assessment in Education: Principles, Policy & Practice* 24.3, pp. 317–350. ISSN: 0969-594X. DOI: 10.1080/0969594X.2017.1319337.

Ballou, Dale (2009). "Test scaling and value-added measurement". In: *Education Finance and Policy* 4.4, pp. 351–383. ISSN: 1557-3060, 1557-3079. DOI: 10.1162/edfp.2009.4.4.351.

Bartholemew, D.J. et al. (2008). *Analysis of Multivariate Social Science Data*. Florida: CRC Press.

Bartholomew, D. et al. (2011). *Analysis of multivariate social science data*. 2nd ed. London: Chapman and Hall.

Bartl, E. and R. Bělohlávek (2024). "Cardinality of fuzzy sets and accumulation of small membership". In: *IEEE Transactions on Fuzzy Systems* 32.6, pp. 3779–3789.

Bartl, E., R. Bělohlávek, and A. Scharaschkin (2018). "Toward factor analysis of educational data". In: *Proceedings of the 14th International Conference on Concept Lattices and their Applications*. Ed. by D. Ignatov and L. Nourine. Olomouc, Czech Republic, pp. 191–206.

Bartl, Eduard and Radim Belohlavek (2011). "Knowledge spaces with graded knowledge states". In: *Information Sciences* 181.8, pp. 1426–1439. ISSN: 00200255. DOI: 10.1016/j.ins.2010.11.040.

Bartl, Eduard, Radim Belohlavek, and Alex Scharaschkin (2018). "Toward factor analysis of educational data". In: *Proceedings of the 14th International Conference on Concept Lattices and their Applications*, pp. 191–206.

Bazan, J., A. Skowron, and R. Swiniarski (2006). "Rough Sets and Vague Concept Approximation: From Sample Approximation to Adaptive Learning". In: *Transactions on Rough Sets V: Lecture Notes in Computer Science 4100*. Ed. by J.F. Peters and A. Skowron. Berlin: Springer, pp. 39–62.

Bedek, M. and D. Albert (2015a). "Applying formal concept analysis to visualise classroom performance". In: *Proceedings of the 11th International Conference on Knowledge Management*. Ed. by T. Watanabe and K. Seta.

Bedek, Michael A and Dietrich Albert (2015b). "Applying formal concept analysis to visualize classroom performance". In: Proceedings of the 11th International Conference on Knowledge Management. Ed. by T. Watanabe and K. Seta.

Belohlavek, Radim (2002). *Fuzzy relational systems: foundations and principles*. New York: Springer (IFSR international series on systems science and engineering, vol. 20).

Belohlavek, Radim, J.W. Dauben, and G.J. Klir (2017). *Fuzzy logic and mathematics: a historical perspective*. Oxford: Oxford University Press.

Birkhoff, G. (1948). *Lattice theory (Revised edition)*. New York: American Mathematical Society.

Black, B. and P. Newton (2016). "Tolerating differences of opinion". In: 17th annual meeting of the Association for Educational Assessment (Europe), November 2016. Cyprus.

Bradley, T-D. (2020). "At the interface of algebra and statistics". PhD thesis. City University of New York.

Bradley, T-D., J.L. Gastaldi, and J. Terilla (2024). "The structure of meaning in language: Parallel narratives in linear algebra and category theory". In: *Notices of the American Mathematical Society* 71.2, pp. 174–185.

Bramley, T. (2007). "Paired comparison methods". In: *Techniques for monitoring the comparability of examination standards*. Ed. by P. Newton et al. London: Qualifications and Curriculum Authority, pp. 246–294.

Bramley, T. and C.L. Vidal Rodeiro (2014). *Using statistical equating for standard maintaining in GCSEs and A levels*. Cambridge Assesment Research Report. Cambridge, UK: Cambridge Assessment.

Bramley, Tom (2011). "Subject difficulty - the analogy with question difficulty". In: *Research Matters: A Cambridge Assessment Publication, Special Issue 2*, pp. 27–33.

Braun, H.I. and P.W. Holland (1982). "Observed-score test equating: A mathematical analysis of some ETS equating procedures". In: *Test equating*. Ed. by P.W. Holland and D. B. Rubin. New York: Academic Press, pp. 9–49.

Briggs, D.C. (2022). *Historical and Conceptual Foundations of Measurement in the Human Sciences*. New York: Routledge. DOI: 10.1201/9780429275326.

Briggs, Derek C. (2013). "Measuring growth with vertical scales: measuring growth with vertical scales". In: *Journal of Educational Measurement* 50.2, pp. 204–226. ISSN: 00220655. DOI: 10.1111/jedm.12011.

Brogden, H. E. (1977). "The Rasch model, the law of comparative judgment and additive conjoint measurement". In: *Psychometrika* 42.4, pp. 631–634. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02295985.

Buntins, M., K. Buntins, and F. Eggert (2016a). "Psychological tests from a (fuzzy-)logical point of view". In: *Quality & Quantity* 50.6, pp. 2395–2416.

Buntins, Matthias, Katja Buntins, and Frank Eggert (2016b). "Psychological tests from a (fuzzy-)logical point of view". In: *Quality & Quantity* 50.6, pp. 2395–2416. ISSN: 0033-5177, 1573-7845. DOI: 10.1007/s11135-015-0268-z.

Bělohlávek, R. (2012a). "Optimal decomposition of matrices with entries from residuated lattices". In: *Journal of Logic and Computation* 22.6, pp. 1405–1425.

– (2012b). "Optimal decomposition of matrices with entries from residuated lattices". In: *Journal of Logic and Computation* 22.6, pp. 1405–1425.

Bělohlávek, R., J.W. Dauben, and G.J. Klir (2017). *Fuzzy Logic and Mathematics: A Historical Perspective*. Oxford, UK: Oxford University Press.

Bělohlávek, R. and G. Klir, eds. (2011). *Concepts and Fuzzy Logic*. Cambridge, MA: The MIT Press.

Bělohlávek, R. and V. Vychodil (2016). "Factorization of matrices with grades". In: *Fuzzy Sets and Systems* 292, pp. 85–97. DOI: 10.1016/j.fss.2015.03.020.

Carnap, R. (1947). *Meaning and necessity: a study in semantics and modal logic*. Chicago: University of Chicago Press.

Carpineto, C. and G. Romano (2004). *Concept Data Analysis: Theory and Applications*. Chichester, UK: Wiley.

Cartwright, N.L. and R. Runhardt (2014). "Philosophy of Social Science: A New Introduction". In: ed. by N.L. Cartwright and E. Montuschi. Oxford: Oxford University Press. Chap. Measurement, pp. 265–287.

Chandler, D. and R. Munday (2011). *A Dictionary of Media and Communication*. Oxford, UK: Oxford University Press.

Cho, E. (2023). "Interchangeability between factor analysis, logistic IRT, and normal ogive IRT". In: *Frontiers in Psychology* 14, p. 1267219.

Choppin, B. (1968). "Item banking using sample free calibration". In: *Nature* 219.5156, pp. 870–872.

– (1985). "A fully conditional estimation procedure for Rasch model parameters". In: *Evaluation in Education* 9, pp. 29–42.

Coe, R. (2008). "Comparability of GCSE examinations in different subjects: an application of the Rasch model". In: *Oxford Review of Education* 34.5, pp. 609–636.

Cresswell, M. J. (1987). "A more generally useful measure of the weight of examination components". In: *British Journal of Mathematical and Statistical Psychology* 40.1, pp. 61–79. ISSN: 2044-8317. DOI: 10.1111/j.2044-8317.1987.tb00868.x.

– (2003). *Heaps, prototypes and ethics: the consequences of using judgements of students' performance to set examination standards in a time of change*. London: Institute of Education.

Cronbach, L.J. and P.E. Meehl (1955). "Construct validity in psychological tests". In: *Psychological Bulletin* 52.4, pp. 281–302.

Cuff, B. M. P., M. Meadows, and B. Black (2018). "An investigation into the Sawtooth Effect in secondary school assessments in England". In: *Assessment in Education: Principles, Policy Practice* 26.3, pp. 321–339. DOI: 10.1080/0969594X.2018.1513907.

Davey, B.A. and H.A. Priestley (2002). *Introduction to lattices and order*. Cambridge: Cambridge University Press.

Davier, A. von, R. Mislevey, and J. Hao, eds. (2021). *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment*. Springer.

Delap, Martin R. (1994). "The interpretation of achieved weights of examination components". In: *The Statistician* 43.4, p. 505. ISSN: 00390526. DOI: 10.2307/2348135.

Department for Education (2013). *Mathematics GCSE subject content and assessment objectives*. URL: www.education.gov.uk/schools/teachingandlearning/qualifications/gcses.

Doignon, J.-P. and J.-C. Falmagne (1999). *Knowledge spaces*. Berlin: Springer.

Domingue, B. (2014a). "Evaluating the equal-interval hypothesis with test score scales". In: *Psychometrika* 79.1, pp. 1–19.

Domingue, Ben (2014b). "Evaluating the equal-interval hypothesis with test score scales". In: *Psychometrika* 79.1, pp. 1–19. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/s11336-013-9342-4.

Dubois, D. and H. Prade (1990). "Rough fuzzy sets and fuzzy rough sets". In: *International Journal of General Systems* 17, pp. 191–209.

Dubois, Didier and Henri Prade (1988). "The treatment of uncertainty in knowledge-based systems using fuzzy sets and possibility theory". In: *International Journal of Intelligent Systems* 3.2, pp. 141–165. ISSN: 1098-111X. DOI: 10.1002/int.4550030204.

Ducamp, A. and J.-Cl. Falmagne (1969). "Composite measurement". In: *Journal of Mathematical Psychology* 6, pp. 359–390.

Eckart, C. and G. Young (1936). "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1.3, pp. 211–218.

Ellerman, D. (2022). "Follow the math!: The mathematics of quantum mechanics as the mathematics of set partitions linearized to (Hilbert) vector spaces". In: *Foundations of Physics* 52.100. DOI: 10.1007/s10701-022-00608-3.

Fattore, M. (2016). "Partially ordered sets and the measurement of multidimensional ordinal deprivation". In: *Social Indicators Research* 128.2, pp. 835–858.

Frege, G. (1892). "Über Sinn und Bedeutung". In: *Zeitschrift für Philosophie und philosophische Kritik* 100, pp. 25–50.

Galton, F. (1869). *Hereditary genius*. London: Macmillan.

Ganter, B. and C.V. Glodeanu (2014). "Factors and skills". In: *Formal Concept Analysis: 12th International Conference, ICFCA 2014*. Cluj-Napoca, Romania: Springer, pp. 173–187.

Ganter, B. and R. Wille (1999a). *Formal concept analysis - mathematical foundations*. Berlin: Springer.

– (1999b). *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer.

Ganter, B. et al. (2017). "An invitation to knowledge space theory". In: *Formal Concept Analysis: 14th International Conference, ICFCA 2017*. Rennes, France: Springer, pp. 3–19.

Garner, L. and G. Jr. Engelhard (2002). "An eigenvector method for estimating item parameters of the dichotomous and polytomous Rasch models". In: *Journal of Applied Measurement* 3, pp. 107–128.

Gelman, A. et al. (2013). *Bayesian data analysis*. 3rd ed. London: Chapman and Hall.

Goertz, G. (2006a). *Social science concepts*. Princeton: Princeton University Press.

– (2006b). *Social science concepts*. Princeton: Princeton University Press.

Goertz, G. and J. Mahoney (2012). *A tale of two cultures: qualitative and quantitative research in the social sciences*. Princeton: Princeton University Press.

Goguen, J.A. (1969a). "The logic of inexact concepts". In: *Synthese* 19.3, pp. 325–373.

– (1969b). "The logic of inexact concepts". In: *Synthese* 19.3, pp. 325–373.

Gottwald, S. (2001). *A treatise on many-valued logics*. Baldock: Research Studies Press.

Gould, S.J. (1996). *The mismeasure of man*. New York: W.W. Norton & Co.

Gray, L. (2020). "Evidence-based policy-making and exam board insider researchers: creating communicative spaces". In: *Assessment in Education: Principles, Policy & Practice* 27.2, pp. 142–159. DOI: 10.1080/0969594X.2020.1749557. eprint: https:

//doi.org/10.1080/0969594X.2020.1749557. URL: https://doi.org/10.1080/0969594X.2020.1749557.

Guttman, L. (1944a). "A basis for scaling qualitative data". In: *American Sociological Review* 9, pp. 139–150.

– (1971a). "Measurement as structural theory". In: *Psychometrika* 36, pp. 329–347.

Guttman, Louis (1944b). "A basis for scaling qualitative data". In: *American Sociological Review* 9.2, pp. 139–150. ISSN: 00031224. DOI: 10.2307/2086306.

– (1971b). "Measurement as structural theory". In: *Psychometrika* 36, pp. 329–347.

Hajek, P. (1998a). *Metamathematics of fuzzy logic*. Dordrecht: Kluwer.

– (1998b). *Metamathematics of Fuzzy Logic*. Dordrecht: Kluwer.

Hampton, J.A. (2006). "Concepts as prototypes". In: *Psychology of Learning and Motivation* 46, pp. 79–113.

Heene, M. (2013a). "Additive conjoint measurement and the resistance toward falsifiability in psychology". In: *Frontiers in Psychology* 4.246.

– (2013b). "Additive conjoint measurement and the resistance towards falsifiability in psychology". In: *Frontiers in Psychology* 4, p. 246.

Heilmann, Conrad (2015). "A New Interpretation of the Representational Theory of Measurement". eng. In: *Philosophy of science* 82.5, pp. 787–797. ISSN: 0031-8248.

Hirth, J. and T. Hanika (2022). "Formal conceptual views in neural networks". Available at https://doi.org/10.48550/arXiv.2209.13517.

Höhle, U. (1996). "On the fundamentals of fuzzy set theory". In: *Journal of Mathematical Analysis and Applications* 201, pp. 786–826.

Hölder, O. (1901). "Die Axiome der Quantität und die Lehre vom Mass". In: *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* 53, pp. 1–46.

Holmes, S., B. Black, and C. Morin (2017). *Marking reliability studies 2017: rank ordering versus marking: which is more reliable?* Tech. rep. Ofqual.

Huang, Baokun et al. (2023). "Competence-based knowledge space theory from the perspective of formal concept analysis". Available at SSRN: https://ssrn.com/abstract=4620449 or http://dx.doi.org/10.2139/ssrn.4620449.

Isham, C.J. (1995). *Lectures on Quantum Theory*. London: Imperial College Press.

Jadhav, C. (24 March 2017). *Levelling the playing field*. Ofqual blog. URL: https://ofqual.blog.gov.uk/2017/03/24/levelling-the-playing-field/.

Joliffe, I.T. (2010). *Principal component analysis*. New York: Springer. DOI: 10.1007/b98835.

Jones, I. and B. Davies (2023). "Comparative judgement in education research". In: *International Journal of Research and Methods in Education* 47.2, pp. 170–181. DOI: 10.1080/1743727X.2023.2242273.

Kane, M. (2008). "The benefits and limits of formality". In: *Measurement: Interdisciplinary Research and Perspectives* 6, pp. 101–108.

Karabatsos, G. (2001a). "The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory". In: *Journal of Applied Measurement* 2.4, pp. 389–423.

– (2018). "On Bayesian testing of additive conjoint measurement axioms using synthetic likelihood". In: *Psychometrika* 83.2, pp. 321–332.

Karabatsos, George (2001b). "The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory". In: *Journal of Applied Measurement* 2.4, pp. 389–423.

Kelly, K.T., M. Richardson, and T. Isaacs (2022). "Critiquing the rationales for using comparative judgement: a call for clarity". In: *Assessment in Education: Principles, Policy Practice* 29.6, pp. 674–688. DOI: 10.1080/0969594X.2022.2147901.

Keynes, J.M. (1921). *A treatise on probability*. London: Macmillan.

Kline, P. (2000). *A Psychometrics Primer*. London, UK: Free Association Press.

Kneale, W. and M. Kneale (1962). *The development of logic*. Oxford: Oxford University Press.

Kolmogorov, Andrei (1954). *Foundations of the theory of probability*. 2nd ed. New York: Chelsea Publishing Company.

Krantz, David H. et al. (1971a). *Foundations of measurement*. New York: Academic Press. ISBN: 978-0-486-45314-9.

Krantz, D.H. et al. (1971b). *Foundations of Measurement. Volume I: Additive and Polynomial Representations*. New York: Academc Press.

Krupka, M. (2008). "Factorization of concept lattices with hedges by means of factorization of residuated lattices". In: *CLA 2008: Proceedings of the Sixth International Conference on Concept Lattices and their Applications*. Ed. by R. Bělohlávek and S. Kuznetsov. Olomouc, Czech Republic, pp. 231–242.

Kyngdon, A. (2011). "Plausible measurement analogies to some psychometric models of test performance: plausible conjoint systems". In: *British Journal of Mathematical and Statistical Psychology* 64.3, pp. 478–497.

Laurence, S. and E. Margolis (1999). "Concepts and cognitive science". In: *Concepts: core readings*. Ed. by E. Margolis and S. Laurence. Cambridge, MA: The MIT Press, pp. 3–81.

Leighton, J.P. and M.J. Gierl (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.

Li, Jiazheng et al. (2023). "Distilling ChatGPT for Explainable Automated Student Answer Assessment". Available at https://doi.org/10.48550/arXiv.2305.12962.

Linden, W van der and R Hambleton (1997a). *Handbook of modern item response theory*. New York: Springer.

Linden, W.J. van der and K.K. Hambleton, eds. (1997b). *Handbook of Modern Item Response Theory*. New York: Springer.

Lord, F.M. and M.R. Novick (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Luce, R.D. and L. Narens (1994). "Fifteen problems concerning the representational theory of measurement". In: *Patrick Suppes: Scientific Philosopher*. Ed. by P. Humphries. Dordrecht: Springer Netherlands, pp. 219–249.

Luce, R.D. and J.W. Tukey (1964a). "Simultaneous conjoint measurement: a new scale type of fundamental measurement". In: *Journal of Mathematical Psychology* 1, pp. 1–27.

Luce, R.Duncan and John W. Tukey (1964b). "Simultaneous conjoint measurement: A new type of fundamental measurement". In: *Journal of Mathematical Psychology* 1.1, pp. 1–27. ISSN: 00222496. DOI: 10.1016/0022-2496(64)90015-X.

Łukasiewicz, J. (1923 [1970]). "A numerical interpretation of the theory of propositions [trans. O. Wojtasiewicz from Interpretacja liczbowa teorii zdań". In: *Jan Łukasiewicz: Selected works*. Ed. by L. Borkowski. Amsterdam: North Holland, pp. 129–130.

Machery, E. (2011). "Concepts: a tutorial". In: *Concepts and fuzzy logic*. Ed. by R. Belohlavek and G. Klir. Cambridge, MA: The MIT Press, pp. 13–44.

Marquer, E. (2020). "LatticeNN: Deep Learning and Formal Concept Analysis". MA thesis. Université de Lorraine.

Maul, A. (2017a). "Rethinking traditional methods of survey validation". In: *Measurement: Interdisciplinary Research and Perspectives* 15.2, pp. 51–69.

Maul, Andrew (2017b). "Rethinking traditional methods of survey validation". In: *Measurement: Interdisciplinary Research and Perspectives* 15.2, pp. 51–69. ISSN: 1536-6367, 1536-6359. DOI: 10.1080/15366367.2017.1348108.

McGrane, J. and A. Maul (2020). "The human sciences: models and metrological mythology". In: *Measurement* 152, p. 107346.

McGrane, Joshua A. and Andrew Maul (2020). "The human sciences, models and metrological mythology". In: *Measurement* 152, p. 107346. ISSN: 02632241. DOI: 10.1016/j.measurement.2019.107346.

Messick, S. (1989). "Validity". In: *Educational Measurement*. Ed. by R.L. Linn. 3rd edition. New York: American Council on Education and Macmillan, pp. 13–104.

Michell, J. (2006a). "Psychophysics, intensive magnitudes and the psychometricians' fallacy". In: *Studies in the History and Philosophy of Biological and Biomedical Sciences* 17, pp. 414–432.

– (2009a). "The psychometricians' fallacy: too clever by half". In: *British Journal of Mathematical and Statistical Psychology* 62.1, pp. 41–44.

– (2012a). "The constantly recurring argument: inferring quantity from order". In: *Theory and Psychology* 22.3, pp. 255–271.

– (2013). "Constructs, inferences and mental measurement". In: *New Ideas in Psychology* 31, pp. 13–21.

– (2021a). "Representational measurement theory: is its number up?" In: *Theory and Psychology* 31.1, pp. 3–23.

Michell, Joel (1990a). *An introduction to the logic of psychological measurement*. London: Taylor & Francis.

– (1990b). *An Introduction to the Logic of Psychological Measurement*. London, UK: Routledge.

– (1994). "Numbers as quantitative relations and the traditional theory of measurement". In: *The British Journal for the Philosophy of Science* 45.2, pp. 389–406. ISSN: 0007-0882, 1464-3537. DOI: 10.1093/bjps/45.2.389.

– (1997). "Quantitative science and the definition of measurement in psychology". In: *British Journal of Psychology* 88.3, pp. 355–383. ISSN: 00071269. DOI: 10.1111/j.2044-8295.1997.tb02641.x.

Michell, Joel (1999a). *Measurement in psychology : critical history of a methodological concept*. eng. Ideas in context ; 53. Cambridge: Cambridge University Press. ISBN: 1-107-11472-1.

– (1999b). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press. ISBN: 978-0-511-15042-5.

– (2002). "Do ratings measure latent attributes?" In: *Ergonomics* 45.14, pp. 1008–1010. ISSN: 0014-0139, 1366-5847. DOI: 10.1080/00140130210166843.

– (2005). "The logic of measurement: A realist overview". In: *Measurement* 38.4, pp. 285–294. ISSN: 02632241. DOI: 10.1016/j.measurement.2005.09.004.

– (2006b). "Psychophysics, intensive magnitudes, and the psychometriciansâĂŹ fallacy". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 37.3, pp. 414–432. ISSN: 13698486. DOI: 10.1016/j.shpsc.2006.06.011.

– (2009b). "The psychometricians' fallacy: Too clever by half?" In: *British Journal of Mathematical and Statistical Psychology* 62.1, pp. 41–55. ISSN: 00071102. DOI: 10.1348/000711007X243582.

– (2012b). "'The constantly recurring argument': Inferring quantity from order". In: *Theory & Psychology* 22.3, pp. 255–271. ISSN: 0959-3543, 1461-7447. DOI: 10.1177/0959354311434656.

– (2014). "The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma". In: *Theory & Psychology* 24.1, pp. 111–123. ISSN: 0959-3543, 1461-7447. DOI: 10.1177/0959354313517524.

– (2020). "Thorndike's *Credo* : Metaphysics in psychometrics". In: *Theory & Psychology* 30.3, pp. 309–328. ISSN: 0959-3543, 1461-7447. DOI: 10.1177/0959354320916251.

– (2021b). "Representational measurement theory: Is its number up?" In: *Theory & Psychology* 31.1, pp. 3–23. ISSN: 0959-3543, 1461-7447. DOI: 10.1177/0959354320930817.

Mikolov, T. et al. (2013). *Efficient estimation of word representations in vector spaces*. arXiv:1301.3781.

Mokken, R.J. (1971). *A theory and procedure of scale analysis, with applications in political research*. De Gruyter Mouton.

Mosier, C. I. (1940). "Psycho-physics and mental test theory: fundamental postulates and elementary theorems". In: *Psychological Review* 47.4, pp. 355–366. ISSN: 0033-295X.

Mosier, Charles I. (1941). "A psychometric study of meaning". In: *Journal of Social Psychology* 13.1, pp. 123–140. ISSN: 0022-4545.

Munroe, Marshall Evans (1965). *Introductory real analysis*. Reading, Mass.: Addison-Wesley Pub. Co.

Newton, P. et al., eds. (2007). *Techniques for Monitoring the Comparability of Examination Standards*. London, UK: Qualifications and Curriculum Authority.

Newton, Paul E. (2007). "Clarifying the purposes of educational assessment". In: *Assessment in Education: Principles, Policy & Practice* 14.2, pp. 149–170. ISSN: 0969-594X, 1465-329X. DOI: 10.1080/09695940701478321.

– (2018). *Grading vocational and technical qualifications*. Ofqual report. Coventry, UK.

Newton, Paul E. and Stuart Shaw (2014). *Validity in educational and psychological assessment*. SAGE Publications.

Newton, P.E. (2020). *Maintaining standards - during normal times and when qualifications are reformed*. Ofqual report. Coventry, UK.

Ofqual (2017a). *GCSE, AS and A level assessment objectives*. Available at www.gov.uk/government/publicat
objectives-ancient-languages-geography-and-mfl/gcse-as-and-a-level-assessment-objectives.

– (2017b). *Inter-board comparability of grade standards in GCSEs, AS and A levels 2017*. Tech. rep. Ofqual.

– (2018). *Marking consistency metrics - an update*. Ofqual research report. Coventry, UK.

– (2019). *National Reference Test Information*. Coventry, UK.

Ofqual (2022). *GCSE (9 to 1) qualification-level conditions and requirements*. Available at www.gov.uk/government/publications/gcse-9-to-1-qualification-level-conditions.

Ostini, R. and M.L. Nering, eds. (2005). *Handbook of polytomous item response theory models*. Routledge.

Partree, B.H., A. ter Meulen, and R.E. Wall (1993). *Mathematical methods in linguistics*. Dordrecht: Kluwer.

Pennington, J., R. Socher, and C.D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, pp. 1532–1543.

Peres, A. (1995). *Quantum Theory: Concepts and Methods*. Dordrecht: Kluwer.

Perline, Richard, Benjamin D. Wright, and Howard Wainer (1979). "The Rasch model as additive conjoint measurement". In: *Applied Psychological Measurement* 3.2, pp. 237–255. ISSN: 0146-6216, 1552-3497. DOI: 10.1177/014662167900300213.

Pollit, Alastair et al. (2008). *Improving the quality of GCSE assessment*. Report commissioned by the Qualifications and Curriculum Authority.

Pollitt, A. and A. Ahmed (2008). *Outcome space control and assessment*. Tech. rep. Hissar, Bulgaria: Paper for the 9th annual conference of the Association for Educational Assessment—Europe.

Rasch, Georg (1960). *Probabilistic Methods for some Intelligence and Attainment Tests*. Chicago, Illinois: University of Chicago Press.

Raykov, Tenko. and George A. Marcoulides (2011). *Introduction to psychometric theory*. eng. New York: Routledge. ISBN: 1-136-90002-0.

Reid, T. (1748 [1849]). "An essay on quantity". In: *The Works of Thomas Reid*. Ed. by W. Hamilton. Edinburgh: Maclachlan, Stuart and Co., pp. 715–719.

Reyment, R.A. and K.G. Jöreskog (1993). *Applied Factor Analysis in the Natural Sciences*. Cambridge, UK: Cambridge University Press.

Rooij, R. van (2011). "Measurement and interadjective comparison". In: *Journal of Semantics* 28, pp. 335–358.

Rudolph, L. (2013). *Qualitative mathematics for the social sciences: mathematical models for research on cultural dynamics*. London: Routledge.

Scharaschkin, A. (2023). "Measuring educational constructs qualitatively". Paper presented at the Annual Conference of the Association for Educational Assessment Europe, Malta, November 2023.

– (2024). "Educational assessment without numbers". In: *Frontiers in Psychology* 15:1399317. DOI: 10.3389/fpsyg.2024.1399317.

Scharaschkin, Alex (2017). "Commentary on Baird, J., Andrich, D., Hopfenbeck, T. N. and Stobart, G., 'Assessment and learning: fields apart?'" In: *Assessment in Education: Principles, Policy & Practice* 24.3, pp. 454–462. ISSN: 0969-594X. DOI: 10.1080/0969594X.2017.1331905.

Scharaschkin, Alex and Jo-Anne Baird (2000). "The effects of consistency of performance on A level examiners' judgements of standards". In: *British Educational Research Journal* 26.3, pp. 343–357.

Schmidt, G. and T. Strohlein (1993). *Relations and graphs: discrete mathematics for computer scientists*. Berlin: Springer.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: Princeton University Press.

Shye, S. and D. Elizur, eds. (1994). *Introduction to facet theory*. Thousand Oaks, CA: Sage Publications, Inc.

Spearman, C. (1904). "'General intelligence', objectively defined and measured". In: *American Journal of Psychology* 15.2, pp. 201–292.

Stefanutti, L. and D. de Chiusole (2017). "On the assessment of learning in competence based knowledge space theory". In: *Journal of Mathematical Psychology* 80, pp. 22–32.

Stenner, A Jackson and William P Fisher (2013). "Metrological traceability in the social sciences: A model from reading measurement". In: *Journal of Physics: Conference Series* 459, p. 012025. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/459/1/012025.

Stevens, S. S. (1946a). "On the theory of scales of measurement". In: *Science, New Series* 103.2684, pp. 677–680.

Stevens, S.S. (1946b). "On the theory of scales of measurement". In: *Science* 103.2684, pp. 677–680.

Sudmann, A. et al., eds. (2023). *Beyond Quantity: Research with Subsymbolic AI*. Bielefeld: transcript Verlag.

Suppes, Patrick (1951). "A set of independent axioms for extensive quantities". In: *Portugaliae Mathematica* 10, pp. 163–172.

Susskind, L. and A. Friedman (2014). *Quantum Mechanics: The Theoretical Minimum*. London: Allan Lane.

Tal, E. (2020). "Measurement in science". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E.N. Zalta. Stanford University.

Tatsuoka, K.K. (2009). *Cognitive assessment: An introduction to the rule space method*. Boca Raton, FL: CRC Press.

Thurstone, L.L. (1927a). "A law of comparative judgement". In: *Psychological Review* 34, pp. 278–286.

– (1927b). "The method of paired comparisons for social values". In: *The Journal of Abnormal and Social Psychology* 21.4, pp. 384–400.

– (1927c). "The method of paired comparisons for social values". In: *Journal of Abnormal and Social Psychology* 21, pp. 384–400.

– (1928). "Attitudes can be measured". In: *American Journal of Sociology* 33, pp. 529–554.

– (1931). "Multiple factor analysis". In: *Psychological Review* 38, pp. 406–427.

Thurstone, L.L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Trimmer, J.D. (1980). "The present situation in quantum mechanics: A translation of Schrödinger's "cat paradox" paper". In: *Proceedings of the American Philosophical Society* 124.5, pp. 323–338.

Uher, J. (2021). "Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies". In: *Journal of Theoretical and Philosophical Psychology* 41, pp. 58–84.

– (2022a). "Functions of units, scales and quantitative data: Fundamental differences in numerical traceability between sciences". In: *Quality and Quantity* 56, pp. 2519–2548.

– (2022b). "Rating scales institutionalise a network of logical errors and conceptual problems in research practices: A rigorous analysis showing ways to tackle psychology's crises". In: *Frontiers in Psychology* 13, p. 1009893.

van Fraassen, Bas C. (2008). *Scientific Representation: Paradoxes of Perspective.* Oxford, UK: Oxford University Press.

Vaswani, A. et al. (2017). *Attention is all you need.* arXiv:1706.03762.

Weyl, H. (1952). *Space, Time, Matter.* New York: Dover.

Wiliam, D. (2017). "Assessment and learning: a long and winding road". In: *Assessment in Education: Principles, Policy and Practice* 24.3, pp. 309–316.

Wiliam, Dylan and Paul Black (1996). "Meanings and consequences: A basis for distinguishing formative and summative functions of assessment?" In: *British Educational Research Journal* 22.5, pp. 537–548.

Wind, Stefanie A. (2017). "An instructional module on Mokken scale analysis". In: *Educational Measurement: Issues and Practice* 36.2, pp. 50–66. ISSN: 07311745. DOI: 10.1111/emip.12153.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Basil Blackwell.

Wolff, J. E. (2020a). *The metaphysics of quantities.* First edition. Oxford scholarship online. Oxford: Oxford University Press. ISBN: 978-0-19-189588-3.

Wolff, J.E. (2020b). *The Metaphysics of Quantities*. Oxford, UK: Oxford University Press.

Zand Scholten, A. (2011). "Admissible statistics from a latent variable perspective". PhD thesis. University of Amsterdam.

Appendix A: Appendix to Chapter 4 [Paper 1]

Quantitative structure

Let P be a property of interest. For example, P could be the level of attainment, such as pass, merit, or distinction, that an assessor gives to each student's portfolio in an art assessment. P could be the mark-level (e.g. a mark out of 25) given to an A level English essay by an examiner. P could be the age of a student, the length of a rod, or the colour of an object. In each case, P can assume, for any particular object, one of a finite or infinite collection of labels, descriptions, or values.

Then the property P is *totally ordered* if there is a binary relation $\geq$ on its values that satisfies the following three conditions. For any values x, y, and z of P,

1. (strong connectivity) either $x \geq y$ or $y \geq x$;
2. (antisymmetry) if $x \geq y$ and $y \geq x$, then $x = y$; and
3. (transitivity) if $x \geq y$ and $y \geq z$, then $x \geq z$.

The property P is *additive* if additionally there is a binary relation $+$ on its values that satisfies the following six conditions. For any values x, y, and z of P,

1. (associativity) $x + (y + z) = (x + y) + z$;
2. (commutativity) $x + y = y + x$;
3. (monotonicity) $x \geq y$ if and only if $x + z \geq y + z$;
4. (solvability) if $x > y$ then there exists a value z such that $x = y + z$;
5. (positivity) $x + y \geq x$;
6. (Archimedean condition) there exists a positive integer n such that $nx \geq y$.

A quantitative property Q (i.e., one whose values satisfy the nine conditions above) is *continuous* if its possible values form a continuum with no 'gaps'. Formally, this requires two further conditions, over and above Q's being totally ordered and additive. Firstly, Q must be *dense,* which means that between any two values or levels of Q, there is another value:

- (denseness) if x and y are values of Q, with $x \geq y$, then there exists a value z of Q such that $x \geq z \geq y$.

Secondly, Q must be *complete*, which means that all sets of values of Q that are bounded above have a least upper bound:

- (completeness) Let $\mathscr{X}$ be a set of values of Q. Then there is a value y of Q such that (i) $y \geq x$ for all values x in $\mathscr{X}$; and (ii) if z is any other value such that $z \geq x$ for all values x in $\mathscr{X}$, then $z \geq y$.

Axioms of additive conjoint measurement

Let X and Y be two sets. Denote by $X \times Y$ the set of all ordered pairs (x, y), where $x \in X$ and $y \in Y$. Let $\succeq$ be a binary relation on $X \times Y$. Then the triple $(X, Y, \succeq)$ is defined to be an *additive conjoint structure* iff $\succeq$ satisfies the following axioms:

1. $\succeq$ is transitive and strongly connected[1].

2. (Independence, or single cancellation) If $(x_1, y_1) \succeq (x_2, y_1)$, then for all $y_2 \in Y$, $(x_1, y_2) \succeq (x_2, y_2)$.

3. (Double cancellation) If $(x_1, y_2) \succeq (x_2, y_1)$ and $(x_2, y_3) \succeq (x_3, y_2)$, then $(x_1, y_3) \succeq (x_3, y_1)$.

4. (Solvability) For all $x_1 \in X$ and $y_1, y_2 \in Y$, there exists an $x_2 \in X$ such that $(x_1, y_1) = (x_2, y_2)$.

5. (Archimedean condition) This essentially ensures that, as Michell (1990a) notes (p.73), 'no value of a quantitative variable is infinitely larger than any other value'. It is more precisely stated in terms of so-called standard sequences, which intuitely capture a notion of differences between values that is necessary if the properties that constitute the conjoint structure are to be quantifiable.

   To be precise, let I be any set of consecutive integers. Then the set $\{x_i : x_i \in X, i \in I\}$ is a standard sequence on X iff there exist distinct $y_1, y_2 \in Y$ such that for all $i \in I$, $(x_i, y_1) \sim (x_{i+1}, y_2)$. Standard sequences on Y are defined analogously. The

[1] Many treatments of the representational theory of measurement, such as Krantz et al. (1971a), call a binary relation that is transitive and connected a *weak order*. The more usual term in the mathematics literature for a relation that satisfies these conditions is a *total pre-order*. Spelling out the definitions of transitivity and strong connectivity, they are that for all pairs (x_1, y_1), (x_2, y_2), $(x_3, y_3) \in X \times Y$:

- if $(x_1, y_1) \succeq (x_2, y_2)$ and $(x_2, y_2) \succeq (x_3, y_3)$, then $(x_1, y_1) \succeq (x_3, y_3)$; and
- either $(x_1, y_1) \succeq (x_2, y_2)$ or $(x_2, y_2) \succeq (x_1, y_1)$.

Archimedean condition is that any bounded standard sequence is finite. Luce and Narens (1994) note that 'In practice, the impact of Archimedeanness is to permit homomorphisms into the real numbers, rather than ordered extensions of the real numbers such as the non-standard reals'. That is, Archimedeanness is required because the representational theory of measurement always aims to represent empirical relational structures as purely *numerical* structures (the real numbers or a suitable sub-structure of the real number field), rather than permitting a wider class of potential representations.

Appendix B: Published version of Chapter 5 [Paper 2]

frontiers | Frontiers in Psychology

TYPE Hypothesis and Theory
PUBLISHED 02 October 2024
DOI 10.3389/fpsyg.2024.1399317

OPEN ACCESS

EDITED BY
Jana Uher,
University of Greenwich, United Kingdom

REVIEWED BY
Robert Mislevy,
University of Maryland, United States
Andreas Sudmann,
University of Bonn, Germany

*CORRESPONDENCE
Alex Scharaschkin
✉ ascharaschkin@aqa.org.uk

RECEIVED 11 March 2024
ACCEPTED 19 September 2024
PUBLISHED 02 October 2024

CITATION
Scharaschkin A (2024) Educational assessment without numbers. *Front. Psychol.* 15:1399317. doi: 10.3389/fpsyg.2024.1399317

Educational assessment without numbers

Alex Scharaschkin[1,2]*

[1]Department of Education, University of Oxford, Oxford, United Kingdom, [2]AQA Education, London, United Kingdom

Psychometrics conceptualizes a person's *proficiency* (or *ability*, or *competence*), in a cognitive or educational domain, as a latent numerical quantity. Yet both conceptual and empirical studies have shown that the assumption of quantitative structure for such phenomena is unlikely to be tenable. A reason why most applications of psychometrics nevertheless continue to treat them as if they were numerical quantities may be that quantification is thought to be necessary to enable *measurement*. This is indeed true if one regards the task of measurement as the location of a measurand at a point on the real number line (the viewpoint adopted by, for example, the representational theory of measurement, the realist theory of measurement as the discovery of ratios, and Rasch measurement theory). But this is not the only philosophically respectable way of defining the notion of measurement. This paper suggests that van Fraassen's more expansive view of measurement as, in general, *location in a logical space* (which could be the real continuum, as in metrological applications in the physical sciences, but could be a different mathematical structure), provides a more appropriate conceptual framework for psychometrics. Taking educational measurement as a case study, it explores what that could look like in practice, drawing on fuzzy logic and mathematical order theory. It suggests that applying this approach to the assessment of intersubjectively constructed phenomena, such as a learner's proficiency in an inherently fuzzily-defined subject area, entails recognizing the theory-dependent nature of valid representations of such phenomena, which need not be conceived of structurally as values of quantities. Finally, some connections are made between this "qualitative mathematical" theorization of educational assessment, and the application of techniques from machine learning and artificial intelligence in this area.

KEYWORDS

theory and philosophy of measurement, psychometrics, educational assessment, van Fraassen, qualitative mathematics, concept lattice, fuzzy logic

1 Introduction

The question of what it could mean to *measure* phenomena that form the basis of theory and debate in the human sciences, such as human attitudes, opinions, dispositions, or psychological or cognitive traits, has been a subject of critical enquiry since at least the mid eighteenth century (Michell, 1999). For example, the question of whether such phenomena could be *quantified* was contested by Reid (1849), even before a clearer definition of "a quantity" had been put forward by Hölder (1901).

This paper considers the question of measuring educational constructs, such as a learner's *ability*, or *proficiency*, or *competence* in a subject, field of study, or educational domain. Many educational tests and assessment procedures—some of them used to make high-stakes decisions about the test-takers—apparently produce, or claim to produce, numerical measurements of such properties, such that learners can be placed on a

quantitative *scale* with respect to them. Psychometrics is the application of statistical methods to the study of psychological and educational phenomena. It relies on the particular mathematical characteristics of quantitative structures (in practice, the real numbers and vector spaces over the reals) to perform calculations and procedures that are used as the warrants for substantive conclusions, such as "how much" ability a student is estimated to have, or how to equate measurements of ability derived from different tests.

The paper argues that the reliance of psychometrics on quantitative structures is grounded in an assumption that *quantification* is necessary to allow *measurement*. It proposes, however, that psychological and educational measurement need not be reliant on numbers. It suggests that van Fraassen's (2008) account of measurement as a process whereby the measurand is located in an appropriate "logical space" is well-suited to serve as a foundation for an account of the measurement of educational phenomena such as students' abilities or competencies in a subject domain—phenomena that are arguably inherently "fuzzy" and multifaceted. Such a logical space *could* be the particular mathematical structure that uniquely characterizes the real numbers (a complete ordered field, in mathematical terminology), but it need not be.

The structure of the paper is as follows. Section 2 briefly outlines the approach to measuring cognitive and educational constructs, by assuming quantitative structure, that became standard in psychometrics over the twentieth century. It summarizes critiques of the quantity assumption, and argues that these critiques have sufficient conceptual and empirical weight to warrant a serious explanation of what an approach to psychological and educational measurement could look like if the assumption is set aside. Taking the example of summative educational assessment in particular, it suggests that in many cases construct validity may be better served by a more generalized view of measurement, of the kind proposed by van Fraassen (2008). Van Fraassen's approach is explained in more detail in Section 3.

Section 4 makes the discussion more concrete by comparing quantitative and qualitative measurement approaches for a toy example of an educational test. This is extended in Section 5 to a consideration of the practicalities—in particular, the computational complexity—of applying qualitative mathematical (fuzzy order-theoretic) methods to the kinds of test response data that arise in real practice. And since traditional methods of analysis of educational assessment data are increasingly being supplemented, or even supplanted, by the application of techniques from natural language processing, machine learning, and artificial intelligence (AI), Section 6 considers some of the connections between educational measurement and AI-enabled classification procedures. Finally, the concluding discussion in Section 7 poses some questions for further research. It concludes that it is worth pursuing further conceptual and technical development of non-quantitative measurement approaches in psychometrics, especially since, with the rapid rise and application of AI (e.g., von Davier et al., 2021), there is a risk that psychometrics is simply replaced with data science—with the loss of substantive theoretical content concerning construct definition and the design of valid measurement procedures. A way forward is for psychometrics itself to develop into a discipline that rests on quantitative measurement when it is appropriate, but does not exclude a broader view.

2 Quantification in psychometrics

2.1 Abilities as latent quantities

Psychometrics normally conceptualizes a learner's *ability* (or *proficiency*, or *competence*) in a domain as a latent numerical quantity, θ (Kline, 2000; van der Linden and Hambleton, 1997). For each learner, a value of θ is calculated from the observed data arising from an assessment (e.g., item response data). The "more θ" a learner has (the higher their value of θ), the "better at" the assessment construct they are taken to be (modulo some "measurement error"). That is to say, the relation of *betterness*, between learners, as to the different levels, states, or configurations of their abilities, is taken to be adequately captured by the relation of *order* ($\geq$) between numerical values. Moreover, to allow a value of θ actually to be derived for each learner, the set of all possible θ-values is normally supposed not only to be totally ordered, but quantitative and continuous.[1] Making these structural assumptions about the property of *ability* enables it to be treated as if it were a real number. Hence the whole array of statistical techniques whose mathematical validity depends on the metric and topological properties of the real numbers (such as factor analysis, item response theory, maximum likelihood estimation, etc.) can be applied to obtain numerical values that are taken to be *measurements of learners' abilities* in the cognitive or educational domain in question.

This paper will argue that one should not think of the "betterness" relation between learners, as to their proficiency in a particular educational domain, as a total order relation (a ranking), in general, but rather as a partial order.[2] Sometimes the way in which the assessment construct is defined will allow learners to be ranked as to their proficiency with respect to that construct. In other cases, it may only be possible to infer, for some pairs of learners, that their proficiency states, or levels, are non-comparable (qualitatively different). This does not preclude the possibility of

1 See the Appendix for definitions of *total order* and *quantity*. Informally, a totally ordered set X is one in which all the members can be ranked—there is an ordering $\geq$ such that either $x \geq y$ or $y \geq x$, for all x and y in X. A property is a quantity if its values are totally ordered and also additive—that is, they can be combined in a way that mirrors the properties of the addition of numbers. Additivity is required for a property's values to form an *interval scale* or a *ratio scale*, in the terminology of Stevens (1946). A quantitative property is *continuous* if its possible values form a continuum with no "gaps".

2 See the Appendix for a formal definition of *partial order*. In essence, when entities are partially ordered, there may exist pairs of entities that are not directly comparable, and the entities cannot necessarily be placed in a single linear sequence (a ranking) with respect to the feature of interest. In educational tests, each individual item (question or task) typically totally orders the respondents with respect to that item (for example "those who got the question right" $\geq$ "those who got the question wrong"; or "those who scored 3 marks" $\geq$ "those who scored 2 marks" $\geq$ "those who scored 1 mark' $\geq$ "those who scored 0 marks") In general, however, the joint result (the product) of all of these total orders is an overall partial ordering of respondents, with some patterns of item responses not being directly comparable with others.

grouping learners together into "coarser" ordinal classes (such as examination grades), such that one can infer that those who "pass" are more proficient than those who "fail", for instance. It just means that, within the "pass" category, there may be some learners whose proficiencies, although both of at least a "pass" level, may be different, and non-comparable. This argument is developed further in Section 4 below.

There is a literature that critically examines the plausibility of assuming quantitative structure for phenomena such as ability (for example, Michell, 2006, 2009, 2012, 2013; Heene, 2013; Kyngdon, 2011; McGrane and Maul, 2020, and from a broader perspective, Uher, 2021, 2022a). One focus of this has been what Michell (2012) calls the "psychometricians' fallacy": the implicit leap that is often made, from maintaining that a property has a totally-ordered structure (that its possible values, states, or levels can be ranked, that is, placed on an *ordinal scale*, as described by Stevens, 1946), to treating it as if it had quantitative structure (as if its values formed an *interval* or a *ratio* scale, in Stevens' typology).

In some cases it is possible to test empirically whether a property whose values are ordered is plausibly likely to have the further structure required for it to be quantitative. This is discussed in Section 2.2.2. Yet at an even more basic level, one might question why a construct such as *ability* with respect to a given cognitive or educational domain (specified in a more-or-less precise way), should even be regarded as a property that necessarily ought to have a totally ordered structure. Must it be a phenomenon that only occurs in such a way that any one person's ability-state is always linearly comparable with (larger than, the same as, or smaller than) any other person's state? Uher (2022b) makes an analogous point with respect to the use of rating scales to "measure" the property of agreement.

If one considers the actual data upon which the inferences derived from educational testing procedures are based, then as Kane (2008) notes, "we are likely to have, at best, a partial ordering, unless we arbitrarily decide that some patterns [of item response] are better than others". In practice, and as discussed further in Section 4, almost all psychometric approaches to working with such partially-ordered data do indeed involve making decisions about how to use the data to generate a total order (with each learner's score being their location with respect to this total order).

The question whether such decisions are indeed "arbitrary" (and if not, which one is best or most appropriate) hinges, again, on how the measurand—each respondent's ability in the domain in question—is conceptualized. This issue is well-described by Maul (2017, p. 60), who notes that

> Any effort to construct a measure of an attribute will have trouble getting off the ground in the absence of a sufficiently well-formed definition of the target attribute, including an account of what it means for the attribute to vary (i.e., what meaning can be attached to claims about there being "more" or "less" of it, between and possibly within individuals) and how such variation is related to variation in the observed outcomes of the instrument (i.e., item response behaviour).

It is suggested in Section 3.2 that questions of this kind form part of what van Fraassen (2008) refers to as the *data model* for the target attribute. It is rather rare for psychometrics textbooks to devote much attention to these theoretical or conceptual issues, however. Often (e.g., Raykov and Marcoulides, 2011) it is stated that psychological and educational measurement is concerned with appraising how individuals differ with regard to hypothesized, but not directly observable, attributes or traits, such as intelligence, anxiety, or extraversion. It is assumed that these traits are in fact quantities (for instance Kline, 2000, p. 18) simply states that "the vast majority of psychological tests measuring intelligence, ability, personality and motivation … are interval scales"), and models are then introduced to relate them to observable data such as test or questionnaire responses in such a way as to enable the numerical latent trait parameters to be estimated, together with measures of precision such as standard errors—all conditional on the adequacy and plausibility of the model that has been assumed. Of course if the model is not adequate as a structural theory of the phenomenon itself, then results may simply reflect artifacts of the model (e.g., consequences—sometimes rather trivial tautologies—that follow from the metric structure of the real numbers), rather than corresponding to valid inferences with respect to the theory of the phenomenon.

Why should a phenomenon such as a learner's proficiency or competence in a particular domain be assumed to have the structure of a total order (let alone a quantity)? The reason probably goes back to a belief fundamental to the early development of psychometrics, that quantitative structure is necessary to enable measurement. For example, Thurstone (1928) claimed that

> When the idea of measurement is applied to scholastic achievement, … it is necessary to force the qualitative variations [in learners' performances] into a quantitative linear scale of some sort.

If "the idea of measurement" entails *locating a measurand at a point on the real number line*, then "forcing" observed qualitative variations to fit a quantitative structure is an understandable approach to adopt (even if it raises questions about validity). Indeed two common theoretical frameworks for psychological and educational measurement—the representational theory of measurement, and Rasch measurement theory—could be construed as concerned with ways to "force" qualitative variation into quantitative form: the former by aiming to define conditions under which qualitative observations can be mapped into numerical structures; the latter by rejecting observations that do not fit an assumed quantitative model. These approaches are unpacked a little in the next section.

2.2 Theories of measurement

2.2.1 The representational theory of measurement

Tal (2020), in his survey of the philosophy of measurement in science, describes the representational theory of measurement (RTM) as "the most influential mathematical theory of measurement to date". Wolff (2020), in a recent structuralist account of quantity and measurement, calls it "arguably the most developed formal theory of measurement". Michell (1990) claimed that it is "the orthodox theory of measurement within the philosophy of science".

The canonical text on RTM (Krantz et al., 1971, p. 9) takes *measurement* to mean "the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful".

RTM supposes that we are given an "empirical relational structure" (itself an abstraction of certain features of an "observed reality"). This structure consists of objects, relations between them, and possibly also ways of combining or composing them. For example in educational measurement contexts, we might take as objects students' responses to a writing task, and consider a binary relation $\succeq$ of *betterness* as being of interest (as in "student X's piece of writing is a better response to the task than student Y's: $X \succeq Y$"). Or we might be interested in how parts of a test or assessment combine (via a binary operation $\bullet$) to form an overall measure. For example, "correctly answering questions 3 and 4 demonstrates a higher level of proficiency than correctly answering questions 1 and 2": $q_3 \bullet q_4 \succeq q_1 \bullet q_2$. We might then wish to investigate whether these aspects of students' responses to tasks—this empirical relational structure—can be mapped to a numerical ordering or scoring system, in such a way that the structure is preserved (e.g., relative betterness between responses is mirrored by the relative magnitudes of the numbers assigned to those responses).

The idea is that if such homomorphisms can be shown to exist, then inferences in the numerical relational structure (normally taken to be the real numbers with the usual order relation $\geq$ and binary operations $+$ and $\cdot$) provide warrants for conclusions in the substantive domain of the empirical relational structure. If, further, we posit that differences in the observed outcomes of an educational assessment procedure, such as the administration of a test or examination, are *caused by* differences in the configurations, between learners, of their "underlying proficiency", then establishing a homomorphism between the empirical relational structure and the real numbers [i.e., establishing that the outcomes can be "placed on an interval (or ratio) scale"] serves to justify the assumption of quantitative structure for this assumed underlying proficiency trait, and hence to enable the measurement of each test-taker's proficiency by locating them at the point on the real line that corresponds to their level of proficiency.

2.2.2 Qualitative relational structures and testing for quantity

The adequacy of RTM as a theory of measurement has been extensively critiqued (see, e.g., Michell, 1990, 2021; see also Luce and Narens, 1994), with commentaries noting that its abstract nature sidesteps the actual process of measuring anything, the construction of measuring instruments, and any discussion of measurement error. The merits of such critiques are not discussed further in this paper, because the position adopted here will be that of Heilmann (2015). Heilmann (2015, p. 789) does not assess RTM as a candidate for a theory of measurement, but rather as a collection of mathematical theorems: theorems whose structure makes them useful for investigating problems of concept formation. He proposes viewing theorems in RTM as

> providing us with mathematical structures which, if sustained by specific conceptual interpretations, can provide insights into the possibilities and limits of representing concepts numerically

He regards RTM as studying not mappings from an empirical relational structure to a numerical relational structure, but rather from a *qualitative relational structure* (QRS) to a numerical relational structure. Taken in that sense, he argues, RTM can provide tools for testing the extent to which abstract concepts (captured or described as qualitative relational structures) can be represented numerically.[3]

Arguably, this is how RTM (including in particular the subset of RTM theorems that form the so-called theory of *conjoint measurement*: see Luce and Tukey, 1964) does in fact tend to be used in the literature exploring the plausibility of assuming quantitative structure for educational, psychological, or social measurands.

For example, Michell (1990) re-analyzed data collected by Thurstone (1927b) regarding judgements as to the seriousness of various crimes. Thurstone (1927a) claimed that his theory of *comparative judgement* enabled the construction of a *quantitative scale* for the measurement of seriousness of crime, by applying the theory to the outcomes of a collection of pairwise comparisons, in which subjects were repeatedly asked which of two crimes presented to them was the more serious. Michell (1990, p. 107) carefully stated the assumptions of Thurstone's theory, and demonstrated by applying results from RTM that "either seriousness of crimes is not a quantitative variable, or else some other part of Thurstone's theory of comparative judgement is false".

van Rooij (2011) applied theorems from RTM to explore whether properties of objects, that manifest linguistically as adjectives with comparative degrees, can be represented numerically, what scale properties may hold for them, and hence whether inter-adjective comparisons (such as "x is P-er than y is Q") can be meaningful. This is analogous to the vexed question, in educational assessment, of inter-subject comparison when it comes to setting and maintaining qualification standards (see, e.g., Newton et al., 2007; Coe, 2008).

Karabatsos (2001, 2018), Kyngdon (2011), Domingue (2014), and Scharaschkin (2023) applied theorems from RTM to the question of testing whether psychometric attributes comply with requirements for quantitative structure, combining the RTM results with a stochastic approach to address expected "measurement error" in most measurement scenarios with reasonable numbers of test-takers and test items. Domingue found that the results of a well-known test of reading showed that it was highly implausible that reading proficiency was a quantitatively-structured variable. Scharaschkin found that the results of a test of physics for school-leavers did not support the assumption of quantitative structure

3 A further extension of Heilmann's position would be to consider mappings from a QRS to another QRS: in other words, to relax the restriction that the "representing" structure should be numerical. Such a generalization might permit both RTM and van Fraassen's approach to be located, from a formal mathematical perspective, within the general theory of structure known as category theory, but will not be pursued here.

for a hypothesized "physics proficiency" construct. On the other hand, he found that the results of a similar test of economics were approximately consistent with an assumption of quantitative structure.

None of these applications require assuming the validity or adequacy of RTM as a substantive theory of measurement—indeed, Michell (2021) explicitly rejects it. Yet they do shed light on the extent to which qualitatively-structured data can be treated *as if* it were a manifestation of quantitatively-structured latent traits, and provide empirical evidence that it is not always valid to do so.

This is relevant to the practice of educational assessment and test construction because most practitioners and test developers probably do work within a pragmatic "as if" framework, as summarized by Lord and Novick (1968, p. 358):

> Much of psychological theory is based on trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behaviour.

Some of the ways in which theories of cognition have been more directly incorporated into the use of quantitative latent variable modeling, and their relation to the ideas considered in this paper, are discussed further in Section 5.4.

2.2.3 Rasch measurement theory

Psychometrics conducted in the Rasch measurement tradition (Andrich and Marais, 2019) takes the view that measurement is only meaningful for quantitative phenomena. Thus, if a putative measurement procedure such as an educational or psychological test yields results that are inconsistent with a underlying quantitative variable, then the procedure is not, in fact, *bona fide* measurement, and requires modification. In practice this means modifying tests by deleting or changing items until a sufficiently good fit to the Rasch model is obtained.[4]

So rather than trying to find a model that fits the data that has been obtained from the administration of a test, the Rasch measurement approach is to try to make the data fit the model. Modifying the measurement instrument to achieve this may come at the cost of severely constraining the theory of (or, in the terminology of Section 3.2, the relevant data model for) the substantive phenomenon or construct of interest. It might be that the construct cannot be sufficiently constrained or re-defined without significantly departing from its underpinning theory of value. In an educational assessment context, this would be the case if making such changes to the assessment instrument would compromise construct validity: the assessors' understanding of what constitute the key attributes of proficiency in the given domain, and how relatively better/worse/different states of proficiency would present with respect to these attributes. In such cases the choice would seem to be either to abandon the idea of measuring the construct at all, or to abandon the restriction of measurement to locating measurands within solely quantitative mathematical structures. This paper explores the latter option.

2.2.4 Measurements as ratios

Michell (1999) traces the evolution of the concept of measurement in psychology since the publication of Fechner's *Elemente der Psychophysik* in 1860. He bemoans the movement away from the conceptualization of measurement that had become standard in nineteenth century physics, namely (Michell, 1999, p. 14) "the discovery[5] or estimation of the ratio of the magnitude of a quantitative attribute to a unit (a unit being, in principle, any magnitude of the same quantitative attribute)". In other words, as elementary physics texts still state, physical quantity = real number × unit, where the real number is the measurement of the physical quantity.

Michell notes (p. 19) that "according to the traditional understanding of measurement, only attributes which possess quantitative structure are measurable. This is because only quantitative structure sustains ratios". He argues that, this being the case, it is incumbent on psychometricians to investigate whether the phenomena they study do, in fact, have quantitative structure, before applying statistical models that assume it. Since in practice this is almost never done, his claim is that, for the most part, "psychometrics is built upon a myth" (Michell, 2012). Once again, the choice appears to be to accept the constraints of the "traditional understanding of measurement", or to explore whether psychometrics could benefit from engagement with a more expansive conceptualization of what it means to measure something. The next section considers such a viewpoint.

3 van Fraassen's account of measurement

3.1 Basic principles and relevance to psychometrics

Bas van Fraassen's (2008) *Scientific Representation: Paradoxes of Perspective* is an empiricist structuralist account of measurement and representation in science. This stance eschews debate about the ontological status of the phenomena or reality that scientific theories describe, and concerns itself rather with elucidation of

4 The Rasch model, also known as the 1-parameter item response model, postulates that the log-odds of a test-taker of ability θ correctly answering an item of difficulty δ is simply $\theta - \delta$ (in the case of a test consisting of a sequence of dichotomously-scored items). There are of course other item response models that postulate additional item parameters, but Rasch theorists hold that the 1-parameter model is theoretically more appropriate as a basis for enabling measurement because it enables, within a given collection of persons and items, so-called invariant comparisons of persons (as to their ability) and items (as to their difficulty): see Andrich and Marais (2019, p. 80).

5 The development of quantum theory in the twentieth century problematized the classical epistemological viewpoint on measurement as "discovery". As Peres (1995, p. 14) observes, "classical physics assumes that the property which is measured objectively *exists* prior to the interaction of the measuring apparatus with the observed system. Quantum physics, on the other hand, is incompatible with the proposition that measurements discover some unknown but pre-existing reality."

what van_fraassen argues is the key aim of developing and testing such theories, namely their empirical adequacy. van Fraassen (2008, p. 2) claims that "measuring, just as well as theorizing, is representing … measuring *locates* the target in a theoretically constructed logical space". To be more precise (p. 164),

> measurement is an operation that locates an item (already classified in the domain of a given theory) in a logical space (provided by the theory to represent a range of possible states or characteristics of such items).

A key point here is the theory-relatedness of measurement procedures. Echoing Maul's (2017) requirements, quoted in Section 2.1, for a "well-formed definition of the target attribute" as fundamental to psychometric measurement, van Fraassen suggests (p. 166) that "once a stable theory has been achieved, the distinction between what is and is not genuine measurement will be answered *relative to that theory*".

It is argued in Section 4 that a candidate theory for the phenomena (proficiency or competence in a domain) that form the subject matter of educational measurement, is a description of what constitutes betterness between learners' possible states or configurations of proficiency in a given domain. "Betterness"—which, as noted in Section 2, may be a more general order relation than a simple ranking—has to be defined in terms of criteria that may, in general, be manifested with *fuzzy degrees of truth* in the responses of learners to tasks that have been designed to provide information about their proficiency in the domain in question.

van Fraassen considers several measuring procedures in classical and quantum physics (p. 157–172 and 312–316), and concludes (p. 172) that they are all "cases of grading, in a generalized sense: they serve to classify items as in a certain respect greater, less, or equal. But … this does not establish that the scale must be the real number continuum, nor even that the order is linear. The range may be an algebra, a lattice, or even more rudimentary, a poset". In fact, Section 4 below considers the case of lattices as logical spaces for educational measurement procedures.[6]

It is worth exploring how van Fraassen's approach could be applied to educational measurement for at least two reasons. Firstly because, as discussed in Section 2.2.2, the mathematically necessary conditions for a learner's proficiency in a given educational domain to have the structure of a quantity often do not hold; and it is not possible to massage the assessment instrument to make them hold without loss of construct validity. In such cases, it would arguably be inappropriate to theorize the construct as quantitative, and hence its measurement as location *on the real line*, rather than in some other, theory-relevant, logical space.

Secondly, the approach of thinking about educational assessment constructs in terms of fuzzy criteria of value (what will count as creditworthy, or indicative of good/bad performance, in relation to what particular domain content) is what *actually happens in practice*, when subject domain experts develop and administer at least one kind of high-volume, high-stakes, educational assessment procedure, namely the public examinations taken by school pupils aged 16 and 18 in the UK. This brings us to a consideration of what van Fraassen calls *data models*.

3.2 Data and surface models

Measurements arise from the results of procedures designed to gather information about a phenomenon of interest. As noted in Section 2.2.2, these entail selective attention to specific features that are deemed to be relevant. That is to say, measuring a phenomenon involves collecting data structured in a specific way. van Fraassen (2008, p. 253) calls such a structure a *data model* for the measurand in question. He notes that

> A data model is relevant for a given phenomenon, not because of any abstract structural features of the model, but because it was constructed on the basis of results gathered in a certain way, selected by specific criteria of relevance, on certain occasions, in a practical experimental or observational setting designed for that purpose.

In educational measurement we have gathered in a certain way (via an assessment procedure such as a test), selected by specific criteria of relevance (construct-relevant criteria: Pollitt and Ahmed, 2008) on certain occasions (at a particular point or points in time), in a practical setting designed for that purpose (e.g., the rules of administration and physical requirements for conducting an examination).

In the case where the test consists of a sequence of dichotomously-scored items $I := \{i_1, \ldots, i_n\}$ administered to a collection $L := \{l_1, \ldots, l_m\}$ of learners, we can think of this measurement setup as a map $V : L \times I \to \{0, 1\}$ that assigns to each instance of a learner encountering an item the valuation 1 if they answer it correctly, and 0 if they answer it incorrectly. Equivalently, we can think of the information collected by the assessment procedure as organized in an $m \times n$ matrix whose (m, n) entry is $V(l_m, i_n)$. There is, however, more structure entailed by the "betterness" ordering within each item (namely that "1" is better than "0") than immediately stands out from simply viewing the data as a table. As discussed in Section 4.2, the totality of the results-plus-valuation-system can be viewed as a lattice (the so-called *concept lattice* for the data table)—and it is suggested in Section 4 that such lattices (generalized to incorporate fuzzy valuations if necessary) form the natural data model for the phenomena that educational measurement procedures, such as tests and examinations, aim to measure.

van Fraassen (2008, p.253) describes constructing a data model as "precisely the selective relevant depiction of the phenomena *by the user of the theory* required for the possibility of representation of the phenomenon." In the context of educational testing, the proficiencies being studied are proficiencies or competencies *with respect to* a specified domain (such as "high school chemistry", or "A level French"). What "good performance" or "good demonstrated attainment" looks like in these domains (and hence what would count as evidence of better or worse levels, or states, or configurations, of learners' *proficiencies*) is always subject to a prevailing understanding or agreement as to what potential aspects

6 Algebras, lattices, and posets (short for partially-ordered sets) are types of mathematical structures. In particular, a lattice is a partially-ordered set (see the Appendix for a definition) in which each pair of elements has a least upper bound and a greatest lower bound.

of the domain are chosen as relevant for discrimination between learners' performances as to their quality. In other words, the criteria for creditworthiness of candidates' responses to tasks in an assessment can be regarded as the selective relevant depiction of the phenomenon of interest, by those members of the competent authority (the "users of the theory") who design, administer, and grade the tests. For that reason, concept lattices derived from the outcome data from the tests, that encode the relationship between learners and the assessment criteria, are appropriate data models.

In practice, van Fraassen (2008, p.167) notes that data models may be "abstracted into a mathematically idealized form" before empirical or experimental results are used to explore theories or explanations, or for substantive purposes. He gives the example of a data model consisting of relative frequencies, which is "smoothed" such that frequency counts are replaced with probabilities. An idealized or simplified version of a data model is called a *surface model* for the phenomenon in question. Surface models are considered further in Section 5.

4 Theories of constructs: comparing item response theory and fuzzy concept analysis

4.1 A small example

Table 1 shows results from an assessment that generates data on each of three items (or attributes) $\{i_1, i_2, i_3\}$ for six learners $\{l_1, \ldots, l_6\}$. Here 0 means "not demonstrated", $\frac{1}{2}$ means "partially demonstrated", and 1 (or $\frac{2}{2}$) means "fully demonstrated".

A traditional psychometric approach to analyzing this kind of data would be to treat each learner's results from the assessment as a vector in $\mathbb{R}^3$, and each learner's proficiency measure as a quantity (a point in $\mathbb{R}$). For example, we could treat the label for each item response category as a number, and add them to get a total score for each learner. This orders learners, with respect to proficiency, equivalently to fitting a Rasch model (a 1-parameter item-response model), since total score is a sufficient statistic for estimating proficiency in this model. Or we could do a principal components analysis and take the projection of each learner's item-response vector onto the component that accounts for the most variance as their proficiency measure (this is equivalent to fitting a 2-parameter item-response model: see Cho, 2023). Doing so for the data in Table 1 yields three components of which the first accounts for 72% of the variance in outcomes, with the other two accounting for 19 and 9%, respectively. We could therefore take the loading (projection) of each learner's results onto the first component as their score on an "underlying" quantitative variable that represents the assessment construct reasonably well. Figure 1 shows how learners' proficiency measures differ depending on the approach taken.

However, in view of the problems associated with assuming quantitative structure for proficiency discussed in Section 2.1 (tantamount, in Section 3.2's terms, to replacing the data model with a radically different surface model), let us consider a non-quantitative approach. If we take each learner's test response not as a vector of numbers, but rather a vector of ordered labels, then

TABLE 1 Data from a test.

\	i_1	i_2	i_3
l_1	0	$\frac{1}{2}$	$\frac{1}{2}$
l_2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
l_3	1	1	$\frac{1}{2}$
l_4	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
l_5	0	$\frac{1}{2}$	0
l_6	$\frac{1}{2}$	1	1

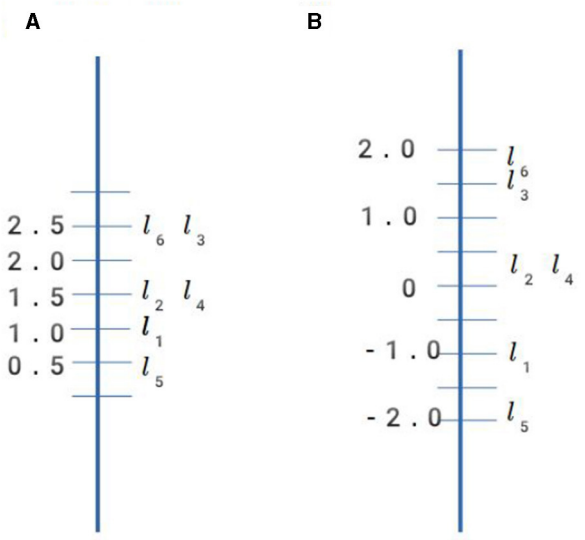


FIGURE 1
IRT-derived proficiency measures. **(A)** Sum score. **(B)** Latent variable score.

the observed data can be characterized as a collection of partially-ordered nodes: a network of "betterness" relations between nodes. In this data model, shown in Figure 2, each node is a *type of performance* on the assessment.

Each type of performance is defined by a collection of *attributes*, that *characterize* it; or (dually) by a collection of *learners*, who *demonstrate* it. The boxes in Figure 2 are the different types of performances on the test. The best performance is at the top of the diagram, and the worst performance at the bottom. Attributes, and learners, may belong to nodes to a *fuzzy degree*. Thus learner 5 belongs to (demonstrates) the lowest type of performance completely (to degree 1). Learners 2, 3, 4, and 6 all demonstrate the highest type of performance to degree 0.5.

Better types of performance are characterized by showing *more* attributes (and, dually, are demonstrated by *fewer* learners) than worse types of performance. An arrow from a box A to a box B means that B is a better performance than A (and by extension better than any performance C such that there is a connected path from C to A). If there is no path between two types of performance, then they are not comparable. Locating a learner (measuring their proficiency), with respect to this data model for the construct which the three-item test aims to assess, then means finding the "highest"

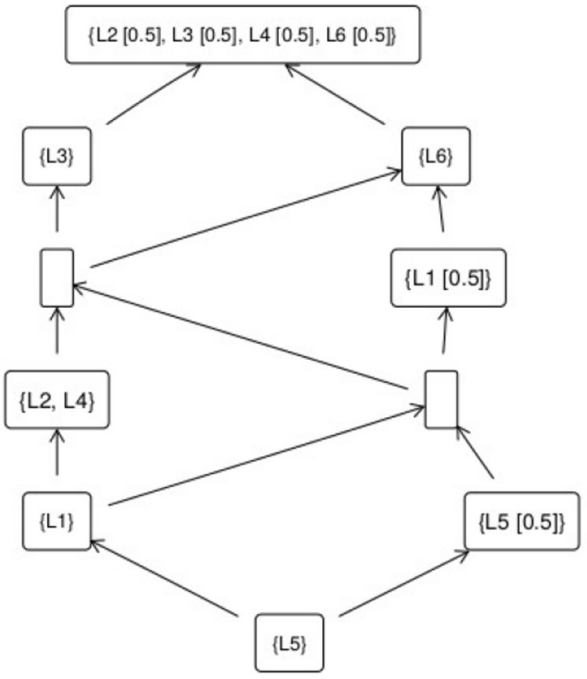


FIGURE 2
Fuzzy concept lattice for assessment data.

node that they belong to in the network. This intuitive description is made more precise in the following section.

4.2 Formal concept analysis and proficiency measurement

Formal concept analysis (Ganter and Wille, 1999; Carpineto and Romano, 2004) is an important development of mathematical order theory that has been applied extensively to fields such as linguistics, political science, information sciences, medicine, and genetics. A recent application (Bradley et al., 2024) is to elucidating the mathematical representation of structure in large language models such as ChatGPT, discussed briefly below in Section 6. It can be thought of as a way of making explicit the information structure that is implicit in a matrix—such as that in Table 1—which relates objects to attributes (or learners to test items). It provides methods to extract the concepts and implications that can be deduced from such data, and introduces a logic to reason and infer new knowledge.

Consider first the case of measuring proficiency in a domain by administering an n-item test to m learners, where each item is dichotomously scored, i.e., for each learner l and item i, it is either the case that l answered i correctly, or that l did not answer i correctly. Given a subset of learners $L_1 := \{l_1, \ldots, l_k\}$, let $I_1 := \{i_1, \ldots, i_j\}$ be precisely those items that all learners in L_1 got correct. Then the pair (L_1, I_1) is an instance of a *formal concept* present in the data. L_1 is called the *extent* of the concept, and I_1 is called its *intent*. We can equally well start with a subset $I_2 := \{i_1, \ldots, i_p\}$ of items, and then form the concept (L_2, I_2), where L_2 is precisely the set of learners who got all items in I_2 correct.

The collection of all formal concepts extracted from a matrix or data table simply restates the information present by virtue of the way the data is structured due to the choice of attributes (test item responses, in this example), and the ordered valuations chosen for attributes (just the two categories $1 \geq 0$ in this case). However, it makes this structure more apparent (and graphically representable, as in Figure 1) because concepts are (partially) *ordered* via the set-theoretic notion of inclusion. A concept (L_1, I_1) is *more general* than a concept (L_2, I_2) if $L_1 \supseteq L_2$ (or equivalently, if $I_1 \subseteq I_2$). The most general concept is the one that has the largest extent (and smallest intent). In test performance terms, the most general concept corresponds to the bottom, or worst, performance: because every other performance has a larger intent (entails more correct items). Similarly, the least general concept (with the smallest extent and largest intent) corresponds to the top, or best, level of performance.[7]

We can think of formal concepts as different ways of performing on the test (i.e., different ways of exhibiting proficiency in the subject domain). Each type of performance—or exhibition of proficiency—can be described *extensively*, by showing the learners who demonstrated it. Or it can be described *intensively*, by showing the item-profiles that characterized it. These two modes of presentation correspond to different ways of training "measuring instruments" (traditionally, human judges; more recently machine-learning methods such as neural nets) to recognize what good/bad performance (high/low proficiency) looks like. One can either give *examples* of a certain kind of performance, until an assessor can correctly classify new instances, or one can give *descriptions* of that kind of performance (in this case, the relevant profile of item responses), to enable new instances to be classified (measured) correctly.[8]

For a small educational measurement procedure of this kind (small in terms of the number of items/tasks/relevant attributes on which data is collected, as well as small in terms of the number of subjects to which it is administered), the qualitative equivalent of a quantitative score is a learner's location in the concept lattice: the highest concept, in the partial order, to whose extent they belong. This level of proficiency is described, not as a numerical "amount" (location on a line), but rather by the intent of the relevant concept: the actual items they mastered (or, more generally, the construct-relevant attributes

7 Normally concept lattices are drawn as so-called Hasse diagrams with the least general concept at the bottom, and the most general concept at the top. An arrow is drawn upwards from concept A to concept B if B is more general than A. In the educational assessment context, we naturally regard the best performance as the *top* concept, which means we need to reverse the usual ordering (in mathematical terms, we use the *dual* lattice). This is done throughout this paper, for example in Figure 2, where the worst level of proficiency (exhibited, to degree 0.5, by learner l_5) is at the bottom of the diagram, and the best level (exhibited by learners l_2, l_3, l_4, and l_6, also to degree 0.5) is at the top.

8 As Weyl (1952, p. 8) noted, "For measurement the distinction is essential between the 'giving' of an object through individual exhibition on the one side, in conceptual ways on the other".

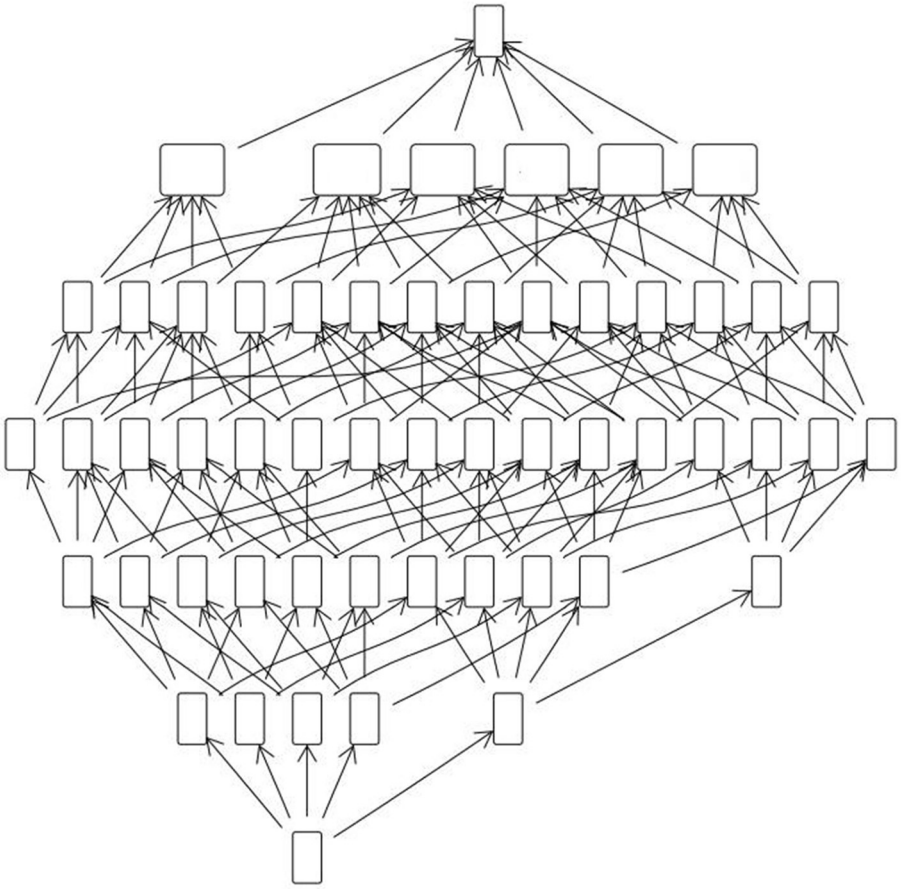


FIGURE 3
Concept lattice for a 5-item test with 100 learners.

their performance demonstrated). For larger (more realistically sized) assessments, the concept-lattice data-model becomes too granular, as shown in Section 5, and we develop a notion of "prototypical" kinds of performances at a manageable number of levels, such that each learner's level, or state, of proficiearency can be described approximately in terms of its qualitatively closest prototype.

Before moving on to that discussion, it is necessary to consider the question of the fuzziness of the criteria that structure data models in many educational measurement procedures.

4.3 Truth degrees and fuzzy concepts

4.3.1 Assessment results as truth degrees

Table 1 illustrates a situation that often obtains in educational assessment. Learners are given tasks, such as questions on a test, and they may be successful in engaging with them *to a certain degree*. The outcome of a learner's interaction with an item is not necessarily captured by the crisp dichotomy of {correct, incorrect}.

The usual way of dealing with this in psychometric models is to model response categories for polytomous items as a sequence of threshold points on a latent quantitative continuum. A learner's response is in a higher category if it results from their proficiency-state being higher than, but not otherwise different from, a learner whose response is in a lower category. Differences in proficiency must be conceived of as differences in degree, not in kind. Yet as Michell (2012, p. 265) notes, in the context of mathematics tests, "the differences between cognitive resources needed to solve easy and moderately difficult items will not be the same as the differences between resources needed to solve moderately difficult and very difficult mathematics items. This observation suggests that abilities are composed of ordered hierarchies of cognitive resources, the differences between which are heterogeneous."

An alternative approach is to start by the viewing the dichotomous situation as providing information about learners' performances in the form of *propositions* of the form "learner l answered item i correctly".[9] This proposition is true just in case the (l, i) entry in the data table arising from the assessment is 1. So we can think of the entries in the table as truth values (with 0 meaning false and 1 meaning true).

9 As Michell (2009) observes, "Tabulated numbers are shorthand for a set of propositions that tell where the numbers came from. Furthermore, deductions from a data set are inferences from these propositions."

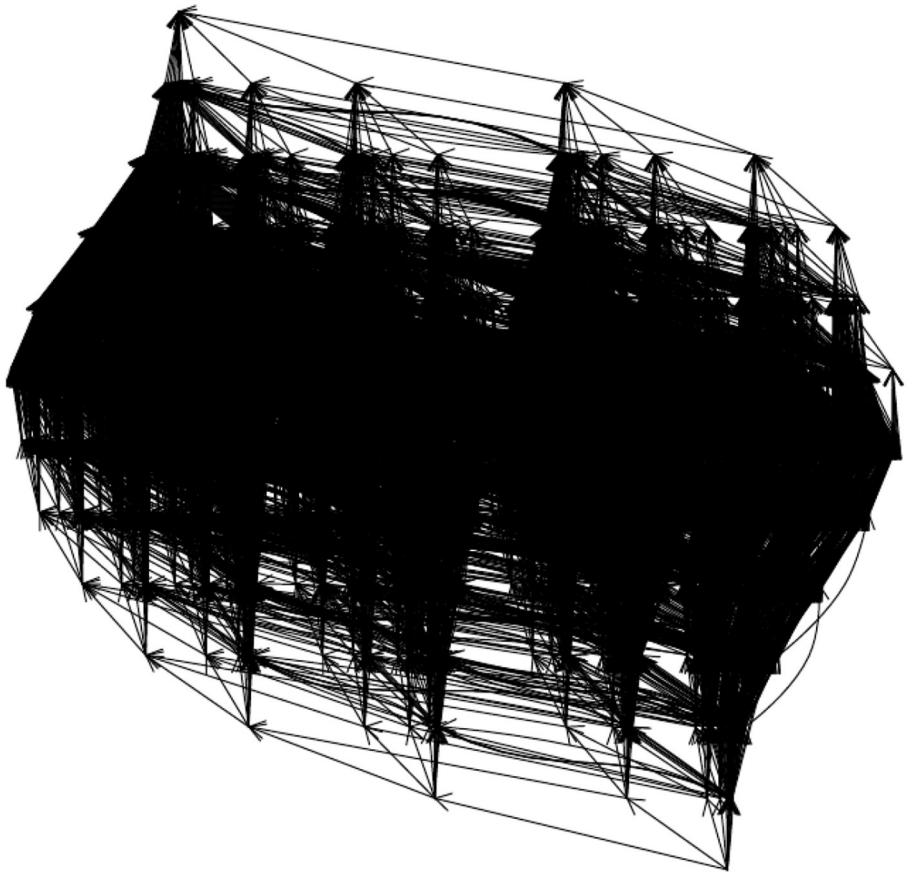


FIGURE 4
Concept lattice for a 12 item test with 200 learners.

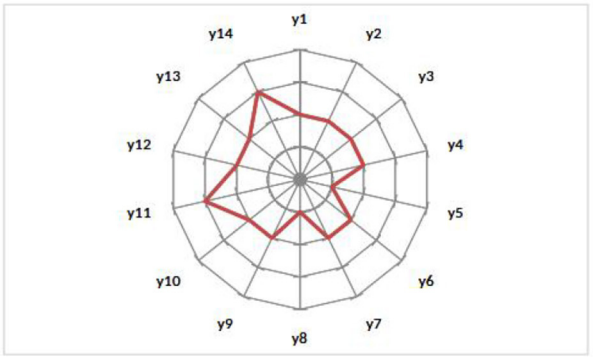
It has long been recognized that, in situations in which there is inherent fuzziness, vagueness, or semantic uncertainty in concepts, bivalent logics, in which the only possible truth values for a proposition are {`false`, `true`} can be unduly restrictive (see e.g., Goguen, 1969; Goertz, 2006; Bělohlávek et al., 2017). *Fuzzy logic* (Hajek, 1998; Bělohlávek et al., 2017) allows propositions to have truth values drawn from ordered sets of *truth degrees*, that can be more extensive than {`false`, `true`}.

Thus we can view the example in Table 1 as providing information about propositions with three truth-degrees, that we could label $\{0, \frac{1}{2}, 1\}$, or {`false`, `partially-true`, `true`}. For example, it is false that learner l_1 demonstrated attribute i_1 (or we could say, she demonstrated it to degree 0), and it is partially-true that she demonstrated attribute i_2 (she demonstrated it to degree $\frac{1}{2}$).

When the outcomes of educational measurement procedures are not completely and crisply dichotomous with respect to all the construct-relevant attributes about which information is collected, the concept lattice for the resulting matrix of fuzzy truth values is itself fuzzy. Objects and attributes belong to concepts with degrees of truth, rather than crisply. In the concept lattice in Figure 2, the label "0.5" after a learner-identifier means that learner belongs to the concept (i.e., has demonstrated that type or level of performance) to degree $\frac{1}{2}$).

Although a discussion of the concept of "measurement error" in psychological testing and educational assessment would take us beyond the scope of this paper, it may be worth clarifying, for the avoidance of doubt, that the application of fuzzy logic in this context is not simply an alternative to using probability theory. Probability is a tool that can be used to study (epistemic) *uncertainty* (the lack of precision that arises from incomplete or poor information), whereas fuzzy logic is a tool that can be used to study (ontological) *vagueness* (the inherent fuzziness, or necessary inexactness, of concepts like "proficiency" in a certain domain). Erwin Schrödinger, when considering what the development of quantum mechanics meant for the measurement of physical phenomena, distinguished these two facets when he noted (Trimmer, 1980; p. 328) that "There is a difference between a shaky or out-of-focus photograph and a snapshot of clouds and fog banks".

The statement "Mary has a fairly good understanding of physics" is vague but certain, whereas "Mary will pass the physics test tomorrow" is precise but uncertain. Working with propositions such as the former (i.e., deploying what Goguen, 1969 calls a "logic of inexact concepts") is core to educational assessment,



Intent of largest factor: 'a medium/reasonable performance across the content and assessment objectives, but characterised by good communication skills on the longer essay items, and less so on the short answers'

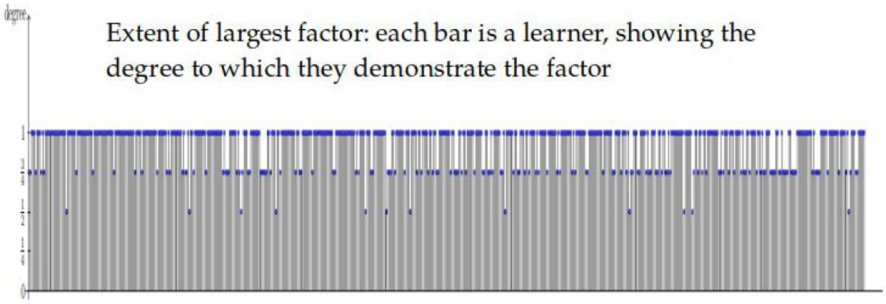


FIGURE 5
Factor representation for fuzzy data.

because of the contestable and intersubjective nature of educational constructs, discussed further in Section 7.2.

4.3.2 Truth degrees and quantities

Buntins et al. (2016) apply fuzzy logic to psychological tests in a somewhat different way to that proposed here. They take the view that scores obtained from a test should not "refer to latent variables but to the truth value of the expression 'person *j* has construct *i*", where a *construct* is defined by a collection of relevant *attributes*, each of which may be *possessed* by a test-taker to a certain degree, and each of which may be *relevant* for the construct to a certain degree. Modeling truth degrees as real-valued quantities in the interval [0,1], they present an algorithm for aggregating them across attributes to arrive at an overall score for each learner: the truth value of the proposition "this learner has the construct". They are careful to distinguish the semantic vagueness of a construct definition (recognized in the use of fuzzy truth values) from the idea of "measurement error".

Buntins et al. claim that this approach "neither relies on latent variables nor on the concept of [quantitative] measurement". However, they do state it is arguable that "although there is no measurement theory involved in the ... formalism, the application to actual test behavior does presume item answers to be assessed on an interval scale level", because "test answers have to be real numbers between 0 and 1, reflecting the subjective truth-values of the corresponding attributes for the tested person … However, these only refer to the item level and do not extend to theories about latent variables."

In fact truth degrees do *not* have to be real numbers between 0 and 1. What is required is that they have a way of being compared with each other—that is, an order structure (which could be a partial order)—and way of being combined with each other. In general these requirements are met by taking them to have the mathematical structure of a so-called complete residuated lattice (Hajek, 1998). Further work on conceptualizing truth degrees—and especially what that means for empirically eliciting them—is important, as touched on in Section 7, but beyond the scope of this paper.

Buntins et al. see their approach "not as opposed to psychometric theory but tr[ying] to complement it with an alternative way to conceptualize psychological tests". By contrast, the approach presented in this paper is suggested not as an alternative to, but an extension of, psychometric theory: one in which quantitative measurement forms an important, but special, case of a more general measurement framework.

4.3.3 Fuzzy relational systems

In summary, the argument in this section is that in general, educational assessment procedures that aim to measure constructs such as proficiency, ability, or competence in a fuzzily-defined domain, generate *fuzzy relational systems*: matrices of truth-values for propositions of the form "learner *l* has demonstrated construct-relevant attribute *i*". As data models, these are equivalent to fuzzy concept lattices: partially-ordered hierarchies, or networks, of types of performance on the assessment, that are discriminable with respect to these construct-relevant attributes. The next section

considers whether these data models can provide insight for realistically-sized assessments.

5 Practicalities of educational assessment with non-quantitative data models

5.1 Granularity of data models

An issue with data models of the kind discussed in the previous section is that their combinatorial complexity increases geometrically with the numbers of learners and construct-relevant attributes of performance (or test items) involved. Figures 3, 4, for instance, show the concept lattices for subsets of outcomes of a physics test.[10] with increasing numbers of learners and attributes. Clearly the information here is too granular to be useful, and we need to simplify or "smooth" it in some way.

For quantitative data models, where learners' test responses as thought of as vectors in n-dimensional Euclidean space, the analogous granuarity-reduction is often performed using latent variable models that aim to find a k-dimensional subspace with $k < n$ (often a one-dimensional subspace, i.e., a line) that is oriented in such a way as most closely to approximate the direction of most of the variation between the positions of these points (possibly subject to some other constraints as well, for certain factor-analytic models: see Bartholemew et al., 2008). Each learner's latent-variable score is then the projection of the vector that represents their test performance onto this subspace. Calculating these scores entails factorizing the (transpose of the) matrix Z of normalized test scores. If there are m learners and n test items, then the $n \times m$ item-by-learner matrix Z^T is factorized into the product of a $n \times k$ item-by-factor matrix L and a $k \times m$ factor-by-student matrix F, plus some error: $Z^T \approx LF$. Then using standard results in linear algebra, it can be shown (e.g., Reyment and Jöreskog, 1993) that the factors are the eigenvectors of the covariance matrix ZZ^T.

5.2 Factorizing qualitative matrices

Bělohlávek (2012) studied the question of factorizing a matrix of fuzzy truth values. Now the matrix product is no longer defined in terms of operations on quantities, but rather in terms of operations on truth values.[11] Let M be an $m \times n$ matrix arising from an educational measurement procedure conceptualized as in Section 4.3, so that M_{ij} is the degree to which learner i displays attribute j. By analogy with the quantitative case, consider an approximate factorization of M into a $m \times k$ learner-by-factor matrix A and a $k \times n$ factor-by-attribute matrix B, i.e., $M \approx A \circ B$. The key theorem in this case, due Bělohlávek (2012), is that *the factors are particular formal concepts* from the concept lattice for M. That is, "picking out key concepts" (particular types of learners' responses to the assessment) is equivalent to "logically factorizing" the matrix of truth-degrees that is the outcome of the measurement procedure.

The factors are the (extents and intents) of specific concepts in the concept lattice for M. The intuition is that, with $M_{ij} = A_{ip} \circ B_{pj}$:

- A_{ip} is the degree to which learner i is an example of (in the extent of) factor p;
- B_{pj} is the degree to which attribute j is one of the manifestations of (in the intent of) factor p;
- $M = A \circ B$ means: learner i displays attribute j if and only if there is a factor (formal concept) p such that i is an example of p (or p applies to i); and j is one of the particular manifestations of p.

Thus, the qualitative analog of projecting a Euclidean space onto a lower-dimensional subspace consists in picking out certain points in a partially ordered set. Specific formal concepts are selected, similarly to the way in which specific vectors—the eigenvectors of the covariance matrix—are selected when learners are scored on quantitative latent variables. The analogs of scores on a latent variable are the degrees to which learners' performances "display" or "participate in" or "reflect" these specific concepts, which may be thought of as *prototype* or *standards of performance* on the construct. They have the advantage, over hypothesized latent variables whose values are abstracted from observed data, that they are directly expressible in terms of the construct-relevant attributes—that is, in terms of the features of learner's responses to assessment tasks that are taken to be important in a "theory" of "what (good) performance means", for the educational construct in question. They can be described both by means of their extent (the collection of actual learners' performances exemplifying the concept/standard in question), and by means of their intent [the collection of (fuzzy) attributes that characterizes the standard in question].

5.3 Measures and meanings: comparing quantitative and qualitative approaches

Bartl et al. (2018) examined this qualitative factor analytic approach to educational assessment data, with the aims of exploring its applicability in practice, and its application to the study of the construct validity of an examination: the degree to which students' responses, assessed as being at a particular level, matched the intentions of the assessment designers in terms of the qualitative performance standard intended to broadly characterize responses at that level. This is the kind of question that is difficult to study using traditional quantitative methods.

10 Part of paper 1 of the AQA A level physics examination taken in 2018. Unusually for an A level assessment, the items here are all dichotomous (multiple-choice questions). The lattices would be even larger if the items admitted fuzzy valuations.

11 The product of two real-valued matrices A and B is defined by setting its (i,j) entry $(AB)_{ij}$ to the inner product of row i of A with column j of B: i.e., $(AB)_{ij} := \sum_{p=1}^{k} A_{ip}B_{pk}$. When the matrix entries are truth values, they are elements of a type of lattice that is equipped with an operation $\otimes$ to combine values. In this case the matrix product $A \circ B$ is defined as $(A \circ B)_{ij} := \bigvee_{p=1}^{k} A_{ip} \otimes B_{pk}$, where $\bigvee$ is the supremum over the indicated set (see Appendix).

The technical issues involved (for example how to determine the coverage and number of factors that broadly explain the data—analogous to a scree plot in quantitative principal components analysis) will not be rehearsed here. See Bartl et al. (2018) for computational details. For a deeper theoretical treatment of the relationship between eigenvectors (of quantitative covariance matrices) and formal concepts (of qualitative matrices of truth values), see Bradley (2020). The key point is that this approach allows drawing out key features associated with responses assigned to a particular level, by the assessment procedure, and an appraisal of the degree to which each learner's performance on the examination embodies or matches those features. Indeed, it "explained" the data (in terms of proportion of data covered or variance explained) as well as standard principal components analysis, but generated factors exemplifying attributes of performance that seemed to be more easily interpretable.

Figure 5 shows an example of this, for the educational measurement data studied by Bartl et al. (2018), in which learners were assessed on 14 fuzzy attributes $\{y_1, \ldots, y_{14}\}$, each of which reflected an aspect of the construct, in this case proficiency in the specific subject of "A level Government and Politics". Each of the attributes corresponds to demonstrating specific types of knowledge and understanding, in accordance with the examiners' agreed understanding of what better/worse proficiency means in this domain. Hence the intent of any given concept can be interpreted by users of the assessment as a description of broadly what that level of proficiency means (and likewise the extent of the concept can be interpreted as an indication of the degree to which each learner has demonstrated that level of proficiency).

The question of the interpretability or explainability of the results of educational measurement procedures—whether those results are numerical scores, or broader grades or levels—is particularly important for high-stakes assessments such as those that underwrite school-leaving qualifications. For learners, clarity about *why* their response to an assessment merited their being characterized as demonstrating a certain level of proficiency is arguably required for reasons of natural justice. For teachers, understanding qualitatively what their students did well, and what they would have to do better to demonstrate more proficiency in a subject domain, is clearly valuable as an input into their future pedagogical practice. Bartl et al. (2018, p. 204) concluded that their approach to qualitative factor analysis yielded "naturally intepretable factors from data which are easy to understand", but that more research is needed both on technical implementation and on the views of learners and teachers.

5.4 Other order-theoretic approaches to educational assessment

In the 1940s Louis Guttman began to develop an approach to psychological measurement (e.g., Guttman, 1944) that led him to think of it as a structural theory (Guttman, 1971), rather than as a process of quantifying amounts of latent traits, and to the development of *facet theory* and *partial order scalogram analysis* (Shye and Elizur, 1994). In the 1980s, Doignon and Falmagne (1999) developed *knowledge space theory*, later evolved into a theory of learning spaces, in which assessment constructs are represented as partially-ordered sets.

Applications of facet theory and knowledge space theory (including related approaches such as Tatsuoka, 2009's *rules space* and Leighton and Gierl, 2007's *cognitive diagnostic models*) normally assume or overlay quantitative latent variable models, to account for "underlying" proficiencies or competencies that determine a learner's progression through such partially-ordered outcome spaces.

However, from the mid 1990s onwards, there has been a strand of research investigating how to extend knowledge space theory to incorporate a focus on skills and competence, leading to the development of *competence-based knowledge space theory* (see e.g., Stefanutti and de Chiusole, 2017). Here, a learner's proficiency or competence is itself conceptualized as a partially-ordered space, rather than a quantity. Ganter and Glodeanu (2014) and Ganter et al. (2017) suggested that formal concept analysis could be applied to study competence-based knowledge space theory, and this is now starting to be done.

For example, Huang et al. (2023) consider how to transform maps from competence-states to "knowledge-states" (types of demonstrated performances) into formal contexts, and hence to represent them as concept lattices. Each node in the lattice then embodies a knowledge-state and a competence-state as its extent and its intent, respectively. This is clearly analogous to the approach set out in Section 4 above.

A very clear application of these methods is to formative, adaptive, assessment and learning systems, where, for instance, they provide an alternative to traditional IRT-based adaptive tests that is more grounded in a theory of learning.

To date there has been less attention to examining summative assessment, and what is often called "educational measurement", from this perspective. Yet, as argued above, application of non-quantitative approaches needs to be investigated here too, since the pragmatic "as if" approach to routine application of latent variable models is not always justifiable.

6 Connections to artificial intelligence

A final reason why it is imperative to pursue research in this area is the rapidly growing application of machine-learning methods, and generative artificial intelligence in particular, in educational contexts. For example, Li et al. (2023) report on using the large language model ChatGPT to score students' responses to (essay style) examinations, and to provide rationales for the scores awarded.

Because the outputs of generative AI applications using large language models are no more than statistically plausible sequences of words, albeit expressed in well-formed natural language, their validity, fairness and reliability is hard to establish theoretically. That is because they are produced using so-called *subsymbolic* approaches to AI (see e.g., Sudmann et al., 2023), such as deep neural nets, rather than *symbolic* methods that aim to use forms of explicit logical inference to arrive at results: analogously to reasoning about a learner's response to a task with reference to criteria for betterness that define the kind of proficiency one intends to measure by administering the task.

An interesting angle opened up by the qualitative measurement approach described above is the possibility of combining formal concept analysis with neural networks to enhance the explainability of, for example, scores derived from applying a classifier based on a large language model to learners' performances on an examination.

Some initial work in this area has been done by Hirth and Hanika (2022) and Marquer (2020), among others. This kind of analysis could complement quantitative approaches to explaining marks or scores awarded to learners' responses, such as dimension-reduction of the high-dimensional vector space that the language model uses to represent linguistic artifacts—such as learners' responses to assessment tasks—as numerical vectors. In fact, Bradley et al. (2024) have recently shown that there is a relationship between quantitative techniques based on linear algebra, such as latent semantic analysis, and formal concept analysis, such that the latter can be seen as a more general form of the former. They have applied formal concept analysis to elucidating how semantics appears to arise from syntax, and to study the structure of semantics, when large language models are used to produce outputs from qualitative data.

Clearly, the practice of educational (and psychological) measurement is changing as technology changes. Tasks can be administered digitally; the widespread availability of devices with reasonable processing power means the possibilities for task design are much more open than they were a decade ago, and they will continue to evolve. The data that is gathered about learners, given their responses to these tasks, can be more unstructured than category-labels or scores: it may be text, audio, or video, and/or representations of such data for example in a vector-space language model. To the extent that human assessors form part of measurement procedures, for example to apply scoring rubrics, they may be partially or wholly replaced by AI.

What remains fundamental, however, is the need to base these measurement procedures in a theory of what defines or constitutes better or worse proficiency, in the domain of interest, and hence what substantive and semantic content is entailed in statements such as "this learner got a score of 137", or "this learner has 1.07 logits of proficiency"; or "this learner has demonstrated three of the four prototypical aspects of proficiency that define a "grade B standard", or whatever — what it means to locate them, via a measurement, at a certain position in a (quantitative or other) space.

7 Discussion

7.1 Qualitative educational assessment is possible in principle, and includes quantitative measurement as a special case

This paper has argued that it is not warranted to assume the phenomena studied in psychometrics, and in educational measurement in particular, are necessarily appropriately conceptualized as quantities. In cases where an assumption of quantitative structure *is* appropriate, then measuring an instance of such a phenomenon means locating it at a point on the real continuum. In cases where the assumption is not appropriate, the idea of measurement becomes, more generally, locating the measurand in a suitable logical space, that is defined in a way that is relevant for the phenomenon.

When the measurand is quantitative and the logical space is the real numbers, the usual methods of psychometric analysis for estimating latent parameters can be deployed. But, *contra* Thurstone (1928), the paper has argued that it is not necessary to "force" theoretically well-supported constructs into a more reductive quantitative form if that is not appropriate. Hence the argument of this paper is not that psychometrics should be replaced, but that its repertoire of measurement approaches should be widened to cope with measurands that are intrinsically non-quantitative in nature.

The paper suggests that the outcomes of educational measurement procedures can be thought of, in general, as fuzzy relational systems; and that fuzzy formal concept analysis is an appropriate tool to describe data models for the measurands they aim to locate. These models instantiate the "betterness" relation for the measurand: they model the notion of "what good performance looks like". Such an account or understanding is prior to, and necessary for, an understanding or agreement as to "what being (more or less) proficient" means, in an educational domain. It forms the theory of the construct (one might say, the theory of *value* for the construct, and hence a foundation for evaluation of construct *validity*).

7.2 Educational constructs are contestable, intersubjective, temporally-located phenomena

These theories of constructs such as proficiency or competence in a domain are necessarily contestable, intersubjectively constructed, and liable to change over time. Intersubjectivity (Chandler and Munday, 2011) refers to the mutual construction of relationships through shared subjectivity. Things and their meanings are intersubjective, within a given community, to the extent that the members of the community share common understandings of them. Thus, the community that constitutes the competent authority for defining an educational construct decides what particular knowledge, skills, and understanding it will encompass, and what will count as better or worse configurations of these aspects as possible ways of being proficient in the domain in question. Thus, for instance, the job of someone marking responses to an examination that is designed to measure that construct is to apply the mutually constructed and agreed standard consistently to each response she marks (irrespective of whether she personally agrees that it is the "right" standard).

We do not have to think of data models that encode these intersubjective constructions as (more or less accurate) representations of some objective or underlying "true" account of the measurand in question. As van Fraassen (2008, p. 260) notes, "in a context in which a given model is *someone's* representation of a phenomenon, there is **for that person** no difference between the question *whether a theory fits that representation* and the question *whether that theory fits the phenomenon*."

7.3 More research is needed on using partial orders in practice, on linking different assessments of the same construct, and on fuzzy valuations

Section 4 argued that in general the data models for measurands such as proficiency in an educational domain are partial orders. This perhaps goes against a relatively strongly ingrained concept of educational assessment as synonymous with *ranking* (e.g., Holmes et al., 2017). Yet in many cases, once a theory of (betterness for) a construct has been settled, rankings are neither necessary nor needed. Two learners' proficiency values may simply be qualitatively different (non-comparable). For instance in Figure 2, this is the case for learners 3 and 6. But both learners 3 and 6 have performed better than learner 1. So if learner 1's performance was sufficient to merit a "pass" grade, let us say (or was picked out as a "pass" grade prototype), then we know that learners 3 and 6 are also sufficiently proficient to be awarded a pass, even though it is not meaningful to say that their actual demonstrated proficiencies were the same, or that either one is more or less proficient than the other. More work is needed on the scope for using visualizations such as concept lattices to help educational assessment designers and teachers engage with and interrogate the outcomes of educational measurement procedures (see, for a start, Bedek and Albert, 2015).

A common application of quantitative latent variable models is to *equating* or *linking* different forms of tests of learners' proficiency in a certain domain. Typically, equating studies are designed to answer questions like "what score on form X of a test is equivalent to (represents the same level of proficiency as) a given score on form Y of the test?". In practical applications in many educational contexts however, such as grading students' responses to school-leaving examinations (Newton et al., 2007), one is not so much interested in constructing a monotone map from scores on X to scores on Y, as in ensuring that the levels or kinds of proficiency demonstrated by students graded, say, A, on this year's examination, are "equivalent", or "of a comparable standard" to the type of proficiency demonstrated by students graded A on last year's examination.

An area for further research is how to implement such comparability studies in the fuzzy-relational approach to educational assessment proposed in this paper. For example one could take the students graded A on each of the two forms of an assessment, and examine the intents of the formal concepts that form their largest factors (cover an appreciable proportion of the data, in the terms of Bartl et al., 2018). Are these sufficiently similar to count as equivalent demonstrations of proficiency, and what criteria should be applied to appraise similarity?

A deeper question is how the truth degrees that summarize each learner's demonstration of each construct-relevant attribute are determined. In some cases this is straightforward in practice (e.g., for dichotomously-classified test items such as multiple-choice questions); but when judges are needed as part of the measurement procedure, different judges may give different truth values, so what counts as a reasonable or acceptable value? A full account of this aspect of qualitative valuation may need to draw on *rough fuzzy logic* (Dubois and Prade, 1990; Bazan et al., 2006), itself an active area of research in machine learning. Certainly more research is needed here.

Having said that, there is strong support for connecting fuzzy relational structures to cognitive theories of concept formation, when exploring the question of how experts—and these days, AIs—learn to categorize (value) responses to tasks, given some prototypical exemplars: see for example Bělohlávek and Klir (2011).

The outcomes of educational measurement procedures are ultimately underpinned by value judgements about exactly what to assess and how to assess it. As Wiliam (2017, p. 312) puts it: "whereas those focusing on psychological assessment tend to ask, 'Is this correct?', those designing educational assessment have to ask, 'Is this good?'". So questions about how to use mathematical methods in these contexts, in a way that leverages their power, but is not unduly reductive, will no doubt always be debated. It is hoped this paper makes a helpful contribution to that debate.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

AS: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

AS would like to thank his Oxford DPhil supervisor Associate Professor Joshua McGrane (Graduate School of Education, University of Melbourne) for helpful comments on an early draft and general discussion and constructive criticism of some of the ideas presented here. He would like to thank his Oxford DPhil supervisor Professor Jenni Ingram (Department of Education, University of Oxford) for overall support with this research. He also thanks the editor and reviewers for their very helpful comments on an earlier draft of this paper.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of

their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1399317/full#supplementary-material

References

Andrich, D., and Marais, I. (2019). *A Course in Rasch Measurement Theory*. Singapore: Springer.

Bartholemew, D., Steele, F., Moustaki, I., and Galbraith, J. (2008). *Analysis of Multivariate Social Science Data*. Boca Raton, FL: CRC Press.

Bartl, E., Bělohlávek, R., and Scharaschkin, A. (2018). "Toward factor analysis of educational data," in *Proceedings of the 14th International Conference on Concept Lattices and their Applications*, eds. D. Ignatov, and L. Nourine (Olomouc), 191–206.

Bazan, J., Skowron, A., and Swiniarski, R. (2006). "Rough sets and vague concept approximation: from sample approximation to adaptive learning," in *Transactions on Rough Sets V: Lecture Notes in Computer Science 4100*, eds. J. Peters, and A. Skowron (Berlin: Springer), 39–62.

Bedek, M., and Albert, D. (2015). "Applying formal concept analysis to visualise classroom performance," in *Proceedings of the 11th International Conference on Knowledge Management*, eds. T. Watanabe, and K. Seta (Osaka).

Bělohlávek, R. (2012). Optimal decomposition of matrices with entries from residuated lattices. *J. Logic Comp*. 22, 1405–1425. doi: 10.1093/logcom/exr023

Bělohlávek, R., Dauben, J., and Klir, G. (2017). *Fuzzy Logic and Mathematics: A Historical Perspective*. Oxford: Oxford University Press.

Bělohlávek, R. and Klir, G. (eds.). (2011). *Concepts and Fuzzy Logic*. Cambridge, MA: The MIT Press.

Bradley, T.-D. (2020). *At the Interface of Algebra and Statistics* (PhD thesis). New York, NY: City University of New York.

Bradley, T.-D., Gastaldi, J., and Terilla, J. (2024). The structure of meaning in language: parallel narratives in linear algebra and category theory. *Not. Am. Math. Soc*. 71, 174–185. doi: 10.1090/noti2868

Buntins, M., Buntins, K., and Eggert, F. (2016). Psychological tests from a (fuzzy-)logical point of view. *Qual. Quant*. 50, 2395–2416. doi: 10.1007/s11135-015-0268-z

Carpineto, C., and Romano, G. (2004). *Concept Data Analysis: Theory and Applications*. Chichester: Wiley.

Chandler, D., and Munday, R. (2011). *A Dictionary of Media and Communication*. Oxford: Oxford University Press.

Cho, E. (2023). Interchangeability between factor analysis, logistic irt, and normal ogive irt. *Front. Psychol*. 14:1267219. doi: 10.3389/fpsyg.2023.1267219

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxf. Rev. Educ*. 34, 609–636. doi: 10.1080/03054980801970312

Doignon, J.-P., and Falmagne, J.-C. (1999). *Knowledge Spaces*. Berlin: Springer.

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika* 79, 1–19. doi: 10.1007/s11336-013-9342-4

Dubois, D., and Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. *Int. J. Gen. Syst*. 17, 191–209. doi: 10.1080/03081079008935107

Ganter, B., Bedek, M., Heller, J., and Suck, R. (2017). "An invitation to knowledge space theory," in *Formal Concept Analysis: 14th International Conference, ICFCA 2017* (Rennes: Springer), 3–19.

Ganter, B., and Glodeanu, C. (2014). "Factors and skills," in *Formal Concept Analysis: 12th International Conference, ICFCA 2014* (Cluj-Napoca: Springer), 173–187.

Ganter, B., and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer.

Goertz, G. (2006). *Social Science Concepts*. Princeton, NJ: Princeton University Press.

Goguen, J. (1969). The logic of inexact concepts. *Synthese* 19, 325–373. doi: 10.1007/BF00485654

Guttman, L. (1944). A basis for scaling qualitative data. *Am. Sociol. Rev*. 9, 139–150. doi: 10.2307/2086306

Guttman, L. (1971). Measurement as structural theory. *Psychometrika* 36, 329–347. doi: 10.1007/BF02291362

Hajek, P. (1998). *Metamathematics of Fuzzy Logic*. Dordrecht: Kluwer.

Heene, M. (2013). Additive conjoint measurement and the resistance towards falsifiability in psychology. *Front. Psychol*. 4:246. doi: 10.3389/fpsyg.2013.00246

Heilmann, C. (2015). A new interpretation of the representational theory of measurement. *Philos. Sci*. 82, 787–797. doi: 10.1086/683280

Hirth, J., and Hanika, T. (2022). Formal conceptual views in neural networks. *arXiv [Preprint]*. doi: 10.48550/arXiv.2209.13517

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. Berichteüber die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. *Mathematisch-Physische Klasse* 53, 1–46.

Holmes, S., Black, B., and Morin, C. (2017). *Marking Reliability Studies 2017: Rank Ordering Versus Marking: Which Is More Reliable*? Coventry, UK: Technical Report, Ofqual.

Huang, B., Li, J., Li, Q., Zhou, Y., and Chen, H. (2023). *Competence-Based Knowledge Space Theory From the Perspective of Formal Concept Analysis*. Available at: https://ssrn.com/abstract=4620449 (accessed August 13, 2024).

Kane, M. (2008). The benefits and limits of formality. *Measur. Interdisciplin. Res. Perspect*. 6, 101–108. doi: 10.1080/15366360802035562

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *J. Appl. Meas*. 2, 389–423.

Karabatsos, G. (2018). On Bayesian testing of additive conjoint measurement axioms using synthetic likelihood. *Psychometrika* 83, 321–332. doi: 10.1007/s11336-017-9581-x

Kline, P. (2000). *A Psychometrics Primer*. London: Free Association Press.

Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement. Volume I: Additive and Polynomial Representations*. New York, NY: Academc Press.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance: plausible conjoint systems. *Br. J. Math. Stat. Psychol*. 64, 478–497. doi: 10.1348/2044-8317.002004

Leighton, J., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

Li, J., Gui, L., Zhou, Y., West, D., Aloisi, C., and He, Y. (2023). Distilling ChatGPT for explainable automated student answer assessment. *arXiv* [preprrint]. doi: 10.18653/v1/2023.findings-emnlp.399

Lord, F., and Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Luce, R., and Narens, L. (1994). "Fifteen problems concerning the representational theory of measurement," in *Patrick Suppes: Scientific Philosopher*, ed. P. Humphries (Dordrecht: Springer), 219–249.

Luce, R., and Tukey, J. (1964). Simultaneous conjoint measurement: a new scale type of fundamental measurement. *J. Math. Psychol*. 1, 1–27. doi: 10.1016/0022-2496(64)90015-X

Marquer, E. (2020). *Latticenn: Deep Learning and Formal Concept Analysis* (Master's thesis). Nancy: Université de Lorraine.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measur. Interdiscip. Res. Perspect*. 15, 51–69. doi: 10.1080/15366367.2017.1348108

McGrane, J., and Maul, A. (2020). The human sciences: models and metrological mythology. *Measurement* 152:107346. doi: 10.1016/j.measurement.2019.107346

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. London: Routledge.

Michell, J. (1999). *Measurement in Psychology*: *Critical History of a Methodological Concept. Ideas in Context* (Cambridge: Cambridge University Press), 53.

Michell, J. (2006). Psychophysics, intensive magnitudes and the psychometricians' fallacy. *Stud. Hist. Philos. Biol. Biomed. Sci*. 17, 414–432. doi: 10.1016/j.shpsc.2006.06.011

Michell, J. (2009). The psychometricians' fallacy: too clever by half. *Br. J. Math. Stat. Psychol.* 62, 41–44. doi: 10.1348/000711007X243582

Michell, J. (2012). The constantly recurring argument: inferring quantity from order. *Theory Psychol.* 22, 255–271. doi: 10.1177/0959354311434656

Michell, J. (2013). Constructs, inferences and mental measurement. *New Ideas Psychol.* 31, 13–21. doi: 10.1016/j.newideapsych.2011.02.004

Michell, J. (2021). Representational measurement theory: is its number up? *Theory Psychol.* 31, 3–23. doi: 10.1177/0959354320930817

Newton, P., Baird, J., Goldstein, H., Patrick, H., and Tymms, P. (eds.). (2007). *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority.

Peres, A. (1995). *Quantum Theory: Concepts and Methods*. Dordrecht: Kluwer.

Pollitt, A., and Ahmed, A. (2008). "Outcome space control and assessment," in *Technical report, Paper for the 9th annual conference of the Association for Educational Assessment–Europe* (Hissar).

Raykov, T., and Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York, NY: Routledge.

Reid, T. (1748 [1849]). "An essay on quantity," in *The Works of Thomas Reid*, ed. W. Hamilton (Maclachlan, Stuart and Co., Edinburgh), 715–719.

Reyment, R., and Jöreskog, K. (1993). *Applied Factor Analysis in the Natural Sciences*. Cambridge: Cambridge University Press.

Scharaschkin, A. (2023). "Measuring educational constructs qualitatively," in *Paper Presented at the Annual Conference of the Association for Educational Assessment Europe* (Malta).

Shye, S. and Elizur, D. (eds.). (1994). *Introduction to Facet Theory*. Thousand Oaks, CA: Sage Publications, Inc.

Stefanutti, L., and de Chiusole, D. (2017). On the assessment of learning in competence based knowledge space theory. *J. Math. Psychol.* 80, 22–32. doi: 10.1016/j.jmp.2017.08.003

Stevens, S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. doi: 10.1126/science.103.2684.677

Sudmann, A., Echterhölter, A., Ramsauer, M., Retkowski, F., Schröter, J., and Waibel, A. (eds.). (2023). *Beyond Quantity: Research with Subsymbolic AI*. Bielefeld: transcript Verlag.

Tal, E. (2020). "Measurement in science," in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta (Stanford, CA: Stanford University).

Tatsuoka, K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. Boca Raton, FL: CRC Press.

Thurstone, L. (1927a). A law of comparative judgement. *Psychol. Rev.* 34, 278–286. doi: 10.1037/h0070288

Thurstone, L. (1927b). The method of paired comparisons for social values. *J. Abnorm. Soc. Psychol.* 21, 384–400. doi: 10.1037/h0065439

Thurstone, L. (1928). Attitudes can be measured. *Am. J. Sociol.* 33, 529–554. doi: 10.1086/214483

Trimmer, J. (1980). The present situation in quantum mechanics: A translation of Schrödinger's "cat paradox" paper. *Proc. Am. Philos. Soc.* 124, 323–338.

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theoret. Philos. Psychol.* 41, 58–84. doi: 10.1037/teo0000176

Uher, J. (2022a). Functions of units, scales and quantitative data: fundamental differences in numerical traceability between sciences. *Qual. Quant.* 56, 2519–2548. doi: 10.1007/s11135-021-01215-6

Uher, J. (2022b). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: a rigorous analysis showing ways to tackle psychology's crises. *Front. Psychol.* 13:1009893. doi: 10.3389/fpsyg.2022.1009893

van der Linden, W., and Hambleton, K. (eds.). (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer.

van Fraassen, B. C. (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

van Rooij, R. (2011). Measurement and interadjective comparison. *J. Semant.* 28, 335–358. doi: 10.1093/jos/ffq018

von Davier, A., Mislevey, R., and Hao, J. (eds.). (2021). *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment*. Cham: Springer.

Weyl, H. (1952). *Space, Time, Matter*. New York, NY: Dover.

Wiliam, D. (2017). Assessment and learning: a long and winding road. *Assess. Educ.* 24, 309–316. doi: 10.1080/0969594X.2017.1338520

Wolff, J. (2020). *The Metaphysics of Quantities*. Oxford: Oxford University Press.

Appendix C: Appendix to published version of Chapter 5 [Paper 2]

Order relations

Let X be a set. A *binary relation* R on X is a collection of ordered pairs of elements of X. We usually write xRy, rather than $(x, y) \in R$, if x is related to y by R. The relation R is said to be:

- *reflexive* if xRx for all $x \in X$;
- *transitive* if for all $x, y, z \in X$, whenever xRy and yRz, then xRz;
- *anti-symmetric* if for all $x, y \in X$, if xRy and yRx, then $x = y$; and
- *strongly connected* if for all $x, y \in X$, either xRy or yRx.

A binary relation that is reflexive, transitive, and anti-symmetric is called a *partial order*. A *partially-ordered set*, or *poset*, is a set with a partial order on it. A partial order that is, in addition, strongly connected is called a *total order*.

An alternative way of defining a partial order is as a directed graph (collection of objects and arrows), in which there is at most one arrow between any two objects. So when a set

of objects X is partially ordered by a relation $\succeq$, we can use '$a \leftarrow b$' as an alternative to '$a \succeq b$', for $a, b \in X$. A total order is the special case in which there is exactly one arrow between any two objects. Hasse diagrams, such as Figure 5.2, depict posets in this way. The Hasse diagram of a total order is a line (a ranking).

Lattices

Let $(L, \succeq)$ be a poset, and let S be a subset of L. An *upper bound* for S is an element $u \in L$ such that $u \succeq s$ for all $s \in S$. u is the *least upper bound* or *supremum* of S, denoted if $y \succeq u$ for all upper bounds y of S. Dually, a *lower bound* for S is a $l \in L$ such that $s \succeq l$ for all $s \in S$, and l is the *greatest lower bound* or *infimum* of S if l for all lower bounds x of S.

A *lattice* is a poset for which every two-element subset x, y has a supremum $x \vee y$ and an infimum $x \wedge y$. A *complete lattice* has a *bottom* (least) element, usually denoted 0 or $\perp$, and a *top* (greatest) element, denoted 1 or $\top$.

Quantitative structure

Let P be a property of interest. For example, P could be the mark-level (e.g. a mark out of 25) given to an English essay by an examiner. P could be the age of a student, the length of a rod, or the colour of an object. In each case, P can assume, for any particular object in its domain, one of a finite or infinite collection of labels, descriptions, or values drawn from a set V.

Suppose there is a total-order relation $\geq$ on the possible values V of P. Suppose, also, that there is a binary operation $+$ on V (i.e. a mapping $+$ that assigns to each pair of

values (u, v), with $u, v \in V$, another value $u + v \in V$) that satisfies the following six conditions. For any $x, y, z \in V$:

- (associativity) $x + (y + z) = (x + y) + z$;
- (commutativity) $x + y = y + x$;
- (monotonicity) $x \geq y$ if and only if $x + z \geq y + z$;
- (solvability) if $x \geq y$ but $x \neq y$, then there exists a $z \in V$ such that $x = y + z$;
- (positivity) $x + y \geq x$;
- (Archimedean condition) there exists a positive integer n such that $nx \geq y$, where the notation 'nx' means $x + x + ... + x$ (n times).

Then P is said to be a *quantity*. A *continuous quantity* Q is a quantity whose possible values form a continuum with no 'gaps'. Formally, this requires two further conditions. Firstly, Q must be *dense*, which means that between any two values or levels of Q, there is another value:

- (denseness) if x and y are values of Q, with $x \geq y$, then there exists a value z of Q such that $x \geq z \geq y$.

Secondly, Q must be *complete*, which means that all sets of values of Q that are bounded above have a least upper bound:

- (completeness) Let X be a set of values of Q. Then there is a value y of Q such that (i) $y \geq x$ for all $x \in X$; and (ii) if z is any other value such that $z \geq x$ for all $x \in X$, then $y \geq z$.

Appendix D: Appendix 1 to Chapter 6 [Paper 3]

Calculating with matrices of truth values

To be used in models of logics, sets of truth-degrees need some (partial or total) *order* structure, that locate truth-degrees between a 'bottom' (usually called `false`, 0, or $\bot$) and a 'top' (`true`, 1, or $\top$). They also need some *compositional* structure to allow them to be logically aggregated or combined across propositions. For instance if proposition P_1 is true to degree t_1, and P_2 is true to degree t_2, then we wish to express the truth degree of their conjunction 'P_1 and P_2' in terms of a composition of the truth degrees t_1 and t_2.

For the applications in this paper, as noted in section 6.2, we only consider degrees of truth that are totally ordered. In general however, following the suggestion of Birkhoff (1948) to develop logics whose truth values may only be partially ordered[1], the mathematical structure that is used to represent degrees of truth is usually taken to be a so-called *complete residuated lattice*. For a formal definition see Höhle (1996), but inuitively this can be thought of as an ordered structure with a bottom (0), a top (1), and operations

[1] Inspired by Keynes (1921), who suggested that modes of probability ('X is more probable than Y') are only partially ordered (as opposed to standard Kolgomorovian probability theory that aims to attach *real-valued* measures to subsets of a universe set).

$\vee$ and $\wedge$ that generalise the ideas of maximum and minimum, respectively; an operation $\otimes$ to combine truth values; and an operation $\rightarrow$ that generalises the idea of implication. An early example of logic using such a structure was developed by Łukasiewicz (1923 [1970]), taking as truth values the real unit interval [0,1] with its usual order relation, and $a \otimes b := \max(0, a + b - 1)$ and $a \rightarrow b := \min(1, 1 - a + b)$.

Within such a structure we have a way of multiplying matrices. Recall that if A and B are matrices over the real numbers, of shapes $m \times k$ and $k \times n$ respectively, their product AB is defined by setting $(AB)_{ij} := \Sigma_{l=1}^{k} A_{il} B_{lj}$. The analogue for the product $A \circ B$, when A and B are matrices over a lattice of truth-degrees, is defined by setting $(A \circ B)_{ij} := \bigvee_{l=1}^{k} A_{il} \otimes B_{lj}$.

A special case is that for matrices of Boolean truth-degrees, where we take 0 to mean false, and 1 to mean true, $(A \circ B)_{ij} = \max_{l=1}^{k}(\min(A_{il}, B_{lj}))$. If we have a linear chain (totally ordered set) of truth-degrees, labelled by numbers drawn from the unit interval [0,1], there are several options available for defining the composition $\otimes$ (see for example Gottwald, 2001). In this paper the so-called Łukasiewicz t-norm is used, which yields $(A \circ B)_{ij} = \max_{l=1}^{k}(\max(0, A_{il} + B_{lj} - 1))$.

Appendix E: Appendix 2 to Chapter 6 [Paper 3]

Latent variable methods and eigenspace decompositions

Linear factor models as matrix factorisations

Latent variable models assume that the observed data arising from an observational or experimental procedure can be explained by a (relatively small) number of *latent factors*. Often the observed data consists of—or is treated as if it consisted of—values of *quantities* (i.e. real numbers, with their associated addition, multiplication, and order properties) for each subject of interest. In educational assessment contexts, the subjects of study could be students, or test items, or both.

In the case of methods such as exploratory factor analysis (EFA), *explaining* the data broadly means approximating it, using a small number of latent variables, in a way that 'accounts for' most of the observed variation. By contrast methods such as confirmatory factor analysis (CFA) *explain* by postulating a model for how the observations arise from, or are caused by, unobserved variables, and examining whether it appears plausible, given the observations.

In practice, the actual models used in both cases can be similar. Namely, the latent factors are assumed to be quantitative properties of the subjects of interest, and each manifest observation for a subject is usually assumed to be (plus or minus some error) a linear combination of the latent factors. The coefficients in the combination are called the weights or *loadings* of each observation on each factor. In the unidimensional case, with just a single latent factor, each subject's loading on the factor is simply the *amount* of the factor that subject *has*.

In educational assessment a common application of such models is to students' scores on tests. Suppose m students sit a test consisting of n items. Let Z be the $m \times n$ matrix whose $(i,j)^{\text{th}}$ entry z_{ij} is student i's standardised score on item j (i.e., if x_{ij} is the score of student i on item j, then $z_{ij} = \frac{x_{ij}-m_j}{s_j}$, where m_j and s_j are the mean and standard deviation, respectively, of the scores for item j). For reasons explained further below, we often work with the transpose (item by student) matrix Z^T, whose $(j,i)^{\text{th}}$ entry z_{ji} is item j's score when administered to student i.

If we now assume there are k ($\leq n$) latent factors that explain the observed outcomes Z^T on the test, then the linear model described above can be written as

$$\begin{aligned} z_{1i} &= l_{11}f_{1i} + l_{12}f_{2i} + \ldots + l_{1k}f_{ki} + \epsilon_{1i} \\ z_{2i} &= l_{21}f_{1i} + l_{22}f_{2i} + \ldots + l_{2k}f_{ki} + \epsilon_{2i} \\ &\vdots \\ z_{ni} &= l_{n1}f_{1i} + l_{n2}f_{2i} + \ldots + l_{nk}f_{ki} + \epsilon_{ni}, \end{aligned} \tag{$\star$}$$

where l_{jr} is the loading of item j on factor r, f_{ri} is the value of factor r for student i, and ϵ_{ji} is the error in approximating student i's observed score on item j by the weighted sum of her scores on the factors.

If we let L be the $n \times k$ (item by factor) matrix of factor loadings, with $(i,j)^{\text{th}}$ entry l_{ij}, and we let F be the $k \times m$ (factor by student) matrix of factor values, with $(i,j)^{\text{th}}$ entry f_{ij}, then we can write equations $(\star)$ more succinctly as

$$Z^T \approx LF, \qquad (\star\star)$$

assuming the error terms are relatively small. In other words, the process of finding the factor loadings reduces to finding a factorisation of the form $(\star\star)$ for the matrix Z^T.

Now, using a fundamental result from linear algebra, the singular value decomposition (Axler, 2023), the matrix Z^T can be factored as

$$Z^T = VSU^T,$$

where U is an $m \times m$ matrix whose columns are the eigenvectors[1] of ZZ^T, V is an $n \times n$ matrix whose columns are the eigenvectors of Z^TZ, and S is a diagonal matrix whose leading diagonal elements are $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots \geq \sqrt{\lambda_p}$, the square roots of the non-zero eigenvalues of the symmetric matrices ZZ^T and Z^TZ. Here $p = \min\{m,n\}$. Normally we have more students than items, so usually $p = n$.

Eckart and Young (1936), in a paper whose foundational importance for psychometrics is symbolised by its being published in the very first volume of *Psychometrika*, showed that the best rank-k approximation to Z^T, where $k \leq n$, is given by using the eigenvectors with the k largest eigenvalues to create the columns of U and V. Here 'best' means with respect to the so-called Frobenius norm for matrices – a natural extension of the

[1] Recall that a vector $\mathbf{v}$ is an *eigenvector* of a matrix M, with associated *eigenvalue* λ, if $M\mathbf{v} = \lambda\mathbf{v}$.

Euclidean norm to $\mathbb{R}^{m\times n}$. One can think of minimising distances calculated using the Frobenius norm as equivalent to optimisation in a least-squares sense[2].

Let U_k be the $m \times m$ matrix whose first k columns are those k eigenvectors of ZZ^T and whose remaining columns are zeros. Let V_k be the $n \times n$ matrix whose first k columns are those k eigenvectors of Z^TZ, with remaining columns zeros, and let S_k be the $k \times k$ diagonal matrix whose diagonal entries are the square roots of the k corresponding eigenvalues, in descending order of magnitude from the top left. Then Eckhart and Young's theorem states that

$$Z^T \approx V_k S_k U_k^T. \qquad (\dagger)$$

Comparing equations ($\star\star$) and ($\dagger$) shows that we can obtain the factorisation of Z^T that is postulated by the linear factor model with k latent factors by taking $L = V_k S_k$ and $F = U_k^T$. Moreover these choices for L and F give the best approximation to the observed data in a least-squares sense.

The unidimensional case: measurements as locations on eigenlines

In particular, consider the common case in educational assessment in which a single latent factor is assumed, that is, that a student's responses to a test can be explained by her *proficiency* (or *ability*, or *competence*) in the domain of interest. Then we want to estimate a proficiency level for each student, given her observed responses. Here a proficiency level means a numerical score, because we are conceptualising proficiency as a quantity. In this unidimensional case, if we let $\mathbf{v}_i$ be the vector of the n item responses for student i, then applying equation ($\dagger$) yields

[2] Eckart and Young arranged the results of an examination or test as an $n \times m$ item-by-student matrix, rather than (as is usual these days) as an $m \times n$ student-by-item matrix. That is why we focus on Z^T rather than Z in the current discussion.

$$\mathbf{z}_i \approx \sqrt{\lambda_1} u_i \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}. \qquad (\ddagger)$$

Equation (‡) says that each student can be located, as to their amount of proficiency in the domain assessed by the test, at a point along the vector $\mathbf{v} := (v_1, v_2, \ldots, v_n)^T$. Effectively, using the Eckart and Young method to estimate unidimensional proficiency scores amounts to projecting the original n-dimensional outcome data, in which each student's item-response vector is thought of as a point in $\mathbb{R}^n$, onto the one-dimensional subspace of $\mathbb{R}^n$ spanned by $\mathbf{v}$.

Now the vector $\mathbf{v}$ is an eigenvector of the symmetric matrix $Z^T Z$, which is, of course, simply the *covariance matrix* of the test scores, usually denoted by Σ. Thus measuring proficiency by fitting a (linear) latent variable model can be seen as finding points on the eigenline (the unidimensional eigenspace) that corresponds to the largest eigenvalue of Σ.

The eigenspaces of Σ can be thought of as 'ways of demonstrating proficiency', or 'key proficiency states', such that each student's actual proficiency is a mixture (combination) of these key states. The relative sizes of the eigenvalues tell us the relative importance of each eigenspace. When λ_1 is relatively large, by comparison with the other eigenvalues, we can essentially just pick out the first eigenline as a single dimension with respect to which students vary in their proficiency levels.

Item response models and eigenvectors as proficiencies

The discussion above has essentially outlined the logic behind the method of principal component analysis. As Joliffe (2010) notes, principal component analysis and (exploratory or confirmatory) factor analysis are different techniques, because of the way the latter models are usually specified under the assumption that the observations are values of underlying continuous random variables with assumed distribution functions. That allows parameters to be estimated using maximum likelihood estimation. Also CFA assumes a model *a priori*, rather than calculating factors to approximate the data. However in practice, using principal components to estimate factors in an EFA or CFA model often gives similar results to the maximum likelihood estimates.

Item response models used in psychometrics, such as the Rasch (1-parameter) model and the 2-parameter logistic model, can be thought of as linear factor models in which it is assumed that there is a link function (normally the logistic function) between the student's item responses and her factor scores. Item responses in turn are regarded as continuous, with threshold parameters for each item that determine when the 'underlying' response of a student to that item is observed as one of the discrete possible values for the item.

The 2-parameter logistic model is essentially interchangeable with a linear factor analysis model with a single factor (see Cho, 2023) with factors and loadings estimated as described above from the dominant eigenvector and eigenvalue of the covariance matrix. Values of the discrimination and difficulty parameters for each item can be obtained from the factor loadings and threshold parameters for the item. Values of the proficiency parameter for each student can be obtained from their factor scores.

The parameters of the 1-parameter logistic model (the Rasch model) can also be estimated from the eigendecomposition of a matrix of item $\times$ item data, but in this case a

matrix of item *comparisons*, rather than *covariances*. In the dichtomous case, let Z be the item responses (i.e. $Z_{ij} = 1$ if student i answered item j correctly, otherwise $Z_{ij} = 0$). Then $Z^T Z$ is a symmetric matrix whose $(i, j)^{\text{th}}$ entry is the number of students who answered both items i and j correctly. Choppin (1968, 1985) shows that the Rasch item difficulty and person proficiency parameters can be estimated from the dominant eigenvector of $Z^T Z$, i.e. the eigenvector corresponding to the largest eigenvalue. Once again, the question of measuring proficiency reduces to finding points on the eigenline (the unidimensional eigenspace) that corresponds to the largest eigenvalue of a symmetric matrix constructed from the assessment outcomes. Garner and Engelhard (2002) extend this estimation method to the polytomous (partial credit) case.

Summary: comparing quantitative and qualitative representations of the concept of proficiency

When working within an assumption of quantitative structure, the key mathematical construct derived from the observed data that provides the mechanism for generating *measurements* of students' proficiencies is the *eigenspace decomposition*—the eigenvectors and their relative dominance—of the matrix of item covariances or comparisons. Section 6.2 of this paper argues that a natural extension of this underpinning quantitative foundation, to the case where constructs are thought of as composed of attributes whose truth values may be fuzzy, replaces the eigenspace decomposition of $Z^T Z$ with a decomposition of the matrix of truth values M in terms of (fuzzy) formal concepts.

Just as the eigenspaces of $Z^T Z$ can be seen as fundamental *dimensions*, or kinds, of proficiency when it is thought of as quantitative (i.e. an amount, or amounts, of something), the formal concepts of M can be thought of as fundamental *types*, or levels, of proficiency when it is thought of as qualitative (i.e. as a conjunction of propositions about what a learner knows, understands, and can do).

In the unidimensional case, a learner's quantitative proficiency is estimated as a *score* (a position on the dominant eigenline). A learner's qualitative proficiency is estimated as a a *degree of membership* of a *particular type of performance* that is described in terms of construct-relevant attributes.