Routledge
Taylor & Francis Group

Research Papers in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rred20

What is a holistic judgement, anyway?

Tony Leech & Sylvia Vitello

To cite this article: Tony Leech & Sylvia Vitello (19 Oct 2023): What is a holistic judgement, anyway?, Research Papers in Education, DOI: 10.1080/02671522.2023.2269960

To link to this article: https://doi.org/10.1080/02671522.2023.2269960

Published online: 19 Oct 2023.

RESEARCH PAPERS IN EDUCATION
https://doi.org/10.1080/02671522.2023.2269960

Routledge
Taylor & Francis Group

What is a holistic judgement, anyway?

Tony Leech and Sylvia Vitello

Research Division, Assessment Research & Development, Cambridge University Press & Assessment, Cambridge, UK

ABSTRACT

Holistic judgement is an appealing approach for many assessment contexts due to its perceived simplicity and efficiency. However, it has been somewhat under-explored conceptually. Drawing on examples from assessment contexts including vocational performance, comparative judgement and the use of holistic judgement in teacher assessment for high stakes grading, we explicate a three-part definition of what holistic judgement constitutes. Holistic judgements result in singular outputs, derive from a comprehensive consideration of relevant constructs and acknowledge that the elements considered within them interconnect. However, holistic judgements may be made using considerably different processes by different judges without contravening this definition, and the ways in which different elements are weighted may vary. We then explore some factors, specific to assessment contexts, that might make holistic judgements more challenging, including materials being very different from one another, non-uniform candidate performance across materials and the presence of construct-irrelevant material. We set this assessment-specific discussion in the context of literatures on decision-making in psychology, medicine and other contexts. We conclude with some recommendations for when holistic judgement should and should not be used, and how holistic judgements can be made less cognitively challenging through the use of appropriate guidance and feedback to judges.

ARTICLE HISTORY
Received 16 September 2022
Accepted 5 September 2023

KEYWORDS
Judgement; holistic judgement; decision-making; assessment

Introduction

Assessment is an important element of education, as it is used for tracking, accrediting and supporting learning, maintaining educational standards and helping make selection decisions for future study or the workplace. Assessment needs to be reliable and valid in order to be meaningful, and it is therefore crucial to critically appraise methods that are used within it. The idea of 'holistic judgement' has considerable appeal in many different assessment contexts. It can seem simpler, more efficient and/or more valid than other approaches for marking or grading assessments (e.g. criterion-by-criterion, or analytical, marking). This is because holistic judgements require a judge (often an examiner or

CONTACT Tony Leech anthony.leech@cambridge.org Research Division, Assessment Research & Development, Cambridge University Press & Assessment, The Triangle Building, Shaftesbury Road, Cambridge CB2 8EA, UK

a teacher) to focus on the *overall quality* of a piece of work or set of pieces of work (such as a student's essay, an exam script or a portfolio of materials). The idea of a holistic judgement means that judges do not provide separate judgements for different aspects of the work (e.g. assessment criterion 1, assessment criterion 2, assessment criterion 3 and so on).

Holistic judgement has a long history in educational assessment. For example, there are cases of holistic scoring being used for large-scale writing assessment in the 1940s (Haswell and Elliot 2019). The Educational Testing Service in the United States started their development of holistic approaches to scoring within the next decade or so (Haswell and Elliot 2019; Hunter, Jones, and Randhawa 1996). Nowadays, the use of holistic judgement (of various forms) can be seen in a diverse variety of assessment contexts, countries and educational systems. It is used in high-stakes and low-stakes contexts, for different types of assessment (e.g. essays, speaking tests, portfolios, performances) and in different assessment processes (e.g. grading assessment responses, judging students' overall abilities in a subject area, marker moderation, comparability of standards). For a range of different examples see (Chambers 2019; Jönsson, Balan, and Hartell 2021; Rodeiro 2022 and Wilson 2016). There is also considerable work on the role of holistic decision-making in, for instance, hiring for jobs and admissions to universities, especially in the United States (Bastedo et al. 2018; Meijer and Niessen 2022; Yu 2018).

Recent events have drawn our attention again to this topic. In the past few years, as assessment researchers, we have worked on different areas of assessment where the notion of holistic judgement has been central to the process we were researching. Most recently, we worked together to analyse the use of Teacher Assessed Grades (TAGs) in England in 2021 for GCSEs and A levels, the major school level qualifications in England, generally taken at the ages of 16 and 18 (Vitello and Leech 2022). In this context, because normal examinations could not take place due to restrictions on school attendance brought in to combat the COVID-19 pandemic, teachers were asked by official government guidance to make holistic judgements of their students' ability in order to award grades to their students (Ofqual 2021). These judgements had to be evidenced across a variety of pieces of performance evidence which may have included past papers, mock exams, homework or other work, as chosen by the school. There were a number of challenges and inconsistencies in this process, which we explore later in this article in more detail.

This made us reflect on a number of other recent research projects we have worked on across various assessment areas which also invoked the idea of holistic judgement. These included projects on comparative judgement[1] in assessment (Bramley and Vitello 2019; Leech and Chambers 2022) and on understanding vocational competence (Vitello, Greatorex, and Shaw 2021). In these projects, we have seen:

(1) increasing applications of holistic judgement to more assessment contexts, and
(2) increasing numbers of recent research studies giving insights into holistic judgement, but that
(3) holistic judgement approaches in assessment remain rather under-specified.

We think the time is ripe for another proper discussion of some implications of the principles of holistic judgement because of the increasing use of the concept in a variety

of different assessment contexts. This is important because, despite many decades of use and research, there is still debate among assessment researchers and professionals about how valid, reliable and fair holistic judgement is and whether it is strong enough on these measurement characteristics to justify using it for assessment (Bloxham et al. 2016; Hamp-Lyons, 2016a, 2016b; Khabbazbashi and Galaczi 2020; van Daal et al. 2019). The critical question still remains: Is holistic judgement appropriate for educational assessments?

Much of the scrutiny and analysis in assessment contexts has focused on statistical measures of reliability, with comparatively less focus on evidence for validity. In addition, criticisms of holistic approaches to assessment have tended to focus on particular implementations, which means it is not always easy to determine whether the issue is operational or conceptual in nature. For example, many research studies have reported that assessors are inconsistent in how they use criteria when making holistic judgements (Hunter, Jones, and Randhawa 1996; Leech and Chambers 2022; van Daal et al. 2019). This inconsistency could, on the one hand, come from specifics of the operational procedure, such as the type of training used or heterogeneity within the group of assessors; on the other hand, it could come from more general characteristics of holistic judgement, such as the cognitive demands it places on assessors. We argue that giving more attention to the fundamental concept of holistic judgement may help us understand the problems better and to develop solutions that may improve the validity and reliability of holistic judgement when it is implemented within an assessment approach.

In this article, we want to unpack the meaning of holistic judgement and elucidate some of the challenges that the use of holistic judgement presents in different assessment contexts, drawing on recent research and debates in this area. It is the issue of under-specification that has implications, both for reliability (in terms of whether different judges would come to the same outcomes), but more especially for validity, particularly in relation to what can be inferred from assessment outcomes. Under-specified judgement procedures could weaken the argument that assessment results deriving from those judgements can be appropriately interpreted.

We hope to offer clarity around what holistic judgement is and is not, and to encourage readers to critically reflect on this approach. We hope to encourage reflection on the assumptions made about it and how appropriate they are, and the potential impacts of using holistic judgement processes in assessment contexts. Such critical reflection is important for revealing the complexities of holistic judgement, both theoretically and in practice. This reflection should, in turn, help organisations produce clearer guidance and training for judges (assessors, examiners and/or teachers) on using this approach in assessment. We believe this should, ultimately, lead to more consistent, effective and appropriate applications of holistic judgement and its outputs, ensuring holistic judgement methods are fit for purpose.

What does 'holistic' mean, first of all?

It is generally suggested that the term 'holism' was coined in 1926 by Jan Smuts in his book Holism and Evolution (Online Etymology Dictionary 2022; Russell 2016). Drawing on the Greek *holos* (meaning 'whole'), Smuts coined this word to refer to his belief that there is a drive for 'wholeness' in the universe. Current dictionary definitions of holism

give a more generalisable concept, applicable across a diverse range of contexts. For example, Cambridge Dictionaries (2022) defines 'holism' as: 'The belief that each thing is a whole that is more important than the parts that make it up.' This general concept of holism (and the term itself) has been adopted in many domains, including philosophy (Esfeld 1998); medicine and healthcare (Russell 2016); psychology (Nisbett et al. 2001); geography (Archer 1995) and education (Forbes and Martin 2004).

Though we can find variations of meaning of 'holism' both between and within domains (e.g. see Russell 2016, for a discussion in medicine), they share many similarities. Central to many interpretations is the idea that parts of a thing are interconnected in such a way that they cannot be understood without reference to the whole – that it is impossible to entirely separate all the different elements. Holistic judgement, then, is a call to introduce complexity and breadth, not inappropriate simplicity, into our understandings of phenomena.

What does holistic judgement mean in assessment contexts?

Within the assessment literature, there are many definitions of holistic judgement, different labels for these kinds of judgements, and the judgements themselves take different forms. Combined with this, various researchers have noted confusion around which assessments are referred to as holistic (e.g. see Harsch and Martin 2013; Haswell and Elliot 2019), despite multiple attempts to classify assessments along some kind of holistic versus analytic spectrum (Hamp-Lyons 1991; Hunter, Jones, and Randhawa 1996; Weigle 2002).

Nevertheless, there is one aspect that is undeniably common across assessment contexts – holistic judgement focuses the assessor on some notion of an 'overall'. Often, the more an assessment is broken down into more criteria, the less it is viewed as holistic. However, focusing on an 'overall' is not enough to define an approach as holistic, which is where some of the definitional confusion may come from. For example, analytic procedures often also result in an 'overall' assessment score, while during holistic protocols assessors may focus on specific assessment criteria to arrive at their overall judgements (Crisp 2008, 2010; Lumley 2002; Sadler 1989). Therefore, there must be something more than that which defines holistic judgements.

Considering interpretations of 'holistic' within and beyond assessment contexts, we propose there are three central concepts that should define holistic judgement in assessment contexts. We argue that all three of these aspects need to be acknowledged and ideally made explicit (that is, defined or explained to users) when holistic judgement is part of an assessment approach. The three aspects are:

(1) The ultimate output of the holistic judgement is singular in nature.
(2) The process involves the combination of a range of information that comprehensively reflects relevant constructs.
(3) The process involves considering the ways that these different parts interconnect.

Figure 1 presents a visualisation of how these three aspects fit together within the process of holistic judgement. The shapes represent different pieces of interconnected

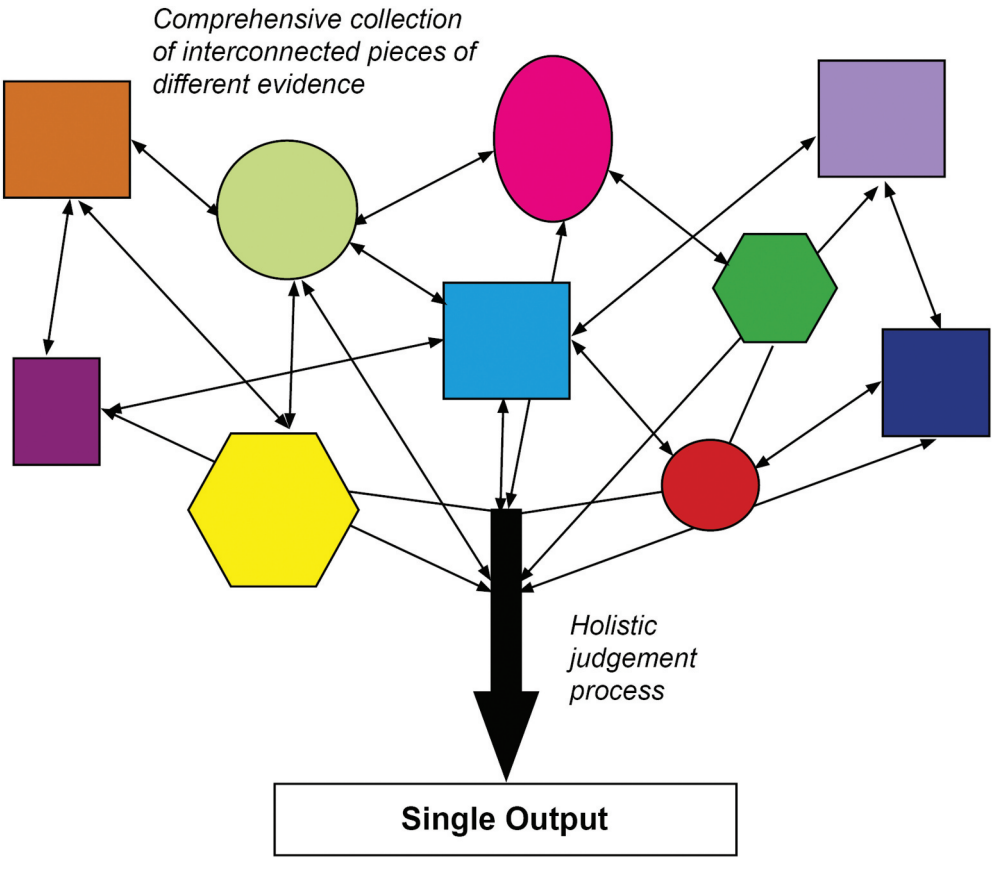

Figure 1. Diagram showing the key elements of holistic judgement under our definition.

information (e.g. assessment evidence, criteria or constructs) that would need to be considered together to reach a holistic judgement. We explain the process in more detail below.

1. The ultimate output of the holistic judgement is *singular* in nature.

While many individual elements are considered during the process of making a holistic judgement (e.g. assessment criteria, features of the students' response), the conclusion of this process is a single output. For example, this could be a grade, a comparative judgement (that one artefact such as an exam script or essay is better than another), a prediction, a selection decision ('yes, you can come to our university') or the like.

This does not necessarily mean that the holistic output has to be *expressed* in a unitary form. It is possible to conceive of holistic judgements expressed as, say, sentences of feedback to a candidate – in other words, a *narrative output*, as opposed to a *discrete unit output* such as a grade. What would make this narrative feedback holistic, in our view, is if it compresses the elements that went into the judgement in a single, unified way. In other words, feedback derived from a mathematics assessment that said something to the test-taker along the lines of 'you are strong in algebra, but in geometry you are less impressive, and you made several errors in

calculus' would be analytical, rather than holistic, as there is no attempt to unify these separate elements into one outcome (say, a judgement of 'mathematics ability' overall). Discrete unit outputs have the advantage that they can be expressed or understood quantitatively and are simple to understand. Narrative outputs, on the other hand, are more nuanced and comprehensive, and can be useful for formative purposes such as giving feedback. Both, however, can be holistic.

The singular nature of holistic assessment results has been the object of some criticisms of holistic approaches to assessment. For example, Hamp-Lyons (1995) has called into question how meaningful and useful an overall result for a candidate's response is, especially for candidates that have uneven profiles of strengths and weakness. One way of overcoming this limitation is to break down very broad holistic judgements into a series of more focused holistic judgements so that they can give information about different features of candidates' responses (e.g. multiple trait assessment developed by Hamp-Lyons 1991). Yet, this type of solution still operationalises holistic judgements as discrete outputs. We argue that, when we view holistic judgement as a broader theoretical concept, one that also encompasses narrative outputs, as long as they are singular, then we increase its potential value and impact as an approach to assessment. For example, in comparative judgement protocols holistic judgements have been combined with narrative feedback to support peer assessment (Potter et al. 2017) and Sadler (2009) argues strongly that this approach can be a 'vehicle for complex learning' (p. 1).

2. The process involves a combination of a range of information that *comprehensively reflects relevant constructs.*

It is difficult to conceive of a judgement being considered truly holistic if it is based only on a very small amount of data, as this would be unlikely to reflect enough relevant variables. For example, a judgement of someone's overall ability in the English language based solely on their ability to read, without reference to other skills such as speaking, listening and writing, would be unlikely to command confidence as a holistic representation. It is necessary for a comprehensive range of relevant information to be included in the set of things judged.

Crucial, of course, to this idea, is some kind of shared understanding of what the relevant constructs, skills and knowledge in each subject actually are. This is perhaps an inherently contested area – as curriculums are socially created and what people consider to be important things to learn or to be able to do in a particular subject vary greatly (including across individuals, subjects, national contexts and so on). Nonetheless, a general (if implicit) principle of the idea of holistic judgement is that such understandings should, on the whole, be more expansive than limited.

In assessment contexts, the relevant elements might be, for example, different items on an exam paper, skills and knowledge learned within a course of study, or different components of a qualification, but we could also consider this list to include such aspects as the social and environmental influences on a candidate's learning, and many other factors. What factors are included would depend on the purpose of the assessment and the specific construct domains.

We do not mean that comprehensive coverage of relevant constructs will be evident at the level of every single topic or item, as this is clearly implausible. Rather, we mean a comprehensive coverage at the level of assessment criteria or construct domains – such as, in the English example above, reading, writing, speaking and listening, for instance.

The notion that holistic judgements take into account a wide range of information is one reason why it has been embraced as an approach in some assessment contexts such as university admissions decision-making (Bastedo et al. 2018) and teacher-based grading of students' abilities in countries like Sweden (Jönsson, Balan, and Hartell 2021).

From an assessment validity point of view, it is often implicitly assumed that, when assessors are asked to make holistic judgements, they will base this on comprehensive and relevant criteria. However, empirical studies have questioned this assumption, finding that assessors who are asked to perform holistic judgements may be driven by a restricted set of the assessment criteria (e.g. Huot 1990) or by a wider set of criteria than those agreed by key stakeholders (for example teachers' grades being affected by non-achievement factors such as students' effort; see Brookhart et al. 2016, for a review).

3. The process involves considering the ways that these different parts *interconnect*.

Finally, a crucial element in holistic judgement, which distinguishes it from analytical forms of judgement, is the notion that it is impossible to wholly distinguish different elements from one another. Instead, there is a sense of a deep and inherent interconnection between the elements. In essay-based assessments within an A level history course (OCR 2021), for example, candidates have to 'demonstrate, organise and communicate knowledge and understanding' in order to 'analyse and evaluate key features' (p.109).[2] In other words, an assessment of the candidate's organisation of the material in their essay cannot easily be distinguished from assessment of their knowledge or their ability to look at and determine the provenance of sources. All the factors are deeply interconnected. Similarly, the way in which competence in a domain is seen to encompass the integration of contextually-appropriate knowledge, skills and psychosocial factors highlights the interconnection of different parts in terms of constituting a competent vocational performance (Vitello, Greatorex, and Shaw 2021).

Here we part company slightly from the approach to holistic judgement taken by the likes of Meehl (1954) and more recently Meijer and Niessen (2022). In this tradition, there is a sense that judgements can be divided into sub-processes and then these can be unified into an overall judgement by mechanical or algorithmic processes. While there are some situations in which this may be possible, our point here is to highlight that this is not always suitable. This is because a proper judgement needs to consider the ways in which features of these sub-criteria interconnect, such that entirely disaggregating them loses something meaningful. In addition, the interactions between these different assessment elements are likely to be dynamic. They are likely to emerge and evolve during the judgemental process as the assessor interprets and reflects on the candidates' work, in much the same way as reading and comprehension is an active process of constructing meaning (McNamara and Magliano 2009). There is, then, value in a truly holistic judgement, if appropriately guided, as we shall go on to discuss further. This tracks the insight of (Eraut 2004, p.262) that judgement is 'schema-driven rather than algorithmic' (p.7) in professional decision-making practice contexts.

How are holistic judgements made in practice?

The nature of *how* the judgemental process is undertaken is not something that by itself defines whether a judgement is holistic or not. As defined by the three aspects described

in the previous section, holistic judgement can be undertaken in a variety of ways. Indeed, the specific processes undertaken by judges may be quite dissimilar, without this necessarily causing variation in outputs – for example, judges may utilise different pieces of evidence at different times, do different tasks (such as judging and justifying judgements) in different orders and at different times in the process, and weigh up evidence differently.

How, therefore, would we expect holistic judgements to be made, given the definition of holistic judgement we have just elaborated? Would this be considered an easy task? Here we give some examples of assumptions held about the process of making holistic judgements within the educational assessment field. Then we focus on three different assessment contexts to discuss what research shows about how holistic judgements are carried out in practice.

Assumptions about how holistic judgements are made in practice

In certain assessment contexts, such as comparative judgement and extended writing, it is common to find the assumption (sometimes with an explicit instruction to judges) that holistic judgements are impressionistic, quick or simple to do (Bramley 2007; Charney 1984; Hamp-Lyons and Davies 2008; Michieka 2010; Pollitt 2012). Some (e.g. Christodoulou 2022) have referred to this, especially in the context of comparative judgement in which it is seen as a major strength of the process, as deriving judgements from judges' 'gut instinct.' This view of holistic judgement bears considerable similarities to how expert intuitive decision-making is considered in the wider psychology literature on human judgement (Erik and Pratt 2007; Eva 2018).

In other assessments where holistic judgements are made, a more systematic process may be expected. That is, criteria are expected to be explicitly focused on and weighed up in some way by the assessor, to reach the single output. This is more likely to be assumed to be the case for contexts where the holistic judgement involves pulling together discrete sets of independent pieces of evidence of performance.

At first sight, this distinction between an intuitive style and a more systematic style of making holistic judgements may seem to resonate with Sadler (1989)'s distinction between 'analytic' and 'configurational' approaches to complex judgements. However, it must be noted that Sadler (1989) specifically described 'analytic' judgements as involving a formula such that 'the global judgment is made by breaking down the multicriterion judgment using separate criteria and then following explicit rules' (p.132). We argue that this kind of analytical approach is not inherently aligned with our core concept of holistic judgement, especially the aspect of criteria interconnectedness. Using formulae may inappropriately constrain the assessment criteria and their interdependencies, which is something that Sadler also cautions against (Sadler 1989, 2009).

For the same reason, we draw a distinction between our description of systematic holistic judgements and the notion of 'mechanical' judgements, as the latter have also specifically been defined as 'approaches that involve applying an algorithm or formula to each applicant's scores' (Kuncel et al. 2013, 1060). The intention of using systematic, rather than those descriptions of analytical or mechanical, approaches to holistic

judgement is different. Judges are still expected to focus explicitly on criteria and different elements of the assessment evidence, but there is no algorithm for reaching their final decision.

Examples of how holistic judgements are made in practice

We now discuss how holistic judgements are actually made in practice. We present examples from three assessment contexts we highlighted earlier: vocational performance, comparative judgement, and the 2021 Teacher Assessed Grades process in England. Although there are various other assessment contexts that use holistic approaches, some of which have been researched much more extensively (e.g. the case of writing assessments), we focus on areas that we are most familiar with and which we have recently researched. These are the contexts that have stimulated our reflections on what holistic judgement means as an assessment approach. In addition, these three contexts are rarely (if ever) discussed together. We find that comparing them alongside each other helps us to appreciate similarities and differences between ways holistic judgements work in practice, which has implications for educational assessment more generally.

Vocational performance

Holistic judgement is an important concept in vocational education and training because of the multifaceted nature of workplace performance (Greatorex, Johnson, and Coleman 2017) and how complex the construct of workplace competence is (Vitello, Greatorex, and Shaw 2021). In some vocational domains, a growing case is being made to increase the status of holistic judgement approaches, with various researchers arguing that they are still undervalued as an assessment tool relative to psychometric methods (see Hodges 2013, for a discussion in relation to medical education).

Vocational assessments are particularly interesting when thinking about holistic judgement not only because of the complexity of the constructs being assessed but also because of other characteristics such as the timeframe over which the performance evidence is gathered or judged and the dynamic nature of the performance evidence. Workplace assessments often involve the gathering and integrating different pieces of information about performance, sometimes across several weeks, months or years. This relates closely to challenges associated with the comprehensiveness and interconnectedness criteria in our definition of holistic judgement. Vocational research has given us important insights into how this information integration may work in practice.

Medical education studies, for example, have shown that assessors' judgements of performance evolve over the course of an assessment: they make interim judgements during the process and use later information to confirm earlier judgements (Chahine, Holmes, and Kowalewski 2016; Yeates et al. 2013). This body of work has also shown that assessors may focus on different task constructs at different stages of an assessment, and weight the constructs differently when arriving at an overall judgement, both of which may vary depending on the students' current level of performance (Chahine, Holmes, and Kowalewski 2016; Ginsburg et al. 2010; Yeates et al. 2013). Growing evidence supports the notion that the process of making such judgements is influenced not only

by cognitive processes, biases and limitations, but also affected by social, cultural and emotional factors (Govaerts et al. 2013; Kogan et al. 2011; Yeates et al. 2013).

Similarly, more recently, 2022, writing in the context of Dutch vocational education, argue that assessment has to be considered as an 'inherently social and relational judgment-based practice' (p.17). They see workplace assessment as a process of making judgements according to multiple overlapping frames of reference (specifically, the vocational, comparative and educational) in which assessors derive what they call an 'overarching image' (in other words, a holistic judgement) of students' abilities, based on mini-judgements of a number of 'different aspects such as technical skills, social interaction, learning curve and personality' (p.12).

Medical education literature has also highlighted challenges in translating judgements into a singular, discrete output. Yeates et al. (2013) observed that 'despite the presence of the rating sheets, these judgements were overwhelmingly phrased in narrative descriptive language rather than either numerical terms, or in terms of the scale descriptors – meets expectations, above, below or borderline for expectations' (p.334). Similarly, Kogan et al. (2011) found medical assessors had different approaches to deciding on numerical ratings from their judgements; some assessors averaged ratings of sub-rating scales, others used non-compensatory approaches and many expressed uncertainty or a lack of understanding of numerical scales. This suggests that the form of the output of the holistic judgement may distort the meaning of the judgement or limit its utility. Narrative judgements may more closely reflect the judgemental process and, thus, its meaning.

Comparative judgement

In a similar way, recent work at Cambridge University Press & Assessment has deepened our understanding of the judgemental processes at work in Comparative Judgement (CJ) processes in assessment contexts (e.g. Chambers and Cunningham 2022; Leech and Chambers 2022; Leech et al. 2022). As an assessment process, Comparative Judgement involves artefacts such as exam scripts being compared directly to one another, rather than individually to a mark scheme. Judges are simply presented with pairs of these and asked to decide which one in each pair is better, creating a single output indication of superiority. Many judgements are made of many artefacts. The resulting decisions as to which artefacts in each pair are better or worse are processed statistically in order to place all the artefacts on a single measurement scale of quality, or to equate two scales of quality to one another.

Comparative judgement has been used in various assessment contexts, including for maintaining exam standards (Bramley 2007), assessing writing in primary and secondary education (Wheadon et al. 2020), and judging digital portfolios of performance (Kimbell 2012). Intrinsic to this approach's logic is the idea that what is being judged is a *holistic* comparison of the overall quality of the artefacts – that is, directly, 'which is better?' More recent work, however, has focused greater attention on the nature of what this holistic judgement involves in terms of procedures.

As Chambers and Cunningham (2022) suggest, 'when making holistic decisions, judges can decide what constitutes good quality; in practice, this conceptualisation can vary across judges' (p.2). The validity of the CJ process is seen to reside in the fact that the individual judges make holistic judgements and that final outcomes are derived from the collective expertise of multiple judges (van Daal et al. 2019). Interestingly, then, judges

may be doing very different things in making their holistic judgements – but this does not seem to have (at least hitherto) caused disquiet so long as overall outcomes are similar enough.

2022 discuss the many different processes undertaken by different judges in a CJ study on a Physical Education (PE) exam. Using a think-aloud protocol, it was found that some judges made direct comparative reference of one exam script to another, and this included systematic question-by-question comparison. Others indicated instead that they were re-marking items in the scripts and then comparing the overall scores. From this and other evidence, 2022 developed a four-dimension model of the factors affecting the outcome of a holistic CJ judgement, in which judges' particular preferences or styles of judgement, such as direct comparative reference and/or re-marking, are one factor. Note that in CJ procedures it is generally not required of judges that they make their judgements in a particular way or that they need to be trained in a specific approach.

2022, as well as Leech et al. (2022), highlight in the CJ studies they discuss that many judges were at least partially *re-marking* the scripts they were judging, and making their judgements of superiority on the basis of whichever script earned a higher mark. In the context of our article, re-marking highlights an interesting tension in the idea of holistic judgement. We would argue that re-marking individual items on paper and then merely summing the scores, as was being undertaken by some judges in these CJ exercises, should not be considered holistic judgement. This is because it fails the third criterion, that of *interconnection*, as all the items are in this context being dealt with independently.

Nonetheless, it is possible for holistic judgements to be the outcome of re-marking approach processes as well. As long as the interrelation of these elements remains a key feature, this would be a holistic judgement. This could occur, for example, as follows:

- the artefact(s) to be judged is/are first divided into separate smaller items/sections/areas of interest, then
- these are individually judged, and then
- a final outcome is arrived at by an (often implicit) process of prioritisation or decision as to which of these is more significant.

These CJ findings raise an important caution with regard to implementing holistic judgement in assessment: despite best intentions, judges may not be able to execute the judgemental process as intended, which may limit what we can infer from their judgement. There may be many reasons for deviation from this process, which we discuss later. There is a similarity between this and the way teachers drew together mini-judgements into a final outcome in the 2021 process for determining Teacher Assessed Grades for high stakes qualifications in England, which we now discuss further.

Teacher assessed grades for GCSEs and A levels in 2021

In England, due to the fact that examinations could not go ahead in 2021 for the second year due to the COVID-19 pandemic, it was decided that teachers would award GCSE and A level grades to their students on the basis of judgements of evidence that students had produced during the programme of study. A crucial element of this process was that the judgements should be holistic. For instance, the exams and qualifications regulator stated that 'each teacher assessed grade or outcome should be a holistic professional

judgement, balancing different sources of evidence' (Ofqual 2021). Teachers were permitted to base their grades on a wide variety of possible sources of performance evidence. Examples of how to reach a holistic decision for different combinations of evidence and circumstances around the evidence and students were provided in official guidance (JCQ 2021). These highlighted how centres could deal with specific situations such as accounting for contextual factors in evidence or replacing evidence due to exceptional circumstances. This was because such exceptional circumstances would reasonably be expected to affect the quality of candidates' performances, and as such factors defining the material to be considered were closely interconnected. Thus, taking this all together, the range of information that teachers were asked to consider in making their grade judgements was greater than the amount of information on which grades are based in normal exam sessions.

In recent work (Vitello and Leech 2022), we investigated some of the specific processes by which holistic judgements were made by these teachers. This study examined documentation submitted by teachers about their grading decisions, which had to include the rationale for their holistic judgement. It was not possible to determine from this evidence how intuitive processes contributed to teachers' decisions. However, the evidence suggested that teachers interpreted 'holistic' in different ways, and some judgements (or at least some teachers' explanations of their judgements) may not be considered 'holistic' by our definition, especially those where teachers did not necessarily integrate all the different pieces of assessment evidence when making judgements.

The findings showed that a variety of different methods were used to prioritise and weight different pieces of evidence against one another (though all were permitted given the flexibility of the guidance). Often the processes used, where rendered explicit (which was somewhat rare), were complicated. For instance, some teachers chose to focus more on results on full exam papers (as opposed to smaller pieces of performance evidence such as homework pieces), while others did not. Some focused on evidence collected more towards the end of the course of study. Some prioritised evidence produced under more formal exam conditions (that is, timed assessments, sat in silence, with questions that had not previously been seen by candidates), as opposed to material produced in other (perhaps more formative and/or classroom-based) contexts. Finally, some reduced the emphasis they placed on evidence produced when students had circumstances that might have impaired their performance (such as recent illness with Covid-19). There was considerable variation in methods used for combining performance information from particular assessments too, including 'best fit' approaches, a focus on consistency, a decision to simply use the highest scores and discard others, and more complex, multifaceted and heterogeneous weightings also being applied. A singular output, that is, a grade, was ultimately produced – but different teachers used very different methods.

Summary across contexts

When considering those three contexts, two key features stand out sharply – the complexity of the process for judges and the level of variation between judges. Consequently, we posit that the idea of a 'judge making a holistic judgement' in one go can be seen as, at best, a convenient simplification, and at worst, a fiction, that obscures considerably more complex, largely implicit, and, crucially, *varied* processes. Similar views have been

expressed by other researchers several decades ago (e.g. Sadler 1989). That is to say, a holistic judgement cannot really be a single, one-step, intuitive process based on gut reaction. Instead, it is more useful to view a holistic judgement outcome as the consequence of the aggregation of a series of micro-judgements, each of which might be quite different for each judge making them (and, indeed, perhaps made using different processes). This is, in other words, a more systematic view of holistic judgement, which, importantly, does not preclude involvement of more intuitive processes. This aligns with research in other assessment contexts which suggest that more complex models of assessor judgement that incorporate different types of judgemental processes are needed to explain empirical evidence of assessors' thought processes and behaviours (e.g. see Crisp 2010, for a model of how assessors mark GCSE and A level exam questions).The three-part definition we explored earlier provides a strong basis for considering in more detail how holistic judgements are made in assessment contexts.

When is holistic judgement difficult?

The question, then, as we have already hinted at, is why do human judges have to make such aggregated micro-judgements in these assessment contexts anyway? What prevents them from performing the idealised 'true' holistic judgement (on the basis of 'gut instinct') that is frequently presupposed? The work (Chambers and Cunningham 2022; Leech and Chambers 2022; Leech et al. 2022; Rodeiro, Lucia, and Chambers 2022; Vitello and Leech 2022) already mentioned offers some suggestions as to why this idealised representation of a judgemental mechanism cannot always be realised. We highlight in Table 1 at least five potential problems derived from these pieces of work that may face those trying to make holistic judgements in assessment contexts and the potential risks to judgement validity.

These risks can be seen as playing the role of 'rebuttal data' within a Toulmin-style validity argument (Kane 2013; Newton 2017; Toulmin 1958). Rebuttal data is any evidence that calls into question a claim that is made from assessment results. The extent that we could make appropriate inferences from assessment results would be limited if the problems mentioned in Table 1 are significant in a particular context. In these cases, we could

Table 1. Potential problems which may render holistic judgement more challenging.

Problem statement	Potential risks to judgement validity
1. Assessment evidence being judged is too large for judges to keep a view of its overall quality in their heads as they assess it	Judges focus (or place more weight) on a small portion of the evidence (e.g. the first paragraph)
2. Assessment evidence tests very different skills, each judged rather independently	Judges judge each skill separately so place different weight on different skills
3. Assessment evidence comprises many different items/sections, each an independent assessment of a particular technique or area of content	Judges judge each item/section separately so place different weight on different items/sections. There is less sense of the interconnectedness of items/sections
4. Candidate performances are non-uniform within assessment evidence (even that which tests the same skills or knowledge)	Judges balance differences in performance differently from one another. Cognitive biases (halo effects, confirmation) create illusory connections between information
5. Construct-irrelevant aspects of the assessment evidence makes it difficult for judges to focus only on relevant evidence	Construct-irrelevant aspects influence judges' judgements of the material

see that holistic judgements may weaken the validity claims made for the assessments in question. Issues with the assessor judgement stage of an assessment protocol have consequences for multiple types of validity inferences (e.g. scoring, generalisation, extrapolation and decision-making in the framework by 2015. However, it has its most direct effects on scoring inferences, which concerns what the assessment judgement (e.g. the student's result) tells us about the quality of the performance that was judged.

Each problem statement focuses onto a different characteristic of the assessment evidence that may be judged during holistic judgement. These characteristics all relate to at least one of the three elements of the definition of holistic judgement that we presented earlier. The first problem concerns the amount of assessment evidence that needs to be processed during a holistic judgement. Recent pieces of research on comparative judgement (e.g. Leech et al. 2022; Rodeiro, Lucia, and Chambers 2022) have highlighted that judges find it difficult to retain a sense of overall quality when judging a large exam script or set of artefacts. This is related closely to the *comprehensive consideration of relevant constructs* criterion we described earlier. The risk this causes for construct representation is clear: if judges cannot consider the whole of a performance, then they are likely to place more weight on only a small portion of it. This means that the judgement is consequently less truly comprehensive, and therefore may fail to be properly holistic. From a validity perspective, this places doubt on what we can infer from these judgement results about the candidates' performance in the assessment and their level of knowledge and skills being assessed.

The second and third problems concern the content of the assessment evidence. Where assessments test very different skills (such as reading and writing), in essentially independent tasks, it can be difficult for judges to know how to weigh up performances on different skills consistently with each other without having an explicit framework for how to do so (e.g. being told that two skills are equal in value). For GCSE English Language papers in England, in which exactly half of each assessment tests reading and writing, this has been highlighted as an issue (Vitello and Leech 2022).

The third problem can be seen as a combination of the first and second problem. It is exemplified by cases like GCSE science and mathematics exam papers, where these papers are constructed as a set of a significant number of small-sized, objectively marked items (by contrast to papers in humanities subjects, where it is more typical for a small number of high tariff, subjectively marked long-answer questions to be used). Because of this structure it is more difficult for judges to draw a sense of the *interconnectedness* of the performances on each (see, for instance, Leech et al. 2022). In addition to this interconnectedness issue, both these second and third problems concern our comprehensiveness criterion primarily, but also relate to the challenge of trying to summarise complex material into a singular output.

The amount of diversity in the assessment evidence may not only arise from the constructs being assessed but by candidates' performance on these constructs, as candidates can perform differently on different content areas. This presents the fourth problem. The more unbalanced a candidate's performance is between content areas, then the more difficult the task of integrating these different pieces of information to arrive at a holistic judgement is likely to be. An unbalanced profile can also interact with certain cognitive biases that influence the way different pieces of information are interpreted and integrated. For example,

certain cognitive biases may create illusory interconnections between pieces of information. The 'halo effect' is one bias whereby experiencing positive impressions early on in evaluations can lead to more positive views of later work from the same student (Bellé, Cantarelli, and Belardinelli 2017) while 'confirmation bias' can mean that initial evaluations encourage us to view later information within that lens and to seek out information that confirms our evaluation (Oswald and Grosjean 2004). The concern is again about how to come to a singular output from this material in a valid way.

Finally the fifth problem we highlight concerns construct irrelevance and the impact this has on the holistic judgement. Assessment evidence that contains features that are not directly related to the construct(s) of significance may make it difficult for judges to focus on the construct-relevant features. As discussed by Chambers and Cunningham (2022) in the context of comparative judgement, judges can be affected by features such as handwriting and whether scripts have more missing answers or incorrect ones. A judgement that, through being influenced by construct-irrelevant factors, misses out reference to a relevant factor falls foul of our *comprehensive consideration of relevant constructs* condition, and as such the judgement is less holistic than it is intended (or assumed) to be.

On the other hand, a judgement including as a factor something irrelevant may not be intrinsically non-holistic when the relevant constructs have also been considered. There is no obvious condition of our holistic judgement definition with which this situation conflicts – but of course the judgement may be purely less valid and of a lower quality.

Also crucial to comprehensiveness is its relationship to the *interconnectedness* criterion. That is to say, relevant constructs have interactions not only with other relevant constructs, but also with irrelevant ones, in a potentially dynamic fashion. This has an impact for judges because information is not processed and interpreted by itself, but by reference to factors previously looked at.

Of course, all these potential problems that may be faced during holistic judgement cannot be wholly distinguished from one another. For example, the reason why assessment material may be likely to consist of many different items or sections may precisely be because each one tests a different skill. Similarly, the fact that it may be very difficult for judges to keep in their heads a sense of the overall quality of an entire exam paper may be because candidates' performances can be non-uniform through them. Unbalanced performances essentially present judges with pieces of qualitatively different information about the candidate's performance that need to be remembered and integrated.

It is naturally the case that those who argue for replacing holistic or clinical judgement with mechanical aggregations would highlight these issues as major challenges to the validity of holistic judgement. However, this is where our definition of holistic judgement is valuable, especially the importance of interconnectedness within that definition. This is because a different – ecological – validity problem would ensue in these assessment contexts were a pure mechanical aggregation methodology approach to be taken, as this approach would not acknowledge the interconnectedness element. As such, pure mechanical aggregation cannot always be the whole solution. Therefore, we argue in the following section that in such contexts there is indeed a value to maintaining the use

of holistic judgement. However, for the validity rebuttals to be dealt with in these contexts, it is necessary to mitigate the issues discussed above.

What does all this mean for using holistic judgement in assessment?

Throughout this article, our aim has not been to argue that a holistic judgement is an impossibility or an entirely useless concept. Far from it, the notion of holistic judgement can be a valuable way of highlighting that judgements in some contexts cannot be made purely analytical without losing sight of the interconnectedness of the factors which are being judged. Nonetheless, being more explicit about the fact that 'holistic judgement' is a convenient way of speaking about a more complex and often implicit process of the aggregation of micro-judgements allows us to understand better what human beings are actually doing when making such judgements, and consequently to better help them do so.

There are also likely to be benefits in terms of considering the best way to describe, and therefore justify, the process. Helping stakeholders appreciate what is going on in a holistic judgement might encourage greater understanding of the output thereof. This would have important accountability and transparency benefits. This would also support stakeholders in making more appropriate inferences from assessment results that have been produced using holistic judgement procedures, enhancing their validity as an assessment approach (Newton 2017). In this final section, then, we discuss approaches to guiding judges in their holistic judgement procedures, and to specifying what they should focus on.

How might we go about this? We have already discussed examples from the medical context where the merits of more holistic and more analytical ways of drawing together information have been long debated. Summing up this argument, Meehl (1954, cited in Ruscio 2003, 39) argues that 'there is no theoretical or empirical reason to suspect that we can combine information in our heads as effectively as we can by using a simple statistical prediction rule'. Such a rule is equal to, in our contexts, a more explicit schema of prioritisation or decision rule within a holistic judgement process. Meehl was writing in the context of making predictions about medical or psychological diagnoses, and argued for formal, mechanical and/or algorithmic procedures to be used in these contexts as opposed to more subjective or impressionistic (holistic) approaches, as he saw mechanical procedures as likely to lead to more reliable decisions about prognosis and treatment. The debate has become known as the clinical-statistical controversy.

In line with Meehl, Ruscio (2003) questions how well human judgement can really handle a 'plethora of information', noting that a true holistic judgement would require knowledge of far more 'unique configurations of information' than have ever existed and demand feats of information integration that are incompatible with our understanding of cognitive limitations. He argues for approaches that 'resemble an additive sum of main effects', which he sees as more likely to provide the requisite predictive validity for clinical diagnosis. That said, he suggests 'limited forms of holism' are useful, especially in situations 'in which there are a few categorical and causally important variables' for judges to focus on and so there is less cognitive load on judges (pp.38–46).

Following from this, we argue that in assessment contexts, there would be a great benefit to instructing those making these holistic judgements more clearly as to how

exactly to draw together these micro-judgements on individual pieces of information. In other words, clear guidance should be provided as to which specific criteria or bits of evidence (Ruscio's 'categorical and causally important variables') are the most important.

There are examples of increased specificity in assessment contexts as well as in holistic review and personnel selection. The work of 2017 on assessment of vocational competence highlights a general view that, where more holistic or global judgements are necessary (because the use of entirely analytical approaches such as checklists misses some important elements of the interconnectedness of competencies), using specified combinations of evidence can outperform individual, idiosyncratic judgemental approaches. However, the quality of the assessment tool remains the most important factor. In addition, in official guidance on a contingency assessment process for 2022 grading in England (DfE/Ofqual 2021), in which teachers would be asked to collect evidence that could support a Teacher Assessed Grade in the event that examinations could not take place, the Department for Education (DfE) and regulator, Ofqual, provided more specific information for teachers as to what this evidence should be than in the aforementioned 2021 TAGs process (Ofqual 2021). The later guidance explicitly favoured assessment materials similar to exam papers for the relevant course (e.g. past papers) and assessment evidence gathered under exam conditions (DfE/Ofqual 2021).

It is perfectly possible that the teachers determining grades in the 2021 TAGs process may have themselves come up with quite precise decision rules, very similar to those outlined in the later 2022 guidance, in order to structure their holistic judgements. However, these rules were not specified at that time. This means many may not have done this, and as such their processes may have been inconsistent or intangible (that is, teachers may have been unable to exactly explain why they graded as they did, leading to accountability issues). Moreover, the introduction of significant variation into processes highlights a potential risk in terms of comparability of standards. The use of a more specified, common process should provide some level of mitigation for these issues.

A critique of the notion that greater specificity is good for holistic judgement in assessment, though, might run as follows. It could be argued that allowing judges maximal flexibility in terms of what they consider to be high quality (and then perhaps using multiple judgements to establish some level of consistency) means that different valid and legitimate views of the construct can be integrated (Bloxham et al. 2016; Hodges 2013). This approach, however, presents a number of risks. It is difficult to justify, in a high-stakes context where grades have significant progression value, that the same piece of work could *legitimately* be viewed very differently by different markers, as this suggests little agreement on the construct. It is valuable, then, for public accountability and hence for the face validity of an assessment, to be able to clarify more clearly what elements within assessments should be most highly prioritised, and for there, consequently, to be greater consistency between judges on this. In validity-argument language, this would allow assessment designers to offer a rebuttal to a critique of the scoring inference of an assessment item that ran along the lines of 'the same performance(s) are rewarded with the same score by different markers or judges.'

We acknowledge that it will never be possible to pre-determine *all* the factors that might be relevant for a judgement, and, thus, it may not be possible to provide an explicit rationale for how judges should formally organise each into

a prioritisation tree for coming to their holistic judgement. Consequently, the above critique can never be entirely rebutted. However, going some way towards that by producing guidance that aligns closely with the constructs intended to be assessed should allow for the features viewed as most significant to be identified and prioritised. It is possible for guidance to be more explicit about what elements judges should and should not focus on, and to what degree. Inasmuch as some of the criteria may be latent or implicit, and therefore not actually possible to predetermine before the judging begins, an iterative feedback process whereby guidance is developed and re-evaluated by judges themselves, taking into account what they are experiencing throughout the judging, may be valuable (Sadler 1989). This may have accountability and consistency benefits, as well as reflecting judges' expertise, and hopefully would maintain a sufficient level of validity for public acceptability.

In a similar way, other methods of making the cognitive processes 'more manageable' for judges should be considered – including the avoiding where possible of 'apples and pears' judgements. For example, as stated above, it is known from various assessment contexts (see, for instance, Leech and Chambers 2022) that judgements become very difficult if the artefacts being considered are very different in form to one another (in terms of factors such as the number and size of items, and their relative weighting). This issue is unsurprising; how *should* one judge a more substantial paper on which a candidate has performed relatively poorly against a smaller one (perhaps on a different content area) on which they have performed better, if one is meant to be judging holistically? The same issue is present when comparing different pieces of material produced by the same candidate, if the materials are very different in form to one another. There is no intrinsically *more* valid way to make such a comparison so judges will apply their own informal rubric in which they determine which to prioritise. We contend this informal rubric should be made more formal by the provision of more explicit guidance, and the comparison simplified by, if at all possible, ensuring the similarity of form between different artefacts.

It is also important that guidance about the judgement process fits our understanding of what judges are actually doing when they judge. Consequently, guidance should be informed by both the best cognitive science on how people schematise and interpret information (Eva 2018; Gingerich et al. 2014), and by a deep understanding of how judges in assessment contexts work. Steps should be taken to reduce the cognitive load for judges. One way is to make the different pieces of material that must be combined accessible in consistent formats using the same viewing platforms. For instance, Bastedo et al. (2022), in the US university admissions context, has explored the use of information dashboards for presenting decision makers with relevant contextual data to support holistic admissions decisions.

Finally, however, if holistic judgement is not necessary in a particular assessment, especially because the material being judged is not sufficiently interconnected (as for instance, it tests a series of very different skills essentially independently), then it should not be used – in these contexts it is perfectly reasonable to mechanically aggregate a series of sub-scores (using an analytical approach). The use of holistic judgement should be restricted to cases where it is genuinely necessary for understanding the interconnected nature of competencies or skills in a particular area. The

more that the relevant skills can be specified in guidance and a prioritisation schema or rough weighting indicated, the more consistent, justifiable and valid, holistic judgements are likely to be.

In short, when assessment designers are considering whether it is necessary for holistic judgement to be part of an assessment, the following rough framework should be employed.

(1) Do you need holistic judgement? The first task is to decide whether to use holistic judgement or another approach to assessment. Holistic judgement is particularly useful when there is interconnectedness within the assessment evidence that needs to be integrated. Alternative assessment approaches (e.g. analytical) should be considered if there is little sense that the features of the assessment material or performance being judged are strongly interconnected, or if there is no need for a singular output from the judgement.
(2) How will you address the five problems affecting holistic judgement? After taking the decision to use holistic judgement, it will be important to consider the five problems we discussed earlier in the specific context of the assessment. How, and to what extent, can these be mitigated in the design of the assessment tasks, in terms of the content to be assessed (and that which should not be assessed, as it is construct-irrelevant) and the volume and structure of evidence produced by a candidate?
(3) How are your assessors making their holistic judgements? Finally, it is important to understand (at least to some degree) the judgemental processes used by judges to help provide guidance on how to weight or prioritise particular aspects of the evidence or performance, as this will likely improve the quality of the judgement. This guidance should be based where possible on analysis of actual judgemental processes – what judges are actually doing as they judge – and may well be best produced by the judges themselves.

Notes

1. Comparative judgement is an assessment approach where judges (e.g. examiners or teachers) are presented with pairs of artefacts (such as students' essays) and simply asked to decide which one in each pair is the better one. The comparative judgements then undergo a statistical process that places all those judged artefacts on a single measurement scale (a scale of quality).
2. OCR is an organisation that provides courses of study and assessments in the United Kingdom in a variety of subjects. It is a part of Cambridge University Press & Assessment.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Tony Leech is a Senior Researcher at Cambridge University Press & Assessment (OCR). He has degrees from the University of Cambridge. His recent assessment research has included projects

on the measurement characteristics of Teacher Assessed Grades, universal design for assessment and comparative judgement.

Sylvia Vitello is a Senior Researcher in the Assessment Research and Development division at Cambridge University Press & Assessment. She holds a PhD in Experimental Psychology from University College London. Her current research focuses on assessment design for general and vocational qualifications as well as broader education topics such as competence, equality, diversity and artificial intelligence.

References

Archer, K. 1995. "A Folk Guide to Geography as a Holistic Science." *The Journal of Geography* 94 (3): 404–411.

Bastedo, M. N., D. Bell, J. S. Howell, J. Hsu, M. Hurwitz, G. Perfetto, and M. Welch. 2022. "Admitting Students in Context: Field Experiments on Information Dashboards in College Admissions." *The Journal of Higher Education* 93 (3): 327–374.

Bastedo, M. N., N. A. Bowman, K. M. Glasener, and J. L. Kelly. 2018. "What are We Talking About When We Talk About Holistic Review? Selective College Admissions and Its Effects on Low-SES Students." *The Journal of Higher Education* 89 (5): 782–805.

Bellé, N., P. Cantarelli, and P. Belardinelli. 2017. "Cognitive Biases in Performance Appraisal: Experimental Evidence on Anchoring and Halo Effects with Public Sector Managers and Employees." *Review of Public Personnel Administration* 37 (3): 275–294.

Bloxham, S., B. den-Outer, J. Hudson, and M. Price. 2016. "Let's Stop the Pretence of Consistent Marking: Exploring the Multiple Limitations of Assessment Criteria." *Assessment & Evaluation in Higher Education* 41 (3): 466–481.

Bramley, T. 2007. "Paired Comparison Methods." In *Techniques for Monitoring the Comparability of Examination Standards*, edited by P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, and P. Tymms, 246–300. London: QCA.

Bramley, T., and S. Vitello 2019. "The Effect of Adaptivity on the Reliability Coefficient in Adaptive Comparative Judgement." *Assessment in Education Principles, Policy & Practice* 26 (1): 43–58. doi:10.1080/0969594X.2017.1418734.

Brookhart, S. M., T. R. Guskey, A. J. Bowers, J. H. McMillan, J. K. Smith, L. F. Smith, M. T. Stevens, and M. E. Welsh. 2016. "A Century of Grading Research: Meaning and Value in the Most Common Educational Measure." *Review of Educational Research* 86 (4): 803–848.

Cambridge Dictionaries. 2022. "Meaning of Holism in English." Cambridge University Press. Accessed June 01, 2022. https://dictionary.cambridge.org/dictionary/english/holism.

Chahine, S., B. Holmes, and Z. Kowalewski. 2016. "In the Minds of OSCE Examiners: Uncovering Hidden Assumptions." *Advances in Health Sciences Education* 21 (3): 609–625.

Chambers, L., and E. Cunningham. 2022. "Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features?" *Frontiers in Education* 7:1–14. https://doi.org/10.3389/feduc.2022.802392.

Chambers, L., J. Williamson, and S. Child. 2019. "Moderating Artwork–Investigating Judgements and Cognitive Processes." *Research Matters A Cambridge Assessment Publication* 27:19–25. https://www.cambridgeassessment.org.uk/Images/542748-moderating-artwork-investigating-judgements-and-cognitive-processes-.pdf.

Charney, D. 1984. "The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview." *Research in the Teaching of English* 18(1): 65–81. https://www.jstor.org/stable/40170979.

Christodoulou, D. 2022. "Give Me Your Answer Do: An Interview with Daisy Christodolou." researchEd, Accessed September 7. https://researched.org.uk/2018/09/26/give-me-your-answer-do-an-interview-with-daisy-christodoulou/ .

Crisp, V. 2008. "Exploring the Nature of Examiner Thinking During the Process of Examination Marking." *Cambridge Journal of Education* 38 (2): 247–264.

Crisp, V. 2010. "Towards a Model of the Judgement Processes Involved in Examination Marking." *Oxford Review of Education* 36 (1): 1–21.
De Vos, M. E., L. K. J. Baartman, C. P. M. Van der Vleuten, and E. De Bruijn. 2022. "Unravelling Workplace educators' Judgment Processes When Assessing students' Performance at the Workplace." *Journal of Vocational Education & Training* 1–20. https://doi.org/10.1080/13636820.2022.2042722.
DfE/Ofqual. 2021. *Consultation Outcome. Decisions on Contingency Arrangements 2022: GCSE, AS, a Level, Project and AEA*. England: Department for Education and The Office of Qualifications and Examinations Regulation.
Eraut, M. 2004. "Informal Learning in the Workplace." *Studies in Continuing Education* 26 (2): 247–273. https://doi.org/10.1080/158037042000225245
Erik, D., and M. G. Pratt. 2007. "Exploring Intuition and Its Role in Managerial Decision Making." *Academy of Management Review* 32 (1): 33–54.
Esfeld, M. 1998. "Holism and Analytic Philosophy." *Mind* 107 (426): 365–380.
Eva, K. W. 2018. "Cognitive Influences on Complex Performance Assessment: Lessons from the Interplay Between Medicine and Psychology." *Journal of Applied Research in Memory and Cognition* 7 (2): 177–188.
Forbes, S. H., and R. A. Martin. 2004. "What Holistic Education Claims About Itself: An Analysis of Holistic schools' Literature." In *American Education Research Association Annual Conference*. San Diego, California
Gingerich, A., J. Kogan, P. Yeates, M. Govaerts, and E. Holmboe. 2014. "Seeing the 'Black box'differently: Assessor Cognition from Three Research Perspectives." *Medical Education* 48 (11): 1055–1068.
Ginsburg, S., J. McIlroy, O. Oulanova, K. Eva, and G. Regehr. 2010. "Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency." *Academic Medicine* 85 (5): 780–786.
Govaerts, M. J. B., M. W. J. Van de Wiel, L. W. T. Schuwirth, C. P. M. Van der Vleuten, and A. M. M. Muijtjens. 2013. "Workplace-Based Assessment: Raters' Performance Theories and Constructs." *Advances in Health Sciences Education* 18 (3): 375–396.
Greatorex, J., M. Johnson, and V. Coleman. 2017. "A Review of Instruments for Assessing Complex Vocational Competence." *Research Matters A Cambridge Assessment Publication* 23:35–42. https://www.cambridgeassessment.org.uk/Images/375448-a-review-of-instruments-for-assessing-complex-vocational-competence.pdf
Hamp-Lyons, L. 1991. "Scoring Procedures for ESL Contexts." In *Assessing Second Language Writing in Academic Contexts*, edited by L. Hamp-Lyons, 241–276. Norwood, NJ: Ablex.
Hamp-Lyons, L. 1995. "Rating Nonnative Writing: The Trouble with Holistic Scoring." *TESOL Quarterly* 29 (4): 759–762.
Hamp-Lyons, L. 2016a. "Farewell to Holistic Scoring?" *Assessing Writing* 27:A1–A2. https://doi.org/10.1016/j.asw.2015.12.002.
Hamp-Lyons, L. 2016b. "Farewell to Holistic Scoring. Part Two: Why Build a House with Only One Brick?" *Assessing Writing* 100 (29): A1–A5.
Hamp-Lyons, L., and A. Davies. 2008. "The Englishes of English Tests: Bias Revisited." *World Englishes* 27 (1): 26–39.
Harsch, C., and G. Martin. 2013. "Comparing Holistic and Analytic Scoring Methods: Issues of Validity and Reliability." *Assessment in Education: Principles, Policy & Practice* 20 (3): 281–307.
Haswell, R., and N. Elliot. 2019. *Early Holistic Scoring of Writing: A Theory, a History, a Reflection*. Colorado: University Press of Colorado.
Hodges, B. 2013. "Assessment in the Post-Psychometric Era: Learning to Love the Subjective and Collective." *Medical Teacher* 35 (7): 564–568.
Hunter, D. M., R. M. Jones, and B. S. Randhawa. 1996. "The Use of Holistic versus Analytic Scoring for Large-Scale Assessment of Writing." *Canadian Journal of Program Evaluation* 11 (2): 61.
Huot, B. 1990. "Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know." *College Composition and Communication* 41 (2): 201–213.

JCQ. 2021. *Worked Examples to Assist with Determining Grades: JCQ Supporting Guidance on the Awarding of Grades for A/AS Levels for Summer 2021*. England: Joint Council for Qualifications.
Jönsson, A., A. Balan, and E. Hartell. 2021. "Analytic or Holistic? A Study About How to Increase the Agreement in teachers' Grading." *Assessment in Education: Principles, Policy & Practice* 28 (3): 212–227.
Kane, M. T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50 (1): 1–73.
Khabbazbashi, N., and E. D. Galaczi. 2020. "A Comparison of Holistic, Analytic, and Part Marking Models in Speaking Assessment." *Language Testing* 37 (3): 333–360.
Kimbell, R. 2012. "Evolving Project E-Scape for National Assessment." *International Journal of Technology & Design Education* 22 (2): 135–155.
Kogan, J. R., L. Conforti, E. Bernabeo, W. Iobst, and E. Holmboe. 2011. "Opening the Black Box of Clinical Skills Assessment via Observation: A Conceptual Model." *Medical Education* 45 (10): 1048–1060.
Kuncel, N. R., D. M. Klieger, B. S. Connelly, and D. S. Ones. 2013. "Mechanical versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis." *Journal of Applied Psychology* 98 (6): 1060.
Leech, T., and L. Chambers. 2022. "How Do Judges in Comparative Judgement Exercises Make Their Judgements?" *Research Matters: A Cambridge University Press & Assessment Publication* 33:31–47. https://www.cambridgeassessment.org.uk/Images/research-matters-33-how-do-judges-in-comparative-judgement-exercises-make-their-judgements.pdf.
Leech, T., T. Gill, S. Hughes, and T. Benton. 2022. "The Accuracy and Validity of the Simplified Pairs Method of Comparative Judgement in Highly Structured Papers." *Frontiers in Education* 7:1–18. https://doi.org/10.3389/feduc.2022.803040.
Lumley, T. 2002. "Assessment Criteria in a Large-Scale Writing Test: What Do They Really Mean to the Raters?" *Language Testing* 19 (3): 246–276.
McNamara, D. S., and J. Magliano. 2009. "Toward a Comprehensive Model of Comprehension." In *The Psychology of Learning and Motivation*, edited by B. H. Ross, 297–384. Online: Elsevier Academic Press.
Meehl, P. E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minnesota: University of Minnesota Press.
Meijer, R. R., and A. S. M. Niessen. 2022. "Personnel Selection as Judgment and Decision Science: An Introduction." *International Journal of Selection and Assessment* 30 (2): 193–194. https://doi.org/10.1111/ijsa.12377.
Michieka, M. 2010. "Holistic or Analytic Scoring? Issues in Grading ESL Writing." *TNTESOL Journal* 3:75–83. https://www.tennesseetesol.org/Resources/Documents/Journal%20and%20Newsletter/TNTESOL%20Journal%20v3%202010%20%20.pdf#page=76 .
Newton, P. 2017. *An Approach to Understanding Validation Arguments*. Coventry: The Office of Qualifications and Examinations Regulation.
Nisbett, R. E., K. Peng, I. Choi, and A. Norenzayan. 2001. "Culture and Systems of Thought: Holistic versus Analytic Cognition." *Psychological Review* 108 (2): 291.
OCR. 2021. *A Level Specification History a H505: For First Assessment in 2017*. England: OCR.
Ofqual. "Guidance: Information for Heads of Centre, Heads of Department and Teachers on the Submission of Teacher Assessed Grades: Summer 2021 (HTML)." Ofqual. Accessed March 21, 2022. https://www.gov.uk/government/publications/submission-of-teacher-assessed-grades-summer-2021-info-for-teachers/information-for-heads-of-centre-heads-of-department-and-teachers-on-the-submission-of-teacher-assessed-grades-summer-2021-html .
Online Etymology Dictionary. 2022. "Holism (n.)." Accessed June 1, 2022. https://www.etymonline.com/word/holism.
Oswald, M. E., and S. Grosjean. 2004. "Confirmation Bias." In *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement, and Memory*, edited by R. F. Pohl, 79–96. New York: Psychology Press.
Pollitt, A. 2012. "The Method of Adaptive Comparative Judgement." *Assessment in Education: Principles, Policy & Practice* 19 (3): 281–300.

Potter, T., L. Englund, J. Charbonneau, M. T. MacLean, J. Newell, and I. Roll. 2017. "ComPair: A New Online Tool Using Adaptive Comparative Judgement to Support Learning with Peer Feedback." *Teaching & Learning Inquiry* 5 (2): 89–113.

Rodeiro, V., C. Lucia, and L. Chambers. 2022. "Moderation of Non-Exam Assessments: Is Comparative Judgement a Practical Alternative?" *Research Matters: A Cambridge University Press & Assessment Publication* 33:100–119. https://www.cambridgeassessment.org.uk/Images/research-matters-33-moderation-of-non-exam-assessments-is-comparative-judgement-a-practical-alternative.pdf .

Ruscio, J. 2003. "Holistic Judgment in Clinical Practice." *The Scientific Review of Mental Health Practice* 2 (1): 33–48.

Russell, G. 2016. "Holism and Holistic." *BMJ* 353:i1884. https://doi.org/10.1136/bmj.i1884.

Sadler, D. R. 1989. "Formative Assessment and the Design of Instructional Systems." *Instructional Science* 18 (2): 119–144.

Sadler, D. R. 2009. "Transforming Holistic Assessment and Grading into a Vehicle for Complex Learning." In *Assessment, Learning and Judgement in Higher Education*, edited by G. Joughin, 1–19. New York: Springer.

Shaw, S., and V. Crisp. 2015. "Reflections on a Framework for Validation – Five Years on." *Research Matters A Cambridge Assessment Publication* 19:31–37. https://www.cambridgeassessment.org.uk/Images/465780-reflections-on-a-framework-for-validation-five-years-on.pdf .

Toulmin, S. E. 1958. *The Uses of Argument*. Cambridge: Cambridge University Press.

van Daal, T., M. Lesterhuis, L. Coertjens, V. Donche, and S. De Maeyer 2019. "Validity of Comparative Judgement to Assess Academic Writing: Examining Implications of Its Holistic Character and Building on a Shared Consensus." *Assessment in Education: Principles, Policy & Practice* 26 (1): 59–74.

Vitello, S., J. Greatorex, and S. Shaw. 2021. *What is Competence? A Shared Interpretation of Competence to Support Teaching, Learning and Assessment*. Cambridge: Cambridge University Press & Assessment.

Vitello, S., and T. Leech. 2022. *What Do We Know About the Evidence Sources Teachers Used to Determine 2021 Teacher Assessed Grades?*. Cambridge: Cambridge University Press & Assessment.

Weigle, S. C. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.

Wheadon, C., P. Barmby, D. Christodoulou, and B. Henderson. 2020. "A Comparative Judgement Approach to the Large-Scale Assessment of Primary Writing in England." *Assessment in Education: Principles, Policy & Practice* 27 (1): 46–64.

Wilson, J., N. G. Olinghouse, D. Betsy McCoach, T. Santangelo, and G. N. Andrada. 2016. "Comparing the Accuracy of Different Scoring Methods for Identifying Sixth Graders at Risk of Failing a State Writing Assessment." *Assessing Writing* 27:11–23. https://doi.org/10.1016/j.asw.2015.06.003.

Yeates, P., P. O'Neill, K. Mann, and K. Eva. 2013. "Seeing the Same Thing Differently." *Advances in Health Sciences Education* 18 (3): 325–341.

Yu, M. C. 2018. "Viewing expert judgment in individual assessments through the Lens Model: Testing the limits of expert information processing." PhD, University of Minnesota.