CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Research Matters

Issue 32 / Autumn 2021

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Proud to be part of the University of Cambridge

Cambridge University Press & Assessment unlocks the potential of millions of people worldwide. Our qualifications, assessments, academic publications and original research spread knowledge, spark enquiry and aid understanding.

Citation

Articles in this publication should be cited using the following example for article 1: Coleman, V. (2021). What is (or are) social studies? *Research Matters: A Cambridge University Press & Assessment publication, 32*, 6–21.

Credits

Reviewers: Matthew Carroll, Vicki Crisp, Gill Elliott, Joanne Ireland, Melissa Mouthaan

Editorial and production management: Anouk Peigne

Additional proofreading: Judith Nial and Alison French

Design: Lente Artemieff

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division, researchprogrammes@cambridgeassessment.org.uk

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

All details are correct at the time of publication in October 2021.

Contents

Research Matters / 32

A Cambridge University Press & Assessment publication

Foreword

By necessity, in response to Covid-19, nations around the world have had to undertake one of the biggest natural experiments in education which we ever have seen. And now we have a huge challenge in trying to understand what actually happened, what impact there has been for young people, schools and society, and what continuing consequences will flow through schooling, society and the economy. But the very nature of the pandemic has meant that undertaking research is extremely difficult, not just because of the fast-paced nature of events and the responses but also because of disruption to researchers' own professional and personal lives. However, we really do need to know how different people were affected, how different approaches to remote learning functioned, and how and over what timeframe the inequalities and differences will play out—important not least for determining appropriate support and action as schooling begins to gear up and remote learning starts to reduce in intensity. Understanding what worked and what didn't (both in terms of attainment and equity) is important for focusing ongoing support to individuals, schools, and to regions, as well as for developing curriculum policy on the use of remote learning—and developing the continuing response to a pandemic which very much is still with us. Our work with Cambridge CEM synthesising research on responses to interrupted learning and probing learning during the pandemic is designed to be both penetrating and timely. There is a very real risk that research lags behind the need to act. Such lags carry a strong risk of policy error—something we anticipated in the 2017 Cambridge Approach to Improving Education. But quite rightly, policy makers cannot wait in current times; swift action is needed and the research community has to step up with rapid research synthesis and empirical enquiry. At the same time, we are sceptical of "Year Zero" thinking—Tom Benton's piece reminds us that we should not forget what we were thinking about and probing prior to 2020. Any "new normal" should be constructed on the very best of what we know about assessment, curriculum, and educational improvement.

Tim Oates, **CBE** Group Director, Assessment Research and Development

Research Matters / 32

A Cambridge University Press & Assessment publication

Editorial

Our first article in this issue, by Tori Coleman, looks at what is meant by the term *social studies*. As you might expect, it's not entirely clear-cut, but investigations like this can help us to avoid misunderstandings and confusions in conversations with people from countries that use the term slightly differently.

Our second and third articles come from a collaboration with our researchers in Cambridge CEM (part of Cambridge University Press & Assessment since June 2019). They look at the impact of the drastic changes to school life created by the lockdowns imposed to manage the pandemic. The article by Joanna Williamson et al. looks at the effects on students—including what they most missed during lockdown and what they now feel they need more of. The article by Chris Jellis et al. considers teachers' perceptions of their wellbeing during and after lockdown.

Producing exam papers is a complex process involving many people with different roles. Everyone is committed to the quality of the end product—but very occasionally errors manage to evade all the checks and appear in front of candidates. But what exactly counts as an error? Our fourth article, by Nicky Rushton et al., looks at definitions of error in other industries and relates them to the perceptions and understanding of error among those with different roles in producing exam papers at Cambridge University Press & Assessment.

Our final article, by Tom Benton, shows that it is important to avoid "wishful thinking" when anticipating the benefits to reliability that adaptive testing might bring, in particular if tests made up of the kind of questions currently used in GCSEs and A Levels were to be administered adaptively.

Tom Bramley Director, Research Division

What is (or are) social studies?

Victoria Coleman[1] (Research Division)

Introduction

As we are a global organisation, our customers and stakeholders often have very different educational traditions and systems. This can cause confusion when the same terms have different meanings in different countries and cultures. One area where we have found this to be the case is when talking about social studies. The aim of this research was to increase our understanding of what is meant by this widely, but variably, used term. Social studies is a subject discipline which has provoked significant dispute over a range of areas, including whether it should exist as a subject at all (Roldao & Egan, 1992). Broadly it is understood as a discipline which includes content from multiple subject fields. However, there is much variation in how social studies is conceptualised, in terms of both terminology and definition, as well as in what subject content it is considered to encompass and how it is structured and organised. This article first outlines the history of social studies as a school subject to understand its origin as a discipline. We will then discuss some of the issues around conceptualising social studies as a school subject, including different terms and definitions that have been used. Following this, we examine different approaches that have been taken to social studies as a subject in terms of what content is included and how it is organised.

History of social studies

To understand social studies as a subject in schools, it is useful to look at the history of its introduction. Most educational historians consider social studies to be an American invention and its origins as a subject can be traced back to the early twentieth century in the United States (US) (Roldao & Egan, 1992; Ravitch, 2003). A 1916 bulletin titled *The Social Studies in Secondary Education* was published by the US Bureau of Education and is viewed as seminal in the development of social studies in the US, and then ultimately around the world (M. Nelson, 1988). History, which was already a subject taught in schools in the US, was then integrated into social studies during the 1930s, alongside content from

1 The work was carried out when the author was a member of the Research Division.

geography and civics. Consequently, history subject content has driven much of social studies content and approach in many contexts.

Following the publication of the 1916 bulletin, social studies became increasingly widespread within education in the US. The global influence that the US had meant that social studies as a school subject also spread to numerous countries, often replacing or combining other subject areas. That said, much of the research on social studies education has been and remains US-centric (Parry, 1999).

This global trend of social studies adoption is illustrated by a comparative study of social science subjects in jurisdictions across the world during the period 1900–86 (Wong, 1991). This study looked at the presence of social science subjects in the curricula (which they defined as history, geography, civics and social studies), grouped into time periods. It found that 11 per cent (of 47 countries) had social studies as a subject in their curricula in the period 1920–44; increasing to 60 per cent (of 77 countries) in the period 1970–86. There was also a corresponding decrease in history (81 per cent to 47 per cent), geography (87 per cent to 47 per cent) and civics (40 per cent to 27 per cent) subjects, illustrating that social studies had replaced these subjects in many contexts.

However, social studies does not have the same pattern of use in all countries. Wong (1991) noted there were trends in the use of social studies curricula which were related to the region and colonial background of the countries. The shift to social studies was not seen in Eastern European countries. Additionally, while social studies largely appeared to replace history and geography in countries which had previously been Anglo–US and Spanish colonies, this was not the case in those with French colonial backgrounds.

Conceptualisations of social studies

Terminology

There is a great deal of inconsistency and ambiguity around terminology and definition of social studies and where it is considered to sit in the curriculum, in terms of overarching learning areas.

- **Social studies, singular or plural?**
  There is variation in whether social studies is used in the singular, to mean a school subject; or in the plural sense (i.e., the social studies), as an overarching term or category that includes several subjects such as history and geography (J. Nelson, 2001; Mutch et al., 2008). In some cases, it has been used in both senses. This issue is interlinked with tensions in how the relationship between social studies and other disciplines is understood. In this review, we will focus on social studies as a school subject.
- **Social studies or social science**
  A second issue is that social studies is sometimes used interchangeably with social science (Hertzberg, 1981), while other times it is seen as a subject within the overarching area of social science. Where both terms are used it is not always clear whether they are being accidentally conflated or whether they

are being used as distinct terms with a specific relationship to one another. This occurred in the New Zealand curriculum framework of 1993 where both were used with no explicit definition or explanation of their relationship to one another, leading to confusion (Sinemma, 2004).

- **Relationship with social science and humanities**
  Another challenge is the lack of clarity about the relationship between social studies, social science, and the humanities. In some cases, social studies is seen as a subject within the overarching learning area of humanities. For example, in the new Curriculum for Wales, humanities is an "Area of Learning and Experience" and social studies is a discipline within this, alongside geography; history; religion, values and ethics; and business studies (Hughes et al., 2020, p.276). In other contexts, social studies is a subject which draws from both the social sciences and the humanities (National Council for the Social Studies [NCSS], 2010).
- **Relationship with science**
  Sometimes, social studies type subjects are grouped into an overarching area with sciences. For example, in the primary curriculum for Ireland, there is an overarching curriculum area of "Social, Environmental and Scientific Education" which has science, geography and history subjects within this (National Council for Curriculum and Assessment, 1999). In other cases, social studies and science content is combined into a single subject, particularly for lower stages of education, for example Japan has a "Life Environmental Studies" subject which combines social studies and science content in grades 1 and 2 (Shimura, 2015, p.152).
- **Relationship with civics/citizenship**
  Civics or citizenship education (amongst other names) often forms a key part of social studies curricula. Citizenship can be used for the purpose of social studies curricula; as a subject discipline that feeds into social studies; or as a standalone subject which exists instead of, or as well as social studies. For example, Singapore has both a social studies subject and a citizenship subject (Brant et al., 2016).
- **Social studies or another term?**
  Social studies is not the only term used when discussing subject curricula in this area (Brant et al., 2016). Sometimes, this is by including another subject in the title, such as "history and social studies" in previous iterations of the Finnish Curriculum (Löfström, 2019, p.89). In others, alternative terms such as *society* are used as the subject name, for example Queensland previously had "Studies of Society and Environment" as an integrated subject area drawing from a variety of subjects typically included in social studies (Brant et al., 2016, p.67).
- **Translational issues**
  It is apparent from literature looking at non-English speaking countries that in many contexts there are subjects which may be translated to mean social studies, but which could arguably be translated to another term. For example, the Danish subject of "samfundsfag" is translated to social studies by the Danish Ministry of Education, but arguably could also be translated

to social sciences (Hansen, 2020, p.96). This introduces further ambiguity in what is understood by social studies.

Definitions

There has also been much variation in how social studies is defined in different contexts, and even at different time points within contexts. J. Nelson (2001) identifies three categories of definitions and gives examples within these:

- Defining social studies in terms of the basic purpose, for example citizenship, social criticism, social responsibility.
- Defining social studies in terms of knowledge structure dimensions, for example history, law education, social science, humanities, integrative social knowledge.
- Defining social studies in terms of instructional or curricular criteria, for example critical thinking, issues-centred, multicultural studies.

While these categories overlap, it highlights that there are various ways of conceptualising social studies. This leads to significant variation in the content and structure of social studies curricula. There is a wide variety of definitions of social studies, and some of these are outlined below.

1916 bulletin definition

The definition given in the 1916 bulletin is useful to refer to as this has had a broad impact on understanding of social studies. They use the definition: "the social studies are understood to be those whose subject matter relates directly to the organization and development of human society, and to man as a member of social groups" (US Bureau of Education, 1916, as cited in M. Nelson, 1988, p.20).

National Council for the Social Studies

The National Council for the Social Studies (NCSS), formed in the US in 1921, is a professional association dedicated to promoting social studies. In 1994 they published *Standards for Social Studies: A framework for Teaching, Learning and Assessment*, which has subsequently been revised, most recently in 2010 (NCSS, 2010), as well as a framework for social studies State Standards (NCSS, 2013). Within their *Standards for Social Studies* document they define social studies as:

> the integrated study of the social sciences and humanities to promote civic competence. Within the school program, social studies provides coordinated, systematic study drawing upon such disciplines as anthropology, archaeology, economics, geography, history, law, philosophy, political science, psychology, religion, and sociology, as well as appropriate content from the humanities, mathematics, and natural sciences. The primary purpose of social studies is to help young people make informed and reasoned decisions for the public good as citizens of a culturally diverse, democratic society in an interdependent world.
>
> (NCSS, 1994, p.3).

The NCSS has been highly influential on understandings of social studies, particularly in the US (Rutherford & Boehm, 2004). The NCSS Standards are intended to support the development and teaching of social studies both as a standalone subject, and for integration into subject discipline-based classes such as history and geography.

Dictionary definition

It is also useful to examine dictionary definitions of social studies as they show how the term *social studies* is commonly understood. The *Cambridge Dictionary* (n.d.) defines it as: "in the US, a course for younger students that includes many of the social sciences". This highlights that social studies is perceived as an American concept in many contexts.

An alternative dictionary definition offered by Merriam-Webster (n.d.) understands social studies as an overarching area rather than a subject and specifies a variety of subjects that fall into this area: "a part of a school or college curriculum concerned with the study of social relationships and the functioning of society and usually made up of courses in history, government, economics, civics, sociology, geography, and anthropology".

Meanwhile, the *Collins English Dictionary* (n.d.) makes a distinction between social studies in Britain and in the US, outlining differences in the subjects that are included in social studies: "In Britain, social studies is a subject that is taught in schools and colleges, and includes sociology, politics, and economics. In the United States, social studies is a subject that is taught in schools, and that includes history, geography, sociology, and politics".

These dictionary definitions highlight some of the differences in understandings of social studies definitions such as whether it is understood as a subject and what other disciplines it is linked to.

Social studies traditions

The aims or purposes of specific social studies curricula play a large role in determining the content included, and in how social studies is conceptualised. There are several "traditions" of social studies curricula that have been identified in the literature (Ross et al., 2014; J. Nelson, 2001). One influential piece of work by Barr, Barth & Shermis (1977, as cited in J.Nelson, 2001) identified three key purposes or traditions of social studies; this was then expanded to five traditions by Martorella (1996, as cited in Nelson, 2001):

- **Citizenship (or cultural) transmission**
  This tradition understands the purpose of social studies as promoting national values and ideas of "good citizenship", with students taught a generally accepted body of factual knowledge. It focuses on promoting cultural and social unity and gives less attention to diversity of experience (Ross et al., 2014; Barr, et al., 1997).

- **Social science**
  This tradition views the purpose of social studies as teaching students the key rules, principles, generalisations, and processes of social science disciplines such as political science, history, economics and geography. This focuses on teaching students techniques for gathering, processing, and applying information. Arguably much of the work in this area has focused on history, education and skills (Barr et al., 1997; Ross et al., 2014).
- **Reflective inquiry**
  This tradition views reflective inquiry as the purpose of social studies, equipping students with decision-making and problem-solving skills to use in their lives (Barr et al., 1997; Ross et al., 2014).
- **Informed social criticism**
  This perspective considers the purpose of social studies as challenging the status quo and addressing injustices. Social studies is intended to provide students with the opportunity and skills to examine and critique the past and present. This tradition gives more weight to teachers' and students' own experiences, including cultural knowledge and understandings (Ross et al., 2014). This tradition is sometimes merged with the reflective inquiry approach (e.g., Barr et al., 1997), as instructional methods used in this tradition focus on encouraging reflective and critical thinking.
- **Personal, social, and ethical development**
  This tradition sees the purpose of social studies as empowering students to face problems in today's world, focusing on helping them to develop a positive self-concept and self-efficacy. It is grounded in ideas of democratic citizenship, highlighting personal freedoms and responsibilities (Barr et al., 1997; Ross et al., 2014).

Many curricula can be said to be drawing from each of these traditions to different extents (Mutch et al., 2008; Barr, et al., 1997), although it is generally agreed that citizenship transmission has historically dominated much of social studies education (Ross et al., 2014).

Structure and content of social studies curricula

Subjects within social studies

Social studies draws from a range of other disciplinary subjects. Therefore, another important part of defining social studies is considering its relationship with other subject disciplines, and which subjects it draws content from. Where a social studies subject exists, there are two broad approaches for including content from subject disciplines within it (Brant et al., 2016; Hughes et al., 2020):

- An interdisciplinary approach—considering the subject disciplines to be related but distinct; this may be done by including them as strands within the social studies curriculum with connections drawn between them.
- An integrated or unified approach—focusing on the skills and types of thinking that are common across the disciplines included within the social studies curriculum.

History, geography and civics/citizenship

Social studies as a subject is predominantly made up from history, geography and civics/citizenship content, particularly in the US (M. Nelson, 1988). This explains why social studies curricula and research globally tend to focus on these areas. An example of this is seen in the curriculum comparisons review by Brant et al. (2016). While they acknowledge that other subjects are present in some social studies curricula, they focus on history, geography, and civics/citizenship as these are often considered core elements of social studies. They found a common tension in whether history and geography have their own disciplinary identity, or whether they are incorporated into an overarching social studies subject, which usually focuses on citizenship. That said, while both history and geography are frequently included in social studies curricula, they are not included everywhere. For example, in Denmark, social studies does not include geography and history, and instead focuses on politics, economics, sociology and international politics (Hansen, 2020).

History has largely been the dominant subject in social studies content, often focusing on transmission of historical knowledge and facts (J. Nelson, 2001; Brant et al., 2016). However, some historians feel very strongly that history should retain its status as a discipline rather than being taught entirely within the framework of a social studies subject or learning area. For example, Smith (2016) argued that including history content in social studies rather than as a standalone subject represents two competing purposes: history for its extrinsic utility, focusing on socialisation, understanding the self and cross-curricular learning; versus history as a discipline, providing an epistemic framework for uncovering the past and the pursuit of rigorous engagement with evidence. Geography is also a common component of social studies, but there is not as much contention around whether it should be a standalone subject or included within social studies. Brant et al. (2016) noted that in many of the jurisdictions, both geography and history were taught as foundations for civics or citizenship. That said, geography is not taught within social studies in all contexts. For example, in Finland it is instead aligned to the sciences, and focuses on physical geography content (Brant et al., 2016).

Civics or citizenship education (amongst other names) often forms a key part of social studies curricula. Social studies is often used to transmit ideologies and belief systems. Brant et al. (2016) found that a focus on citizenship, and promotion of national identity and sometimes patriotism were present to varying degrees in the different curricula. A further challenge here is defining what citizenship means. In many countries there is a specific emphasis on ideas of democratic citizenship (J. Nelson, 2001), however this is not relevant in all contexts.

There have been changes in how themes of citizenship are represented in social studies courses. Wong's (1991) review of social studies curricula found that while ideas of instilling national spirit, pride and patriotism had historically been the focus of curricula in many countries, social science curricula (including social studies) have increasingly shifted to ideas of responsible citizenship. Similarly, Lerch et al. (2017) reviewed history, civics and social studies textbooks from 78 countries from 1950 to 2011 and found that while references to social structures like democracy remained, there was an increase in references to human agency

and rights. They argued that this represents a core cultural shift from ideas of obedient citizenship to ideas of active and empowered individuals with rights and responsibilities. Similarly, themes around the environment are increasingly seen in social studies. For example, Bromley et al. (2011) analysed social studies textbooks in 65 countries and found that themes of environment had increased, alongside a shift to emphasising social issues and human rights above national citizenship. Additionally, many researchers have called for social studies to move from ideas of national citizenship to ideas of global citizenship instead (e.g., Myers, 2006).

Other subjects linked to social studies

While history, geography and civics are very commonly included and central in social studies curricula this literature review found reference to social studies as encompassing content from a wider variety of subjects and topics. As discussed, the NCSS references social studies as drawing from "anthropology, archaeology, economics, geography, history, law, philosophy, political science, psychology, religion, and sociology, as well as appropriate content from the humanities, mathematics, and natural sciences" (NCSS, 1994, p.3). Other subjects that have been discussed in the context of social studies include criminology (Solhaug et al., 2020) and tourism (Jaber & Marzuki, 2019). Altogether, there is a great deal of variation in which subjects are discussed as part of social studies in both research and curriculum.

Subjects and stages

Another area where there is variation, is whether social studies is taught as a subject across both primary and secondary levels. Some contexts have social studies as a subject throughout schooling, for example in Alberta (Hayward et al., 2018). However, in many curricula there is social studies as a subject at primary school level; with subjects such as history and geography taught separately from lower or upper secondary (Brant et al., 2016; Hughes et al., 2020; Hayward et al., 2018). Sometimes these replace social studies, in other cases social studies continues to be taught alongside these additional disciplinary subjects (e.g., New Zealand). Where social studies is replaced or supplemented by subject disciplines at later stages of learning, this can be done in one level or over several (Hayward et al., 2018).

Components of a social studies curriculum

There is much variation in what is included within social studies curricula (Hayward et al., 2018; Hughes et al., 2020; Brant et al., 2016). There are different approaches to balancing knowledge and skills, with some social studies curricula also including values or dispositions (Hughes et al., 2020). Hayward et al. (2018) report that there was an overall tendency to emphasise "inquiry" skills, particularly in some jurisdictions. In some contexts, social studies curricula focus on knowledge content (Brant et al., 2016). Vogler and Virtue (2007) suggest that many social studies frameworks in the US are overloaded with knowledge content, and so teachers focus on factual content, rather than higher level thinking skills. This issue has also been noted in Canada (Brant et al., 2016). Many curricula bring together the subjects that can make up social studies through focusing on the skills that they

have in common, rather than bodies of knowledge. In some contexts, this focus on skills has been contentious. For example, in New Zealand in the 1990s, a draft for a new social studies curriculum provoked criticism that there was an emphasis on skills at the expense of content (Crittenden, 1998).

Hayward et al.'s (2018) review also looked at the presence of "big ideas" (also referred to as key areas, themes or ideas) within the curricula, how prominent their role was, and whether they were subject-specific or not. They considered big ideas to be core concepts which underpin the curriculum. They noted big ideas in various forms in Scotland, Singapore, Australia, Ontario, and British Columbia. In many cases, these big ideas spanned across grade levels, while in others (e.g., Ontario) there were also specific big ideas at each grade level.

Literacy

There is also discussion about the place of literacy or language arts within social studies. In the US, there has been an increased emphasis on integration of literacy with social studies. This has been partly because the Common Core State Standards have called for greater literacy integration into science and social studies, and accordingly have published English language arts standards for history / social studies (Lee & Swan, 2013). However, there are concerns that integrating literacy into the social studies curriculum can lead to a focus on teaching literacy skills, such as reading comprehension, at the expense of social studies specific skills and knowledge (McGuire, 2007).

Curriculum models

Social studies does not have intrinsic levels of progression to the same extent that subjects such as mathematics do. There is limited research on progressions in social studies, as the focus tends to be on progression within individual disciplines such as history and geography (Hughes et al., 2020). Consequently, there is much variation in how social studies curricula are organised and structured (Brant et al., 2016). Various curriculum models have been used, which also relate to the different traditions of social studies:

- **Expanding horizons model**—This model has been influential in social studies, particularly in the US (Rutherford & Boehm, 2004). There are various interpretations and terms used for this approach including "expanding environments" and "widening interests" (LeRiche, 1987). Broadly, this model argues that children should move from the known to the unknown, beginning by learning about familiar contexts, and expanding from the self out to the world. This model has been subject to various criticisms; it is argued that it is based on outdated theories of child development (LeRiche, 1987), that it does not apply well to the study of history (Krahenbuhl, 2019) and that it needs to be modified to take into account the technological and social changes of the modern world (Clarke et al., 1990).
- **Chronology model**—An approach commonly used in the US has been termed the "chronology model" (Rutherford & Boehm, 2004). This is rooted in the origin of social studies in the US and tendency to use history as the

central component of social studies. This model centres on history, with the curriculum organised by historical periods, with content from other disciplinary subjects such as geography linked to these. Consequently, it is criticised for allowing history to dominate social studies and treating other disciplinary content as secondary.

- **Core knowledge sequence**—This approach stems from the work of Hirsch and is a content-based approach to social studies (Rutherford & Boehm, 2004). A key limitation is how to select the content, and skews in focus can arise based on the disciplinary background of the curriculum developers. It has also been criticised for not taking into account cognitive dimensions of learning, and for focusing on the aggregation of knowledge. Additionally, there are concerns that it may focus on the "knowledge of the powerful" (Brant et al., 2016). "Knowledge of the powerful" refers to the idea that those who have the power in society define what is considered knowledge and who has access to it (Young, 2012).
- **Cognitive taxonomies**—Cognitive taxonomies such as Bloom's taxonomy have been used for social studies, organising the curriculum around types of thinking (Brant et al., 2016). However, this approach has been criticised for failing to include substantive knowledge and there are concerns that it does not consider that there are domain-specific dimensions to conceptual knowledge (Brant et al., 2016).
- **Subject-specific disciplinary thinking**—Another approach has been to consider progression in terms of subject-specific disciplinary thinking. This approach focuses on mastery of the concepts and processes, or epistemologies that are core to the particular disciplines included in the social studies curriculum. However, there are concerns that this artificially separates disciplines, with many competencies not unique to specific disciplines, and misses the opportunity to draw connections between these (Brant et al., 2016).
- **Body and form**—Brant et al. (2016) suggest that another approach is understanding knowledge as both "body and form", such as in "historical literacy" (Lee, 2011, as cited in Brant et al., 2016). There is limited explanation about this approach, but they suggest that it understands a subject as being both a body of knowledge, and a form of knowledge, with these interacting with each other and being of equal importance.
- **Spiral curriculum**—Several references were found to the use of spiral curriculum models for social studies. Broadly, this approach suggests that rather than being organised by disciplines or chronology, content should be organised in "spirals" of key concepts and skills, with progression from familiar concepts and skills to increasing abstraction (Parry, 1999, 2007; Matrai & Szebenyi, 1987).
- **NCSS National Curriculum Standards and C3 Framework**—The NCSS National Curriculum Standards and C3 Framework can also be used as an organising model for social studies. The National Curriculum Standards (2010) were developed to provide a conceptual framework for conceptual design and development of social studies curricula. The standards are

organised around 10 thematic strands. The C3 Framework was developed later and is intended to provide guidance on key concepts and inquiry skills that should be incorporated into social studies curricula, in the US context, including more discipline specific guidance (NCSS, 2013). The C3 Framework and National Curriculum Standards can be used alongside each other, with the C3 Framework building upon the National Curriculum Standards, as well as giving more disciplinary specific guidance (Herczog, 2013).

Representations of gender and race in social studies curricula

It is important to consider representations of gender within social studies. Bernard-Powers (2001) discusses the role for social studies in gender equity, through considering representations of gender dynamics and identities, highlighting gendered issues, and addressing gendered knowledge in social studies curricula. However, she argues that while the issue of gender equity in social studies has been discussed for decades, there remains substantial work to be done on this. Similarly, Engebretson (2014) critiques the NCSS standards for social studies for representations of gender. She highlights that they do not give explicit guidance on gender, focus on binary representations of gender, and have a gender imbalance with men over-represented among the notable people in the content they recommend covering.

Race and ethnicity have often been neglected in social studies (Branch, 2004; Howard, 2003); however, it is important that these are considered as part of social studies work. Branch (2004) highlights that although teachers may avoid discussing race and ethnicity in the classroom, social studies has an important role to play in affirming students' racial and ethnic identities and experiences. Howard (2003) suggests that social studies is well placed to discuss and address issues of race and that it has an important role to play in helping students to understand societal issues of inequality, discrimination, and racism. He argues that since social studies encompasses and often focuses on issues relating to citizenship, it is important that race and racism are discussed as part of this.

In some contexts, such as the US and Australia, where there are indigenous or aboriginal populations that have been marginalised, indigenous perspectives have been excluded or characterised in problematic ways. For example, Sharp (2013) reviewed references to Indigenous Australians in social studies textbooks in Australia from the 1960s to 1980s. She found a tendency towards tokenistic mentioning, and representing Indigenous Australians monoculturally. Where Indigenous Australians were discussed, they were often "othered" and shown as primitive or savage.

Controversial topics

A significant challenge for social studies curricula is how to handle topics that are controversial. This can be linked to the need to discuss race and ethnicity as outlined above; such topics are often avoided due to discomfort with the subject matter. It can be challenging to decide whether an issue is controversial or not. Issues that are considered as non-controversial by some, may be viewed as controversial by others, and this is ultimately affected by underlying ideologies

and bound by place and time. Whether topics are presented as controversial or not within a curriculum frames them as either open to discussion, or as being closed and conventionally agreed upon. Camicia (2008) argues that curriculum developers must continually consider the question "controversial to who, where and when" (p.312).

Colonialism is one example of a topic that may be considered controversial. Masta (2016) considered the issue of colonialism being taught in social studies curricula in a classroom in the US. They found it was often erased (not discussed at all) or normalised (presented as an inevitable and usual process). In social studies, there can be a tendency to attempt to avoid or to deliver a "neutral" approach to challenging topics such as colonialism, rather than engage in critical analysis, but by erasing or normalising colonialism the curriculum can perpetuate harmful colonial ideologies that marginalise ethnic minorities and allow damaging colonial legacies to continue.

Conclusion

This review sought to understand the various ways social studies has been defined and conceptualised as a school subject. Overall, there is a lack of consensus around terminology and definition of social studies and a great deal of variation in how it has been conceptualised and approached in different contexts. History, geography, and civics/citizenship seem to be the subject content most frequently included in social studies. However, there are numerous other subjects that have been included, and expectations about what is part of social studies vary across contexts. The diversity of understandings and approaches to social studies pose a challenge for educators and researchers as it is difficult to compare social studies across different countries and cultures, and to define social studies in an overarching sense. When discussing social studies, it is crucial to clearly and explicitly define the way in which it is being conceptualised. This is important in order to avoid misconceptions arising due to differences in how social studies is commonly understood in different contexts.

References

Barr, H., Graham, J., Hunter, P., Keown, P., & McGee, J. (1997). *A position paper: Social studies in the New Zealand school curriculum.* Prepared for the New Zealand Ministry of Education by the School of Education, University of Waikato, New Zealand. https://nzcurriculum.tki.org.nz/content/download/569/4032/file/social-studies-positions.doc

Bernard-Powers, J. (2001). Gender in the social studies curriculum. In E. W. Ross (Ed.), *The social studies curriculum: Purposes, problems, and possibilities* (pp.177–199). State University of New York Press.

Branch, A. J. (2004). Modeling respect by teaching about race and ethnic identity in the social studies. *Theory & Research in Social Education, 32*(4), 523–545. https://doi.org/10.1080/00933104.2004.10473268

Brant, J., Chapman, A., & Isaacs, T. (2016). International instructional systems: social studies. *The Curriculum Journal, 27*(1), 62–79. https://doi.org/10.1080/09585176.2015.1134340

Bromley, P., Meyer, J. W., & Ramirez, F. O. (2011). The worldwide spread of environmental discourse in social studies, history, and civics textbooks, 1970–2008. *Comparative Education Review, 55*(4), 517–545. http://doi.org/10.1086/660797

Cambridge Dictionary. (n.d.). Social studies. In *Cambridge Dictionary.* Retrieved December 12, 2020, from https://dictionary.cambridge.org/dictionary/english/social-studies

Camicia, S. P. (2008). Deciding what is a controversial issue: A case study of social studies curriculum controversy. *Theory & Research in Social Education, 36*(4), 298–316. https://doi.org/10.1080/00933104.2008.10473378

Clarke, G., Sears, A., & Smyth, J. (1990). *A proposal to revise the elementary social studies curriculum.* Faculty of Education, University of New Brunswick, Canada. https://files.eric.ed.gov/fulltext/ED318661.pdf

Collins Dictionary. (n.d.). Social studies. In *Collins Dictionary*. Retrieved December 12, 2020, from https://www.collinsdictionary.com/dictionary/english/social-studies

Crittenden, B. (1998). Social studies: The plan for New Zealand's schools. *Agenda, 5*(2), 189–200.

Engebretson, K. E. (2014). Another missed opportunity: Gender in the national curriculum standards for Social Studies. *Social Studies Research & Practice, 9*(3), 21–34. http://www.socstrpr.org/wp-content/uploads/2015/01/MS06578Engebretson.pdf

Hansen, M. (2020). Social studies in Denmark: A country report. *Journal of Social Science Education, 19*(1), 95–117. https://doi.org/10.4119/jsse-1581

Hayward, L., Jones, D. E., Waters, J., Makara, K., Morrison-Love, D., Spencer, E., Barnes, J., Davies, H., Hughes, S., Jones, C., Nelson, S., Ryder, N., Stacey, D., Wallis,

R., Baxter, J., MacBride, G., Bendall, R., Brooks, S., Cooze, A., ... Wardle, G. (2018). *CAMAU project: Research report: Learning about progression – Informing thinking about a curriculum for Wales*. University of Glasgow and University of Wales Trinity Saint David. https://eprints.gla.ac.uk/163362/

Herczog, M. M. (2013). The links between the C3 framework and the NCSS national curriculum standards for social studies. *Social Education, 77*(6), 331–333. https://www.socialstudies.org/system/files/publications/articles/se_7706331.pdf

Hertzberg, H.W. (1981). *Social Studies Reform 1880-1980.* SSEC Publications.

Howard, T. (2003). The dis(g)race of the social studies: The need for racial dialogue in the social studies. In G. Ladson-Billings (Ed.), *Critical race theory perspectives on the social studies: The profession, policies, and curriculum* (pp.27–43). Greenwich, CT: Information Age Publishing.

Hughes, S., Makara, K., & Stacey, D. (2020). Learning progression in the humanities: identifying tensions in articulating progression in humanities in Wales. *The Curriculum Journal, 31*(2), 276–289. https://doi.org/10.1002/curj.28

Jaber, H. M., & Marzuki, A. (2019). Improving awareness of tourism education among students in intermediate and secondary schools in the Kingdom of Saudi Arabia: Experts' social studies curricula point of view. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies, 10*(3), 351–359. https://tuengr.com/V10/351.pdf

Krahenbuhl, K. S. (2019). The problem with the expanding horizons model for history curricula. *Phi Delta Kappan, 100*(6), 20–26. https://doi.org/10.1177/0031721719834024

Lee, J., & Swan, K. (2013). Is the common core good for social studies? Yes, but... *Social Education, 77*(6), 327–330. http://www.c3teachers.org/wp-content/uploads/2016/05/Lee-Swan-CommonCore.pdf

Lerch, J., Bromley, P., Ramirez, F. O., & Meyer, J. W. (2017). The rise of individual agency in conceptions of society: Textbooks worldwide, 1950–2011. *International Sociology, 32*(1), 38–60. https://doi.org/10.1177/0268580916675525

LeRiche, L. W. (1987). The expanding environments sequence in elementary social studies: The origins. *Theory & Research in Social Education, 15*(3), 137–154. https://doi.org/10.1080/00933104.1987.10505542

Löfström, J. (2019). Yhteiskuntaoppi: Social studies in Finland: A country report. *Journal of Social Science Education, 18*(4), 22–101. https://doi.org/10.4119/jsse-1583

Masta, S. (2016). Disrupting colonial narratives in the curriculum. *Multicultural Perspectives, 18*(4), 185–191. https://doi.org/10.1080/15210960.2016.1222497

Matrai, Z., & Szebenyi, P. (1987). *The basic principles and model of the integrated spiral social science programme.* https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/1096/file/Matrai_Szebenyi_Social_Science_Programme.pdf

McGuire, M. E. (2007). What happened to social studies? The disappearing

curriculum. *Phi Delta Kappan, 88*(8), 620–624. https://doi.org/10.1177/003172170708800815

Merriam-Webster. (n.d.). Social studies. In *Merriam-Webster Dictionary*. Retrieved December 12, 2020, from https://www.merriam-webster.com/dictionary/social%20studies

Mutch, C, Hunter, P, Milligan, A, Openshaw, R, Siteine, A. (2008). *Understanding the social sciences as a Learning Area – A position paper.* New Zealand Ministry of Education. http://nzcurriculum.tki.org.nz/content/download/2609/34003/file/SSPP%20final%2031%20July%2009%5B1%5D.doc

Myers, J. P. (2006). Rethinking the social studies curriculum in the context of globalization: Education for global citizenship in the US. *Theory & Research in Social Education, 34*(3), 370–394. https://doi.org/10.1080/00933104.2006.10473313

National Council for Curriculum and Assessment. (1999). *Curriculum Online*. Government of Ireland. https://www.curriculumonline.ie/Primary/Curriculum/

National Council for the Social Studies. (1994). *Expectations of excellence: Curriculum standards for the social studies.* https://files.eric.ed.gov/fulltext/ED378131.pdf

National Council for the Social Studies. (2010). *National curriculum standards for social studies: Executive summary.* https://www.socialstudies.org/standards/national-curriculum-standards-social-studies-executive-summary

National Council for the Social Studies. (2013). *The college, career and civic life (C3) framework for social studies state standards: Guidance for enhancing the rigor of K-12 civics, economics, geography and history.* https://www.socialstudies.org/standards/c3

Nelson, J. L. (2001). Defining social studies. In W. B. Stanley (Ed.), *Critical issues in social studies research for the 21st Century* (pp.15–37). Information Age Publishing.

Nelson, M. R. (1988). *The social contexts of the committee on social studies report of 1916.* https://files.eric.ed.gov/fulltext/ED315329.pdf

Parry, L. (1999). Transplanting practices in social studies: an historical case study of curriculum development and reform efforts in Australia, 1969–1989. *Theory & Research in Social Education, 27*(4), 505–528. https://doi.org/10.1080/00933104.1999.10505892

Parry, L. (2007). Selling the new social studies in Australia: Experiences, perceptions and challenges in the professional development of teachers in the 1970s and 1980s. *Journal of In-Service Education, 33*(1), 7–21. https://doi.org/10.1080/13674580601157554

Ravitch, D. (2003). A brief history of social studies. In J. S. Leming, L. Ellington, & K. Porter (Eds.), *Where did social studies go wrong?* (pp.1–6). Thomas B. Fordham Foundation. https://files.eric.ed.gov/fulltext/ED481631.pdf

Roldao, M. D. C., & Egan, K. (1992). *The social studies curriculum: The case for its*

abolition [Paper presentation]. Annual meeting of the American Educational Research Association. https://files.eric.ed.gov/fulltext/ED350208.pdf

Ross, E. W. Mathison, S., & Vinson, K. D. (2014). Social studies curriculum and teaching in the era of standardization. In E. W. Ross (Ed.), *The social studies curriculum: Purposes, problems and possibilities* (4th ed., pp.25–49). State University of New York Press, Albany.

Rutherford, D. J., & Boehm, R. G. (2004). Round two: Standards writing and implementation in the social studies. *The Social Studies, 95*(6), 231–238. https://doi.org/10.3200/TSSS.95.6.231-238

Sharp, H. L. (2013). What we teach our children: A comparative analysis of Indigenous Australians in socia studies curriculum, from the 1960s to the 1980s. *Social and Education History, 2*(2), 176–204. http://dx.doi.org/10.4471/hse.2013.11

Shimura, T. (2015). Primary geography education in Japan: Curriculum as social studies, practices and teachers' expertise. *Review of International Geographic Education Online, 5*(2), 151–165. https://eric.ed.gov/?id=EJ1158052

Sinemma, C. E. L. (2004). *Social sciences, social studies or a new term? The dilemma of naming a learning area.* (Curriculum Marautanga project). New Zealand Ministry of Education. https://nzcurriculum.tki.org.nz/content/download/570/4035/file/dilemma-naming-learning.pdf

Smith, J. (2016). What remains of history? Historical epistemology and historical understanding in Scotland's Curriculum for Excellence. *The Curriculum Journal, 27*(4), 500–517. https://doi.org/10.1080/09585176.2016.1197138

Solhaug, T., Borge, J. A. O., & Grut, G. (2020). Social science education (Samfunnsfag) in Norway: A country report. *Journal of Social Science Education, 19*(1), 47–68. https://www.jsse.org/index.php/jsse/article/view/1748/3555

Vogler, K. E., & Virtue, D. (2007). "Just the facts, ma'am": Teaching social studies in the era of standards and high-stakes testing. *The Social Studies, 98*(2), 54–58. https://doi.org/10.3200/TSSS.98.2.54-58

Wong, S. Y. (1991). The evolution of social science instruction, 1900–86: A cross-national study. *Sociology of Education, 64*(1), 33–47. https://doi.org/10.2307/2112890

Young, M. (2012). What are schools for? In H. Daniels, H. Lauder, & J. Porter (Eds.), *Knowledge, values and educational policy: A critical perspective* (pp.20–28). Routledge.

Learning during lockdown: How socially interactive were secondary school students in England?

Joanna Williamson (Research Division), Irenka Suto, John Little, Chris Jellis (Cambridge CEM), and Matthew Carroll (Research Division)

Introduction

In England, one of the Government's responses to the COVID-19 pandemic, alongside broader national restrictions, was for schools (both state and independent) to be closed to all but the children of key workers and a small number of other children identified as vulnerable. The working population was also expected to work from home if they were able to. This situation came to be known colloquially as "lockdown". There were two separate school lockdowns, the first starting in March 2020 and a second in January 2021. During the second lockdown in particular, schools were expected to make proactive provision for student learning to continue (Montacute & Cullinane, 2021; Williamson, 2021; Leahy et al., 2021), and the emphasis for many was moved to independent learning and home schooling.

For many students, the closing of schools caused serious upheaval in their studies. The advent of widespread schooling at home is commonly believed to have placed great burdens on individual students who often had to take much more responsibility for their own learning than they had done previously. There was a much greater reliance on technology, broadband internet access and the presence and availability of appropriate devices (laptops, tablets and phones). There was the problem that these resources were often shared with other members of the family, including, potentially, parents working from home. Many students were also impacted by repeated periods of self-isolation from anyone outside their own household, due to close contact with confirmed COVID-19 cases (often within school), particularly in the periods of time between and following the national lockdowns.

In an attempt to find out more about the experiences of secondary school students and their teachers during the lockdown period in early 2021, and to compare these experiences with those during their subsequent return to school, we conducted research investigating behaviours and attitudes during

this extraordinary time. In this article, we report on the data we collected from students on their social interactions. Our study of teachers' wellbeing is reported separately (Jellis et al., this issue).

What do we know about student experiences during lockdown?

The unprecedented nature and size of the pandemic meant that most governments had extremely difficult decisions to make, in an unusually short timescale. The United Kingdom (UK) Government's decision to close schools in England in 2020 and 2021, and to expect most students to continue their education from their homes with as much support as could reasonably be put in place by their school, caused a great deal of concern[1] (e.g., Andrew et al., 2020). There was speculation as to how the lockdowns would affect the education and ultimately the life chances of school students, particularly those who were close to taking their GCSE or A Level examinations. In addition, teachers were placed in a position where they were required to plan, deliver their lessons, and mark work in ways that were unfamiliar to them. Again, there were concerns about the quality of teaching and marking under these conditions (Howard et al., 2021, pp.60–61). Consequently, a number of research projects were commissioned, both nationally and internationally, to discover and evaluate any learning loss, or lost opportunities for learning and therefore loss of learning progress, that would potentially occur.

An early study in Norway (Bubb & Jones, 2020) concluded that, although working remotely had put greater pressures on students and teachers, it also provided opportunities not previously apparent. Pupils reported on the autonomy they had gained and that it allowed them to make more decisions for themselves as to when and how to do things. Teachers too, reported pedagogical benefits, including the ability to spread their attention equally among their students rather than tending more to the most demanding. In England, recently published reviews have emphasised that student experiences of learning during lockdown were highly diverse (Leahy et al., 2021; Howard et al., 2021). During the first national lockdown, research confirmed that socio-economically disadvantaged students were spending far less time learning than less disadvantaged peers (Leahy et al., 2021), and were less likely to be learning through live online lessons. Comparatively, little research has been carried out into teaching and learning during the second national lockdown specifically (Howard et al., 2021), but the available evidence suggests that students on average spent more hours per day learning than during the first lockdown (Leahy et al., 2021), and that live online lessons were available more frequently and to more students than during the first lockdown (Nelson et al., 2021; Teacher Tapp, 2021). Despite improved remote teaching provision and specific efforts to mitigate inequalities, discrepancies in student experiences remained. Leahy et al. (2021, p.7) report that "students from middle-class families [were] nearly 1.5 times more likely to be spending more than five hours per day learning than students from working-class families" during the second lockdown, and students in more deprived schools remained far less likely

1 Similar decisions were made in Scotland, Wales and Northern Ireland, engendering similar concerns.

than students in the least deprived schools to have access to digital devices at home (Coleman, 2021; Nelson et al., 2021). This was presumed to be an important factor in accessing and engaging with remote provision: Nelson et al. found that "students in the most deprived schools were still less likely than students in the least deprived schools to attend the online lessons (59% and 78%, respectively), and return set work (47% and 67%, respectively)" (Nelson et al., 2021, cited by Howard et al., 2021, p.39). Although socio-economic factors were judged to be dominant, both recent reviews (Leahy et al., 2021; Howard et al., 2021) stress the difficulty of generalising about student experiences to any groups of students, due to extensive variation at multiple levels—regional, local, school and student.

In terms of the impact of these lockdown experiences on learning, some studies (Kuhfeld & Tarasawa 2020; Pensiero et al., 2020) drew on previous research of learning loss during the summer holidays to estimate the learning loss that had occurred during the lockdown period. The Pensiero study made estimates of learning loss for school students in the UK, with the caveat that actual figures would be highly dependent on the socio-economic group in which the student fell. Far less loss was predicted for the higher socio-economic groups: around 14 per cent of a standard deviation, with 28 per cent for the lower socio-economic groups. The Kuhfeld study from the United States of America (USA) reported extremely early (in April 2020) and suggested that, based on summer learning loss studies, students would return after lockdown with around 70 per cent of the expected learning in reading and around 50 per cent in mathematics. However, since this was based on summer learning loss research, it assumed that no learning at all took place during the time in lockdown, which may not have been the case. Another study, a joint project between the Education Endowment Foundation (EEF) and the Fischer Family Trust (Weidmann et al., 2021) related student performance on the standardised Progress in Reading and Language Assessment (PIRA) and Progress in Understanding Mathematics Assessment (PUMA) tests compared to a previous test taken pre-pandemic in April 2019; the authors found no measurable difference in pupil performance. In contrast, a separate study commissioned by the EEF (Rose et al., 2021, p.1) reported a large degree of "loss", amounting to a loss of two months' progress in both mathematics and reading.

Learning loss aside, another aspect concerning researchers, teachers, parents, and other stakeholders has been the wellbeing of students during this unprecedented time. Several studies have attempted to address this concern. A qualitative study by researchers at the University of York (Kim et al., 2020) highlighted anxieties that teachers have had about their students during lockdown, particularly those whose parents were key workers (in professions such as health and social care, food provision, the police force and other public services) and were left alone at home all day. Other concerns revolved around a perceived lack of engagement among students, work not being completed, and some groups not being responsive at all, and hard to reach. A short briefing note produced by the University College London (UCL) education unit (Moss et al., 2020) raised concerns about student wellbeing and welfare due to pressures on parents' reliance on food banks, long working hours or job loss. In addition, they mentioned issues with those families who were technologically disadvantaged,

who had no internet access and were hard to reach. A large study from the south of England (Mansfield et al., 2021) covering 19,000 school pupils across the age ranges concluded that lockdown had a greater negative impact on students in secondary schools compared with those from primary schools. Those from secondary schools scored lower for happiness, management of schoolwork and loneliness.

The story, therefore, is a mixed one. The evidence to date suggests that for some students, supported by both family and socio-economic advantages such as access to technology and a working space at home, the impact of lockdown on wellbeing and learning loss may have been fairly modest. For those who are technologically disadvantaged, or for whom family life has been more substantially impacted (e.g., via COVID-19 illness itself, parental career loss or reduced income, or parental overwork / risk in a key worker role), the situation is potentially far more bleak.

The present study

In the present study, we explored secondary school students' reflections on their lockdown learning experiences following the second national lockdown in England, which took place from early January 2021 until March 2021. We focused upon the extent and nature of their social interactions, which we hypothesised had changed markedly compared with during normal, pre-pandemic schooling. In normal, non-pandemic schooling, students interact with peers and teachers within lessons and school-directed activities (e.g., assemblies, school-based sports), but school attendance may involve numerous other social interactions besides these, such as interacting with non-teaching staff, or socialising with friends while travelling to and from school. As well as contributing to the nature of pedagogy and learning, social interactions affect interpersonal wellbeing. This is known to be an important component of overall wellbeing for school students; that is, of how they feel about themselves and their schools (McLellan & Steward, 2015). Key questions of interest related to whether there were differences in the patterns of social interactions of those students who learned mostly or entirely at home and those who spent time in school during lockdown (perhaps due to their parents' occupations or being identified as vulnerable).

Method

We devised and administered a short survey for secondary school students, with data collection taking place in May 2021. Respondents were recruited via a post on the Cambridge CEM[2] website asking for volunteer schools in England to take part in the research. Responding schools were sent letters of invitation explaining the research. Those agreeing to participate were provided with a link to the survey and were asked to allow their Year 10 to Year 13 students aged between 14

2 Cambridge CEM (Centre for Evaluation and Monitoring) is a leading provider of assessment and monitoring systems including baseline, attitudinal, diagnostic and entrance tests.

and 18 years to complete it.

The survey took approximately 10 minutes to complete. Students were asked to which of the four year groups (Year 10, Year 11, Year 12 or Year 13) they belonged. They were then asked:

- Where did your lessons take place during lockdown? (i) Mostly or entirely at school; (ii) Mixture of school and home; or (iii) Mostly or entirely at home.

The survey then suggested nine types of activity in which students might reasonably be expected to engage, either at home or at school. The first four of these were types of learning activities that occur in most lessons in English schools in "normal" times: whole class activities, working in small groups, working pairs, and working independently. They vary in terms of the number of people with whom students have the opportunity to interact. The fifth activity type, that of one-to-one conversations with teachers, also relates specifically to schooling (both at home and in the school building) but is not necessarily an activity within lessons: one-to-one conversations between teacher and student might take place (briefly) within a lesson, as a passing conversation in the school corridor, at the start or end of a lesson, or as a separately scheduled meeting. The remaining four activity types were: exploring new ideas and areas of interest, spending time with family, spending time with friends (this can be online), and relaxing / doing leisure activities alone. These activities were not related solely to schooling, although all but the final activity type could be interpreted by students to relate to both educational and non-educational activities. They were chosen in order to capture any potential changes in social interactions more generally. As noted previously, school attendance in pre-pandemic times occasioned social interaction in more than just classroom-based activities, and so to investigate the changes to social interaction associated with changes to schooling it was important not to limit the survey's scope to school-directed or solely educational activities.

For each of the nine activity types, students were invited to respond to three questions:

- How much time did you spend on the following activities during lockdown, compared with normal schooling outside lockdown?
- How helpful were these activities to you during lockdown?
- How much of these activities do you think you need over the coming months, compared with how much of them you had during lockdown?

In Question 2, students were not asked to distinguish between academic progress and wellbeing, but to make an overall judgement reflecting the extent to which each activity had been worthwhile. In Question 3, similarly, we expected respondents to think holistically. Responses were given using 5-point Likert scales. For Question 1, the response options ranged from "Much less time" to "Much more time" with an "Unsure" option. For Question 2, the response options ranged from "Really unhelpful" to "Really helpful" with an "Unsure/not applicable" option (abbreviated to "Unsure/NA" throughout). For Question 3, the response options ranged from "Much less" to "Much more" with an "Unsure" option. The survey did not allow respondents to skip questions.

Participants

Just over 600 students from eight different schools in England took part in the survey. Three schools, labelled A, B and C, provided the majority of the replies, each contributing over 100 responses. The remaining five schools (grouped as Schools R) had a total of 124 student responses among them. Responses were received from students across Years 10 to 13, with more responses from students in Years 10 and 12. The breakdown by school and year group is shown in Table 1. The sample was not nationally representative of England's school population, with independent school students over-represented.

Table 1: Students surveyed by school and year group.

	School type	Number of responses				
		Year 10	Year 11	Year 12	Year 13	Total
School A	State-maintained, single sex	86	47	41	11	185
School B	Independent	108	0	0	48	156
School C	Independent	52	12	48	27	139
Schools R	2 Independent schools; 1 Academy; 1 Free school; and 1 FE College	18	20	66	20	124
Total		264	79	155	106	604

Results and discussion

Locations of lessons

As mentioned previously, students were first asked where their lessons had taken place during lockdown. As would perhaps be expected, the distribution is heavily skewed towards students who spent their lockdown time at home. Eight of the 14 students in the sample who answered "Mostly or entirely at school" were in a single school. Nationally, the average rate of on-site school attendance for secondary school students during the early 2021 lockdown was 5 per cent [3]; the average rate that could be expected from the survey respondents (assuming near full-time attendance from 2.3 per cent and some attendance from a further 20 per cent) does not seem too dissimilar.

Table 2: Where did your lessons take place during lockdown (3 groups)?

Location	N responses	Percentage
Mostly or entirely at school	14	2.3%
Mixture of school and home	121	20.0%
Mostly or entirely at home	469	77.7%

3 https://explore-education-statistics.service.gov.uk/find-statistics/attendance-in-education-and-early-years-settings-during-the-coronavirus-covid-19-outbreak/2021-week-3

The focus of this investigation was on examining the differences between students' remote learning experiences during lockdown and their subsequent experiences upon their return to school. Given this focus and the very few responses from students whose lockdown lessons took place mostly or entirely in school, we decided that we would combine the first two rows of Table 2 and distinguish just two groups in our analyses: students whose lockdown learning took place mostly or entirely at home (77.7 per cent) and students whose lockdown learning was not mostly at home, that is, included time in school (22.3 per cent).

Overview of students' experiences of activity types

The overall profile of the 604 participating students' responses to the three questions about activity types is shown in Figure 1. The first column of bars gives an overall impression of how much time was spent on each activity type during lockdown, compared with during normal, pre-pandemic schooling. In line with our hypothesis, students reported that both the extent and nature of their social interactions changed markedly.

Strikingly, around two-thirds of students reported spending much more time working independently. This finding coheres with Bubb and Jones (2020) who reported increased autonomy among school students in Norway. Conversely, a similar proportion reported spending either less time or much less time working with others, in small groups or as a whole class. Together, these findings may suggest that for many students, there were large parts of the school day during which remote interactive teaching via Zoom[4] and other technologies did not take place. Lessons may have been provided in written format or pre-recorded, taking the form of lectures rather than interactive sessions. The biggest drop in frequency of schooling activity was for working in pairs. This is perhaps unsurprising, given the likely difficulties around arranging this during remote learning. It appears to have been replaced by independent working, rather than by whole class activities via Zoom, for example. Also, two-thirds of students reported spending less time or much less time in one-to-one conversations with teachers.

4 Zoom is a popular video teleconferencing software program used widely during the pandemic.

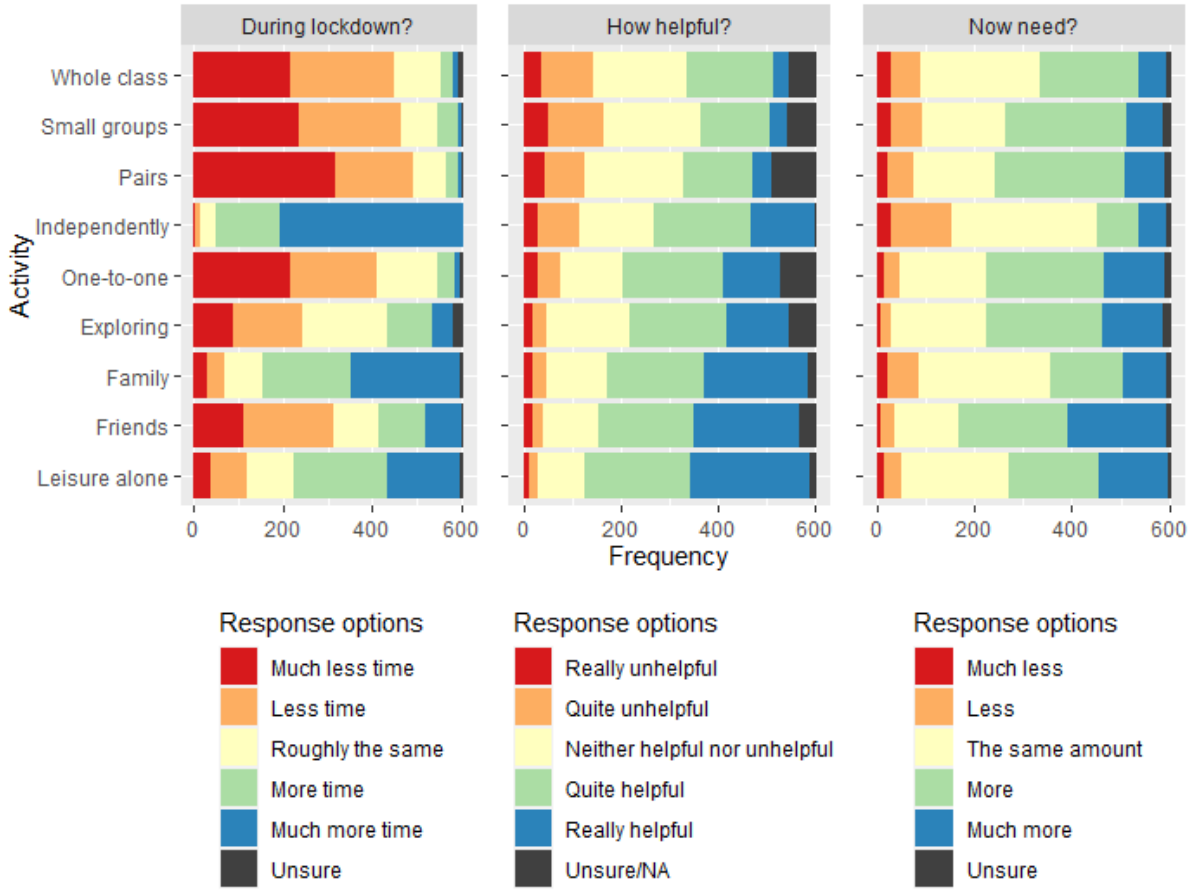


Figure 1: Visual representation of responses to the three main student activity questions.

The second column of Figure 2 indicates how helpful the students found these types of educational activities during lockdown. Perhaps surprisingly, over half of the students found working independently to be helpful or really helpful. Again, however, this is in line with Bubb and Jones's (2020) finding of enthusiasm during the pandemic for increased autonomy among Norwegian students. This can be taken as a positive finding, given how much this type of activity was reported to have increased. One-to-one conversations with teachers during lockdown were also reported to be helpful or really helpful by over half of the students. Since students reported spending less time in such conversations, it does not come as a surprise that the majority reported needing more or much more time for one-to-one conversations with teachers once back in school.

Interestingly, almost a sixth of students responded "Unsure/NA" to the question of how helpful they found working in pairs. Together with the large reported drop in the frequency of pair-work during lockdown, the high level of "Unsure/NA" responses could indicate that this type of interactive learning ceased completely for many of these students. Indeed, Figure 1 gives the general impression of an association between the proportion of "Unsure/NA" responses to each question on helpfulness (within column 2) and the change in frequency of the activity type to which it relates (column 1). This would suggest that many students selected this response because the activity was very much reduced or even ceased altogether during lockdown, rather than because it occurred but they were uncertain of

its value. The outcomes of an analysis of the "Unsure/NA" data supported this idea: the mean scores for the time spent on activity types were in almost all cases much lower for the "Unsure/NA" respondents than for the rest of the cohort who provided a measure-based answer.

The students were least positive about the helpfulness of working in small groups during lockdown; approximately a quarter of them reported that this activity was either unhelpful or really unhelpful. This could indicate that this activity type works least well in a remote format, possibly due to technological difficulties, the engagement levels of other students, or limits around how well teachers can monitor groups. However, without an indication of students' views on small group working in pre-pandemic schooling, firm conclusions cannot be drawn.

Looking beyond the types of learning activities that occur in most English classrooms in normal times, the first column of Figure 2 shows that during lockdown, over two-thirds of the students reported spending more time with their families. The second column indicates that most of these students found this to be helpful or really helpful.[5] Almost two-thirds of the students reported spending more time or much more time alone. This finding aligns with (Kim et al., 2020) who highlighted teachers' concerns that some of their students were left alone for long periods during lockdown. Around half of the students reported spending less time or much less time with friends, and approximately a third reported spending less time or much less time exploring new ideas and areas of interest. The latter reported decrease in time could be hypothesised to be due to reduced time with friends and / or the reduction in interactive learning activities described above. However, it was beyond the scope of the present study to test these hypotheses. It is apparent, however, that the reported increase in time spent working independently was not associated with students spending more time exploring new ideas and interests.

The third column of Figure 2 indicates what the students thought they needed over the coming months, compared with how much of the activities they had during lockdown. It can be seen that the students were broadly positive or neutral about all the activity types included in the survey. That is, there was an overall desire for more time on all activities, with no net negative responses to any activity type. The students' responses were least positive for working independently: over a quarter thought they needed less or much less of this activity and approximately half were neutral about it. Arguably it is surprising that so few students wanted less of it, given how much independent working they had experienced during lockdown. It is possible that many became used to it and discovered its value during that time, but exploring this possibility was outside the scope of our research.

5 While it is possible that the few students who reported spending less time with their families found this to be helpful, the number of such students is too small to alter this interpretation of the figure.

Students' experiences of activity types by year group

Table 1 showed that there were respondents from all four targeted year groups: over 250 were Year 10 students, around 80 were in Year 11, roughly another 150 were Year 12s, and just over 100 were Year 13s. Figure 2 shows students' responses broken down by year group, including the "Unsure" and "Unsure/NA" responses. As in Figure 1, the responses to each item are presented as a multicoloured bar. This time, however, each bar represents a year group rather than all respondents, and the coloured sections of each bar represent proportions rather than numbers of responses. This is to facilitate comparisons, since N varied across the year groups (see Table 1).

Overall, the figure shows a high level of consistency across year groups. Despite this broad similarity, there are some emergent patterns. In particular, the extent to which students reported much less, or less, time in one-to-one conversations with teachers decreased with increasing student year group, suggesting that older students received closer to "normal" levels of one-to-one conversations with teachers than their younger peers. The proportion of students reporting that they found one-to-one conversations with teachers helpful during lockdown also increased with age, with younger students more likely to respond "Unsure/NA" (a logical response for those not experiencing much or any of this activity) or that it was unhelpful. Figure 2 also shows some association between increasing student year group and wanting more time to explore new ideas, and wanting to spend more time with friends.

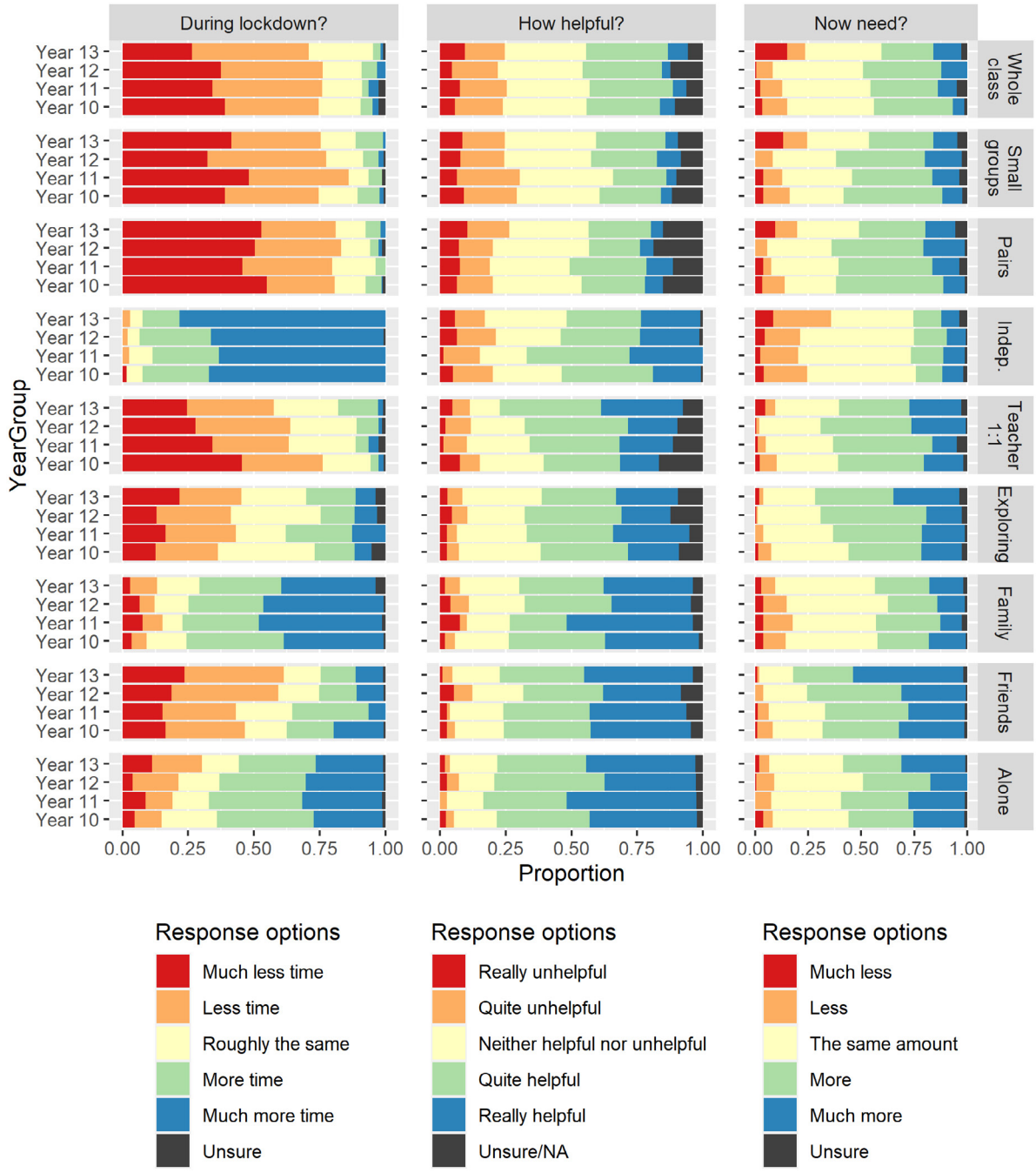


Figure 2: Responses to the three main student activity questions, by respondent year group.

To check for between-school variation, responses were also compared by school (Appendix 1). There was a high level of similarity across schools in the extent to which respondents found activities helpful, and the time they wanted to spend on activities once back at school. Responses about the time spent on independent working during lockdown varied very little between schools, but there were moderately large differences in the amount of time spent on one-to-one conversations with teachers and in small-group working.

Relationship between lockdown location and the types of activity students spent more or less time on during lockdown

Key questions of interest related to whether there were differences between those students who learned mostly or entirely at home and those who spent time in school during lockdown. Accordingly, Figure 3 shows students' responses to the question of how much more or less time they spent on each activity (Question 1), according to the location of their lockdown learning. Note that in Figure 3, proportions of respondents have been centred on zero to emphasise the reported decreases or increases in time spent on the activity types during lockdown. Bars shifted to the left of the zero line indicate a balance of decreasing time on that activity type, while bars shifted to the right indicate a balance of increasing time on that activity type.

Strikingly, Figure 3 shows that both groups of respondents showed similar patterns for the different activity types. That is, the broad patterns identified in Figure 1 held for both groups, a finding also supported by the similarity of means and standard deviations of responses from each group (see Table 3, Appendix 2).

As explained in the Method section, the first five activity types (Figure 3) relate to schooling and occur in most English classrooms in normal times. Students who attended school at least some of the time during lockdown reported spending similar amounts of time on each of these five activity types to those students who learned mostly or entirely from home. The general trend of spending less time in interactive learning activities within lessons (whole class, small groups, and pair-work) and more time working independently was common to both groups, as was spending less time in one-to-one conversations with teachers. It follows that the balance of activity types for those attending school during lockdown appears to have been quite different from what it had been prior to the pandemic. The nature of face-to-face schooling appears to have changed in the direction of remote schooling. This may be because teachers wanted to treat their students as fairly and consistently as possible, and / or did not have time to prepare pedagogical activities in multiple formats.

Nonetheless, some small but potentially relevant differences were observed. Respondents who had worked mostly at home gave a larger proportion of "Much less time" responses for pair-work, small group work, and one-to-one conversations with teachers. Such patterns may be expected due to the reduced social contact associated with being predominantly at home. Although the differences are relatively minor, the different groups of respondents did experience their lessons during lockdown a little differently.

The differences between the two groups were just as small for the four activity types that did not relate solely to schooling. That is, for spending time with friends and family, exploring new ideas, and spending time alone, the distributions of time were broadly similar for students who learned mostly or entirely at home and those who spent time in school during lockdown. Two small but unsurprising differences in the groups can be seen in Figure 3. Respondents who had worked mostly from home gave a larger proportion of "Much less time" responses for time with friends, and a larger proportion of "Much more time" responses for time with family.

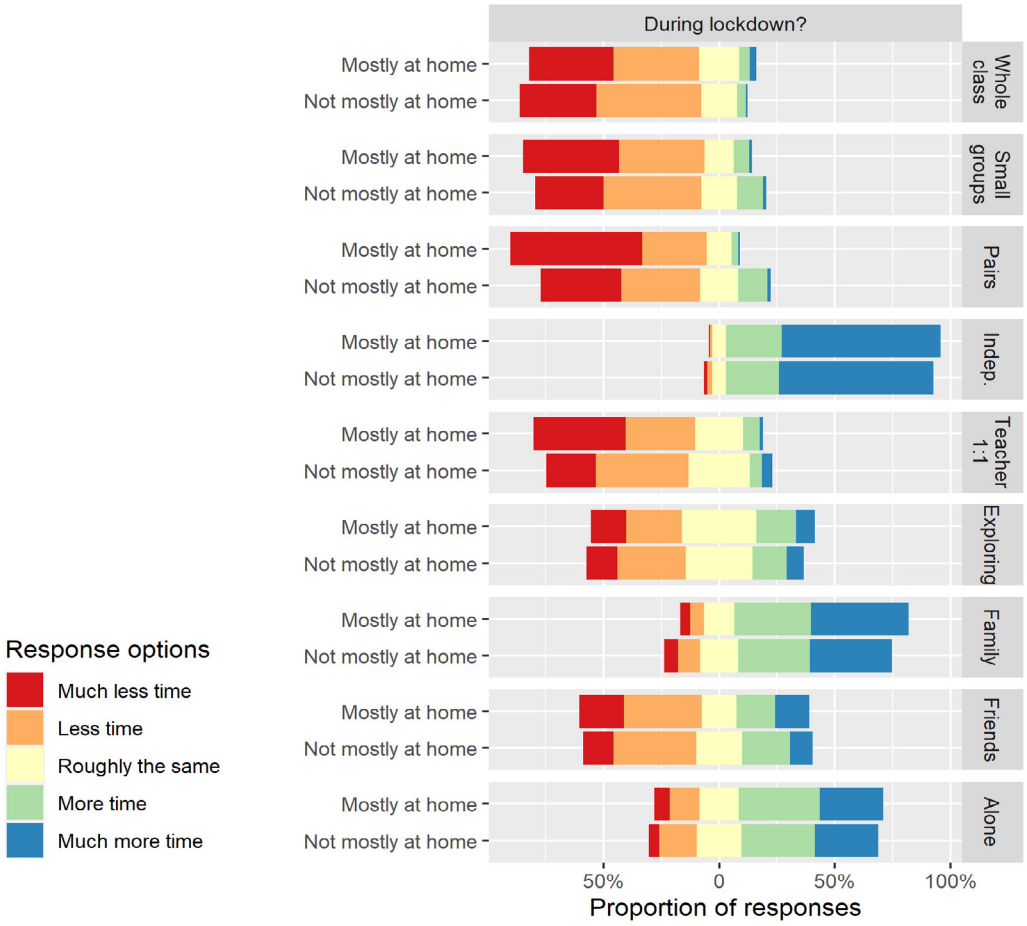


Figure 3: Responses to Question I ("How much time was spent on the activities during lockdown?"), broken down by the main location of learning during lockdown. Bars are expressed as proportions of respondents in each group, and are centred on zero so that bars to the left indicate a reduction in time spent, and bars to the right indicate an increase in time spent; the further left the bar lies, the greater the proportion of respondents that spent less time on that activity.

Relationship between lockdown location and which types of activity were found helpful

The analyses described above establish that there were only minor differences in the time spent on different activity types between the different groups of students. It is feasible, however, that even if broad patterns of time use were similar, students' experiences of those activity types may have differed depending on the location of lockdown learning. Accordingly, Figure 4 breaks down responses to Question 2 ("How helpful was the activity?") by location of learning. Note that interpretation of the figure is the same as for Figure 3, but as relatively more respondents answered "Unsure/NA" for this question, each bar does not sum to 1.

Figure 4 shows that the perceived helpfulness of activity types was similar between the groups, but some slightly larger differences were apparent than in Figure 3. Students who spent lockdown mostly at home gave much greater proportions of "Really helpful" responses to time spent with friends and time spent alone. This finding is most explicable for time spent with friends, as students spending most of their time at home would see friends less, making any social time much more valuable. The finding that time alone was also more helpful was less expected, and perhaps relates to having to share space with other family members. This highlights the multiple functions of time at school, providing social time alongside learning, but also opportunities for young people to have space and time to themselves. In terms of teaching activities, some differences were evident in small group work and pair-work, where students who were mostly at home showed greater proportions of negative responses. This could relate to the challenges of conducting such activities online, where technical limitations hinder "natural" group interaction, or to the simple fact that the activities were less frequently conducted under remote learning. Work in pairs again showed the biggest difference between group mean scores (Table 4), Appendix 2, while the only activity type to get a mean score lower than 3, indicating a net negative opinion, was working in small groups for those students who spent lockdown at home.

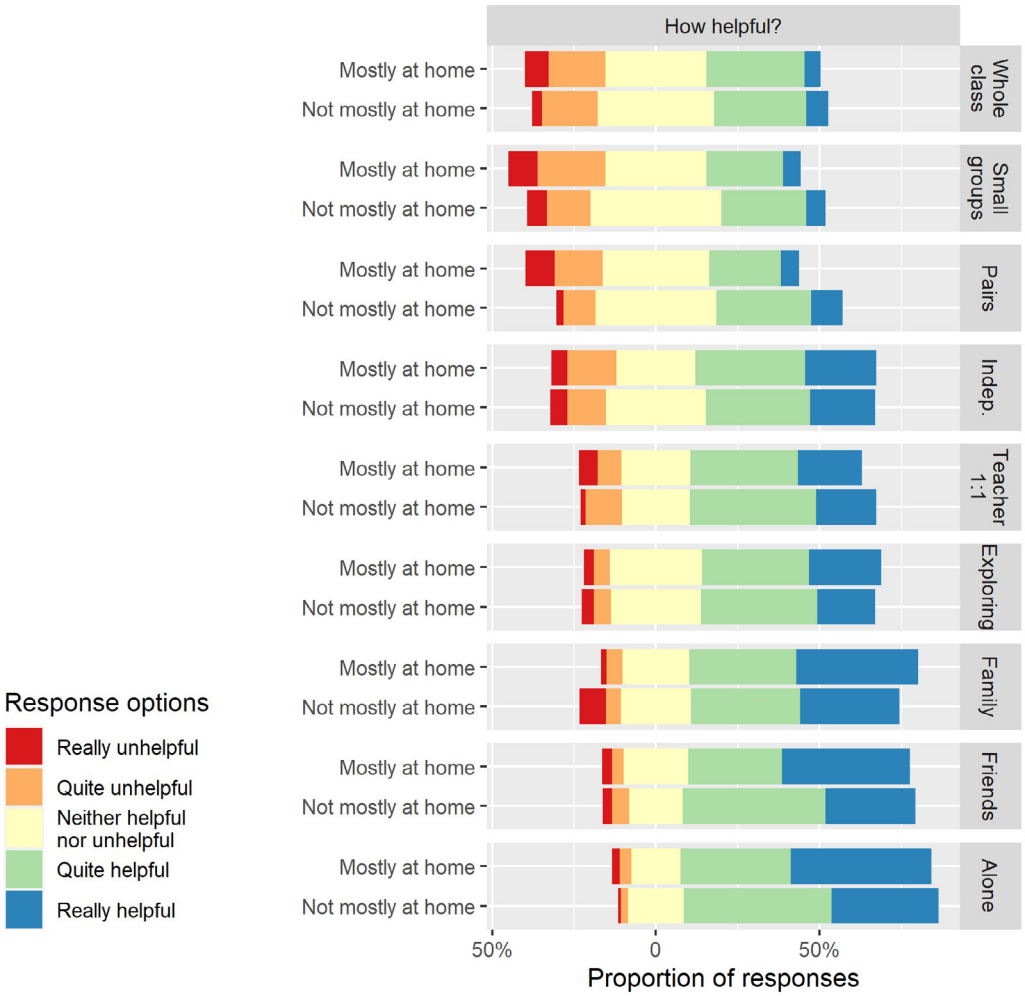


Figure 4: Responses to Question 2 ("How helpful were the different activities?"), broken down by the main location of learning during lockdown. Bars are expressed as proportions of respondents in each group and are centred on zero so that bars to the left indicate more unhelpful activities, and bars to the right indicate more helpful activities. Note that as the proportion of "Unsure/ NA" responses varied between activities and these responses were removed, bars are of different lengths.

Relationship between lockdown location and which types of activity students wanted more of afterwards

The final question in the survey related to which activities students wanted more of, once schools had fully reopened. Figure 5 breaks down responses to Question 3 into the two groups considered so far. Interestingly, although this analysis of the previous questions identified some differences between the groups, this plot highlights just how similar the two groups were in respect of what activities they wanted more of. As noted in the whole-sample analysis, there was an overall desire for more time on all activities except independent working, with neither group showing an overall negative response to any activity.

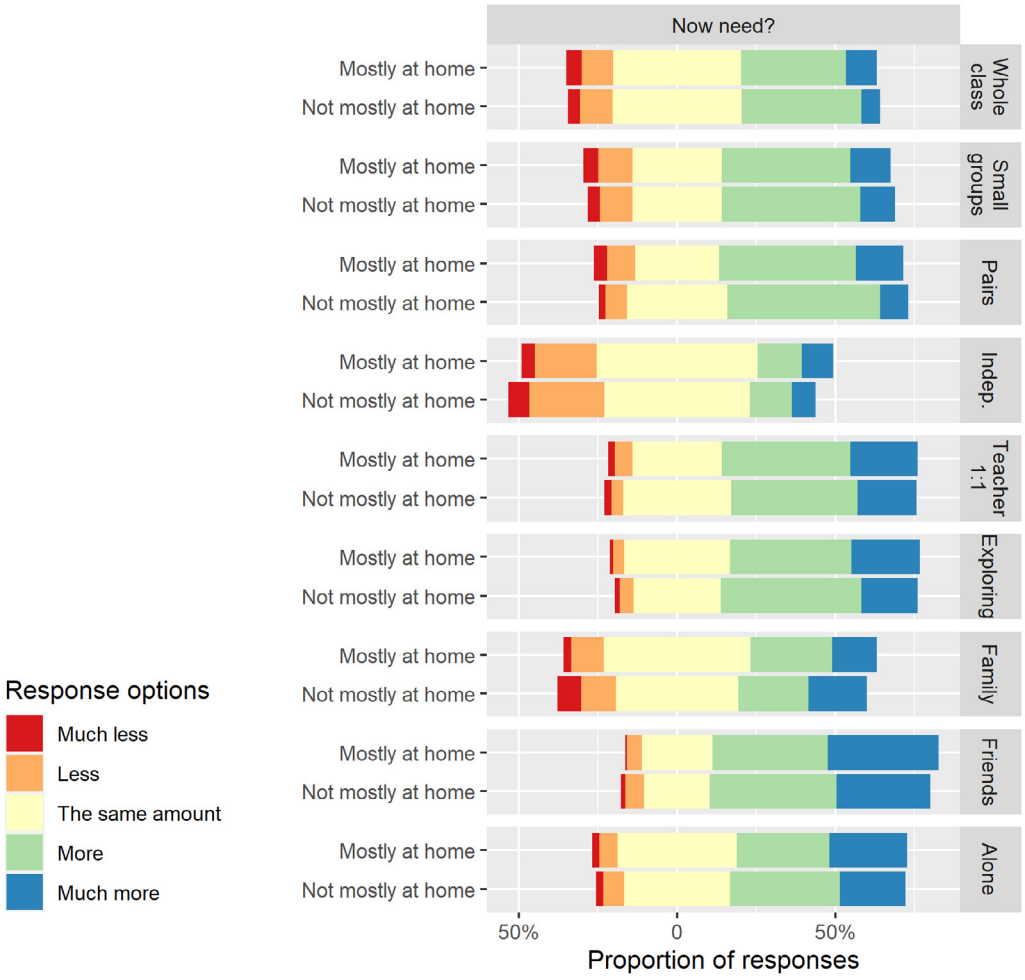


Figure 5: Responses to Question 3 ("How much time should be spent on the activity types in coming months?"), broken down by the main location of learning during lockdown. Bars are expressed as proportions of respondents in each group, and are centred on zero so that bars to the left indicate activity types where less time should be spent, and bars to the right indicate activity types where more time should be spent.

Intriguingly, one of the more evident differences comes from "time with family", where students who spent some time at school gave a slightly greater proportion of "Much more time" responses, although this is also offset by a greater proportion

of "Much less time" responses. Indeed, when mean values are compared (see Table 5, Appendix 2), the values are identical for six of the nine activities, and for the three showing any difference, the difference is only 0.1 or 0.2.

The lack of difference between the groups is, in itself, an interesting finding. It was anticipated that the different experiences of students under lockdown would lead to differing needs going forward. However, these findings imply that the students effectively wanted the same things once schools reopened, regardless of where they spent lockdown. This may, therefore, show that students predominantly wanted a return to "normality" following the challenges of lockdown, rather than missing specific aspects of their school experience.

Limitations

When evaluating the results and discussion above, the study's main limitation should be borne in mind. Although the overall data set was of a reasonable size and was collected from eight different schools, over three-quarters of the respondents to our survey attended just three schools. Thus, the sample was not nationally representative of England's school population. The survey was conducted at a time when teachers were extremely busy collating their students' performance data to provide them with GCSE and A Level grades. Although this is likely to have influenced many schools' decision to participate, delaying the study to a less busy time would have had a negative impact on the validity of the data collected, since students' memories of lockdown would probably have faded.

A further limitation of the data relates to the number of student respondents who participated in face-to-face schooling during lockdown. Nationally, the rate of on-site attendance for secondary school students was low (5 per cent), and our respondents included very few whose lockdown learning took place mostly or entirely in school, although a larger group reported a mix of learning from home and in school. Whether students who attended face-to-face schooling during lockdown were slightly under- or over-represented in our data, the conclusions that can be drawn are limited by the fact that the absolute number of responses from such students was small.

General discussion and conclusions

The results of this study indicate that the 600 students who took part experienced a marked decrease in the extent and type of their social interactions during England's lockdown of early 2021. While we cannot say how typical these changes were or what happened nationally, we were unable to identify a compelling reason to assume that they were substantially different. For example, if substantial reductions in pair-work and small group work were experienced in the well-resourced schools in our study, then it seems likely that many schools in the state-maintained sector found it similarly difficult to continue these interactive activities during lockdown.

Reductions of this kind are of great concern given that the pedagogical benefits of peer tutoring are very well established. Peer tutoring includes a range of approaches within pair-work and small group work, in which students provide

each other with explicit teaching support (Education Endowment Foundation, 2021). Pre-COVID-19, its introduction in schools had been found to have an average positive effect equivalent to approximately five additional months' progress, with low-attaining students and those with special educational needs making the biggest gains (Education Endowment Foundation, 2021). Opportunities of this kind appear to have reduced in 2020 and 2021, even among vulnerable children who attended school during lockdown and potentially need it most.

Just as importantly, our findings of reported reductions in the extent and range of interactive activities, both during and outside of schooling, offer a powerful explanatory mechanism for the decreases in the wellbeing of young people that have been reported since the pandemic struck (for example, Office for National Statistics, 2020). As explained previously, interpersonal wellbeing is known to be an important component of overall wellbeing for school students (McLellan & Steward, 2015). We would suggest that it is an important topic for further pandemic-related research.

Finally, perhaps the most positive finding of our study was a strong general trend for students wanting more of all the activity types explored, except independent learning (although even for independent learning, students seemed to think post-lockdown levels were about right, and over half found it helpful during lockdown). Since this finding could be an effect of school type, it would be interesting to research this further among a larger, nationally representative sample of students. Could there really be an increased desire and respect for education as a result of the lockdown, among students as well as those parents who had to home-school?

Ethical considerations

The research was conducted in full accordance with the principles stated in the research team's institutional *Research Ethics Guidance* document. This included obtaining the necessary consent for students' participation (from parents / guardians, or from students themselves if aged 16+) and protecting their privacy.

Acknowledgements

We would like to thank Filio Constantinou and Kate Bailey for their input into the design of the survey. We are also grateful to Hannah North, Antje Diestelhorst and Kayleigh Lauder for their administrative support.

References

Andrew, A., Cattan, S., Costa-Dias, M., Farquharson, C., Kraftman, L., Krutikova, S., Phimister, A., & Sevilla, A. (2020). *Learning during the lockdown: real-time data on children's experiences during home learning* (IFS Briefing Note BN288). The Institute for Fiscal Studies. https://ifs.org.uk/uploads/BN288-Learning-during-the-lockdown-1.pdf

Bubb, S., & Jones, M. A. (2020). Learning from the COVID-19 home-schooling experience: Listening to pupils, parents/carers and teachers. *Improving Schools, 23*(3), 209–222. https://doi.org/10.1177/1365480220958797

Coleman, V. (2021). *Digital divide in UK education during COVID-19 pandemic: Literature review.* Cambridge Assessment Research Report. https://www.cambridgeassessment.org.uk/Images/628843-digital-divide-in-uk-education-during-covid-19-pandemic-literature-review.pdf

Education Endowment Foundation. (2021). *Peer tutoring.* https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/peer-tutoring/#closeSignup

Howard, E., Khan, A., & Lockyer, C. (2021). *Learning during the pandemic: review of research from England* (Ofqual/21/6803/4). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/998935/6803-4_Learning_during_the_pandemic-_review_of_research_from_England.pdf

Jellis, C., Williamson, J., & Suto, I. (2021). How well do we understand wellbeing? Teachers' experiences in an extraordinary educational era. *Research Matters: A Cambridge University Press & Assessment publication, 32,* 45–66.

Kim, L., Dundas, S., & Asbury, K. (2020). *'I think it's been difficult for the ones that haven't got as many resources in their homes': Teacher concerns about the impact of COVID-19 on pupil learning and wellbeing.* Department of Education, University of York. https://psyarxiv.com/wsyqk/download?format=pdf

Kuhfeld, M., & Tarasawa, B. (2020). *The COVID-19 slide: What summer learning loss can tell us about the potential impact of school closures on student academic achievement.* NWEA. https://files.eric.ed.gov/fulltext/ED609141.pdf

Leahy, F., Newton, P., & Khan, A. (2021). *Learning during the pandemic: quantifying lost time* (Ofqual/21/6803/2). Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1000351/6803-2_Learning_during_the_pandemic-_quantifying_lost_time.pdf

Mansfield, K., Jindra, C., Geulayov, G., & Fazel, M. (2021). *Self-reported wellbeing and sample characteristics in a survey of 19000 school pupils during the first UK COVID-19 school closures.* https://psyarxiv.com/gtbfm/download?format=pdf

McLellan, R., & Steward, S. (2015) Measuring children and young people's wellbeing in the school context. *Cambridge Journal of Education, 45*(3), 307–332. https://

www.doi.org/10.1080/0305764X.2014.889659

Montacute, R., & Cullinane, C. (2021). *Research Brief: January 2021: Learning in Lockdown*. The Sutton Trust. https://www.suttontrust.com/wp-content/uploads/2021/01/Learning-in-Lockdown.pdf

Moss, G., Bradbury, A., Duncan, S., Harmey, S., & Levy, R. (2020). *Responding to COVID-19, Briefing Note 2: Learning after lockdown*. https://discovery.ucl.ac.uk/id/eprint/10111678/1/Moss_Briefing%20Note%202%20Responding%20to%20COVID-19%20learning%20through%20disruption_final.pdf

Nelson, J., Andrade, J., & Donkin, A. (2021). *The impact of Covid-19 on schools in England: experiences of the third period of partial school closures and plans for learning recovery. Graphs and commentary on questions posed to the NFER Teacher Voice*. Omnibus Survey panel, March 2021. National Foundation for Educational Research. https://www.nfer.ac.uk/media/4435/the_impact_of_covid_19_on_schools_in_england.pdf

Office for National Statistics. (2020). *Coronavirus and the social impacts on young people in Great Britain: 3 April to 10 May 2020*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/

Pensiero, N., Kelly, A., & Bokhove, C. (2020). *Learning inequalities during the COVID-19 pandemic: how families cope with home-schooling*. University of Southampton research report. https://doi.org/10.5258/SOTON/P0025

Rose, S., Twist, L., Lord, P., Rutt, S., Badr, K., Hope, C., & Styles, B. (2021). *Impact of school closures and subsequent support strategies on attainment and socio-emotional wellbeing in Key Stage 1*. Interim Pap, 1. Education Endowment Foundation and National Foundation for Educational Research. https://www.nfer.ac.uk/impact-of-school-closures-and-subsequent-support-strategies-on-attainment-and-socio-emotional-wellbeing/

Teacher Tapp. (2021). *Vaccines, meetings in lockdown, and a rare upside of the pandemic!* February 2021. https://teachertapp.co.uk/vaccines-meetings-lockdown-rare-upside-pandemic-more-money-boiler/

Weidmann, B., Allen, R., Bibby, D., Coe, R., James, L., Plaister, N, & Thomson, D. (2021) *COVID-19 disruptions: Attainment gaps and primary school responses*. Education Endowment Foundation. https://dera.ioe.ac.uk/37913/1/Covid-19_disruptions_attainment_gaps_and_primary_school_responses_-_May_2021.pdf

Williamson, G. (2021, January 6). *Education Secretary statement to Parliament on national lockdown*. GOV. UK. https://www.gov.uk/government/speeches/education-secretary-statement-to-parliament-on-national-lockdown

Appendix 1: Students' experiences of activity types by school

Figure 6 shows students' responses broken down by school. As in Figure 2, the coloured sections of each bar represent proportions rather than numbers of responses in order to facilitate comparisons, since N varied across schools (Table 1).

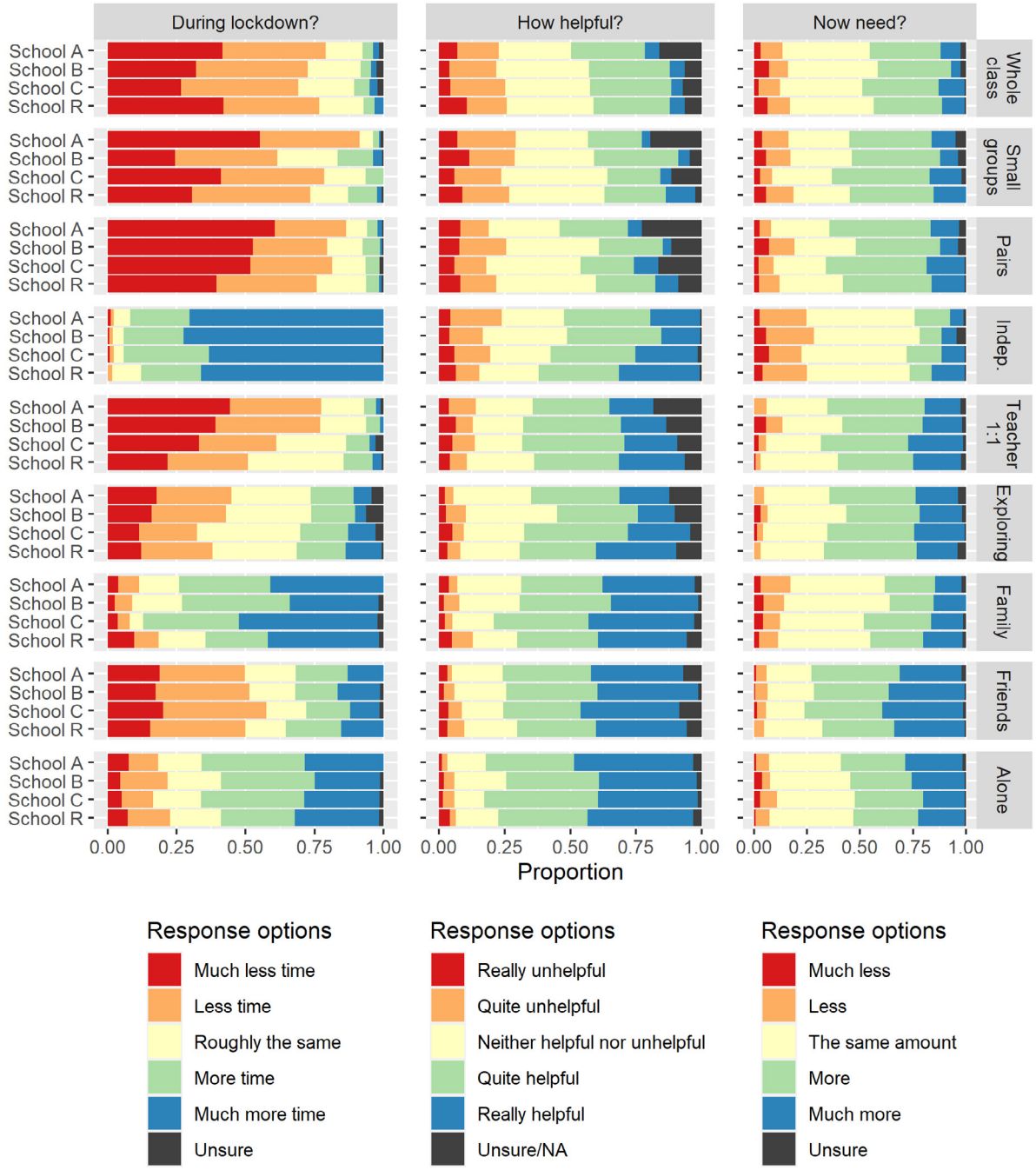


Figure 6: Responses to the three main student activity questions, by school.

Appendix 2: Responses summarised by lockdown location

To further support comparisons between students whose lockdown lessons took place mostly or entirely at home, and those who spent at least some time in school, Tables 3–5 show the mean and standard deviation of scores from each group, for each activity type.

Table 3: Mean responses to Question 1 (how much time was spent on the activities during lockdown compared with normal schooling outside lockdown). Values are derived from scoring 1 for much less time, 2 for less time, 3 for similar, 4 for more time, and 5 for much more time. Hence, a mean less than 3 indicates less time overall on that activity, and a mean greater than 3 indicates more time overall.

Activity	Not mostly at home		Mostly or entirely at home	
	Mean	SD	Mean	SD
Whole class activities	1.9	0.8	2.0	1.0
Working in small groups	2.1	1.0	1.9	1.0
Working in pairs	2.1	1.1	1.6	0.8
Working independently	4.5	0.8	4.6	0.7
One-to-one conversations with teachers	2.3	1.0	2.0	1.0
Exploring ideas/ interests	2.7	1.1	2.8	1.2
Time with family	3.8	1.2	4.0	1.1
Time with friends	2.8	1.2	2.7	1.3
Leisure activities alone	3.6	1.2	3.6	1.2

Table 4: Mean responses to Question 2 (how helpful were the different activity types). Values are derived from scoring 1 for really unhelpful, 2 for unhelpful, 3 for neither helpful nor unhelpful, 4 for helpful, and 5 for really helpful. Hence, a mean less than 3 indicates an unhelpful activity overall, and a mean greater than 3 indicates a helpful activity overall.

Activity	Not mostly at home		Mostly or entirely at home	
	Mean	SD	Mean	SD
Whole class activities	3.2	0.9	3.1	1.0

Working in small groups	3.1	1.0	2.9	1.1
Working in pairs	3.4	0.9	3.0	1.1
Working independently	3.5	1.1	3.5	1.1
One-to-one conversations with teachers	3.7	1.0	3.6	1.1
Exploring ideas/ interests	3.7	1.0	3.7	1.0
Time with family	3.8	1.2	4.0	1.0
Time with friends	3.9	1.0	4.0	1.0
Leisure activities alone	4.1	0.8	4.1	1.0

Table 5: Mean responses to Question 3 (how much time should be spent on the activity types in coming months). Values are derived from scoring 1 for much less time, 2 for less time, 3 for a similar amount of time, 4 for more time, and 5 for much more time. Hence, a mean less than 3 indicates students wanted to spend less time doing that activity type, and a mean greater than 3 indicates students wanted to spend more time doing that activity type.

Activity	Not mostly at home		Mostly or entirely at home	
	Mean	SD	Mean	SD
Whole class activities	3.3	0.9	3.3	1.0
Working in small groups	3.5	1.0	3.5	1.0
Working in pairs	3.6	0.8	3.6	1.0
Working independently	2.9	1.0	3.1	1.0
One-to-one conversations with teachers	3.7	0.9	3.7	0.9
Exploring ideas/ interests	3.8	0.9	3.8	0.9
Time with family	3.3	1.1	3.4	0.9
Time with friends	3.9	0.9	4.0	0.9
Leisure activities alone	3.7	1.0	3.7	1.0

How well do we understand wellbeing? Teachers' experiences in an extraordinary educational era

Chris Jellis (Cambridge CEM), Joanna Williamson (Research Division), Irenka Suto (Cambridge CEM)

Introduction

Much has been reported in the news and in academic circles of the effects of the COVID-19 pandemic, particularly on student wellbeing and learning in the light of the national lockdowns that resulted. Across the UK, schools were closed to most children from mid-March 2020 until at least June 2020, although the great majority of children did not return to school until September 2020 (Children's Commissioner, 2020). A second national closure took place from early January 2021 until early March 2021. During these periods, home schooling, supplemented by distance learning through the use of collaborative technologies such as Zoom, became the norm[1]. While there has been considerable interest in the effects of the pandemic and school closures on children, with so-called "learning loss" a particularly salient concern (Kuhfeld & Tarasawa, 2020; DfE, 2021; Weidmann et al., 2021, p.9), rather less attention has focused on the wellbeing of teachers and school leaders. Teaching has become bound up with the availability of broadband, knowledge and understanding of technology and the ability to control student behaviour and motivation remotely (Coleman, 2021). Additionally, many teachers have been expected to collect evidence of student knowledge and understanding in order to justify teacher assessed grades in the absence of England's usual high stakes external examinations for GCSE and A Level. Undeniably, teaching experiences have changed substantially.

In this article, we report on a study of teachers' wellbeing. We surveyed teachers about their experiences and concerns during and after England's second national school closure, during early 2021. Our aim was to improve understanding of how teachers had been impacted in these unprecedented times, and of the kinds of support that they may need.

1 While schools were closed to most children, school attendance was still permitted for some specific groups, including students from disadvantaged backgrounds, students with certain special educational needs, and the children of key workers. However, attendance statistics indicated that even among those children permitted to attend school during the national lockdowns, only a small minority actually did so (Children's Commissioner, 2020, p.1).

What is teacher wellbeing and why is it important?

The term *wellbeing* is all around us. Predominantly, it has been adopted by the media to denote such things as fitness, lifestyle, diet and good mental health. However, the psychological definition of wellbeing is not so far-ranging. Diener (2000) defines subjective wellbeing as being equivalent to the concept of living a good life, or colloquially, "happiness". At the heart of wellbeing is the concept of agency, that is, the power people have to determine their own thoughts and actions. It brings together concepts from self-determination theory (SDT), motivation theory and self-efficacy theory.

In SDT, Deci and Ryan (1985) postulated that human beings have three inherent psychological needs: competence, relatedness and autonomy:

- Competence is the ability to deal effectively with the world around you.
- Relatedness is the ability to share experiences with other people and to develop a sense of belonging.
- Autonomy concerns itself with the ability to act according to one's own sense of needs and values.

Motivation theory is a complex subject and is largely outside the scope of this article but some of its main features are associated with self-worth, subject mastery, intelligence and ability and attribution. For teachers, there are two areas where motivation could be considered particularly important: (i) their own motivation to teach; and (ii) motivating their students. In his book *Teaching and Researching Motivation* (Dörnyei, 2001, p.158), Zoltan Dörnyei stated that the combination of teacher motivation and student motivation among practising teachers has not been studied widely. He identified that teachers have to draw on a wide range of motivational skills in order to teach effectively, and these are affected by factors such as the institutional environment, time available, stress and autonomy.

Self-efficacy theory, a term first used by Albert Bandura (1977), concerns the way individuals evaluate their experiences and thoughts through a process of self-reflection. It is based on the premise that, although previous action is often considered to be the best predictor of future achievement, more important is a person's assessment of their own ability to carry out a particular task successfully.

Although general wellbeing is an established concept, the area of teacher wellbeing is not so well defined. This led to researchers considering the particular case of teacher self-efficacy, and various instruments have been created to measure this trait. Tschannen-Moran and Woolfolk Hoy (2001) reviewed several different measures and used their findings to develop their own teacher self-efficacy scale: the Ohio State Teacher Efficacy Scale (OHSTES).

Gordon and Debus (2002) argued that teachers with high self-efficacy are likely to engage in a wider range of teaching practices than teachers with low self-efficacy. It also affects teachers' responses to:

- the outcomes of pupil learning tasks

- their use of novel teaching practices
- their responses to children who are difficult to teach
- their inclusion of children with disabilities
- their level of stress and their satisfaction with the teaching profession.

Measuring teacher self-efficacy however, has been the source of much discussion and confusion (Henson, 2001), particularly as the results are so interwoven with teacher learning strategies and motivational style.

Noticing that there was a need for a specific teacher wellbeing scale that encompassed these other concepts, Rebecca Collie developed such a scale for her PhD thesis (Collie, 2014) and further refined it over the next few years into the Teacher Well Being Scale (TWBS), a well-regarded survey instrument consisting of 16 questions (Collie et al., 2015, p.745). Collie's scale, on which this study is based, drew from the above-mentioned concepts of SDT, self-efficacy theory and motivation theory and proposed three teacher-specific factors of wellbeing:

- ***Organisational wellbeing*** concerns the environment in which teachers work and the relationships they form with their colleagues in school.
- ***Workload wellbeing*** concerns the time available to carry out the marking, teaching and administrative work allocated to them.
- ***Student interaction wellbeing*** covers areas such as student behaviour and motivation, interactions with students and classroom management.

An important characteristic of the TWBS is that it takes a practice-oriented approach to measuring teacher wellbeing; that is, it focuses on the determinants of wellbeing rather than attempting to assess indicators or outcomes of wellbeing (e.g., life satisfaction) directly (Collie et al., 2015, pp.745-746). The practice-oriented approach has been shown to assess wellbeing reliably (Organisational wellbeing $\alpha = 0.84$, Workload wellbeing $\alpha = 0.85$ and Student interaction wellbeing $\alpha = 0.82$) (Collie et al., 2015, p.748) and offers the additional benefit of identifying the factors that might be relevant in trying to improve it. This was a particularly important benefit for the present research, given the highly unusual circumstances in which teachers—as well as the rest of society—found themselves. The TWBS has recently been used in several different studies and circumstances (Fox et al., 2020; Yeo, 2021) to collect self-report data on teachers' wellbeing. It has been proved to be a robust and reliable measure that is easy to administer, yielding data that is straightforward to analyse.

Teacher wellbeing and demographic characteristics

What would we expect to see in our research based on the work of Collie and others? Christian Gloria and colleagues (2013) posited that positive affect (one's ability to face life with a positive outlook and interact positively with others) was positively correlated with resilience and negatively correlated with burnout. They also found that positive affect was more common among more experienced teachers and stress was more common among female teachers. Conversely, Collie et al. (2015) found higher levels of teacher wellbeing for older

and *less* experienced teachers compared with younger and more experienced teachers, and no effect by gender. A further study (van Petegem et al., 2005) looked directly at gender, parental status, job security and years of experience in relation to teacher wellbeing. They found a positive relationship between years of experience and teacher wellbeing, a positive relationship between teacher wellbeing and positive attitudes towards their students, and a negative link between teacher wellbeing and teacher dissatisfaction. They also noted that teachers who had children of their own tended to display higher levels of wellbeing. So, the situation is complex.

Teacher wellbeing and COVID-19

Given the relatively small amount of research into teacher wellbeing in general, it is unsurprising that little has been published about the effect of COVID-19 on teacher wellbeing. A working paper produced by University College London (UCL) (Allen et al., 2020) reported on the increased stress and work-related anxiety experienced by head teachers, who were expected to lead teams in ways that called on access to skills and resources that may not have been readily available. This increased stress was not found to be reflected by classroom teachers, who, although expected to teach in very different ways, did not have the stresses of managing students in the classroom. Allen and her colleagues' findings were that teacher wellbeing, as measured using the Warwick–Edinburgh Mental Wellbeing Scale, had not changed between October 2019, before the pandemic and April 2020, when the UK national lockdown was very well established. Another study by Collie (2021) based on data from Australian schools found (fairly unsurprisingly) that teachers were much more stressed if they were teaching both remotely and in school than if they were teaching remotely only. A recent study by Kim et al. (2021) from York University in the UK, highlighted that the stresses are not spread equally, with primary school head teachers and senior leaders being more stressed than their secondary school counterparts, largely because there are fewer of them in a typical primary school to shoulder the burden.

Method

Nine schools (eight English, one Welsh) were recruited through the Cambridge Centre for Evaluation and Monitoring (CEM)[2] website (www.cem.org) with a view to learning more about teacher and student wellbeing during the second national COVID-19 lockdown. This report concentrates on the experiences and concerns of teachers, which were collected using an online survey. The survey was based on the TWBS instrument (Collie et al., 2015). It also included a single, open question designed to allow teachers freedom to express their concerns and reflect on their wellbeing during lockdown as compared with their perceptions post lockdown.

The TWBS was developed in Canada. In the present study, we made some very

2 Cambridge CEM (Centre for Evaluation and Monitoring) is a leading provider of assessment and monitoring systems including baseline, attitudinal, diagnostic and entrance tests.

small modifications to it, to adapt its language for use in the UK. We then used it to collect data on teachers' perceptions of their wellbeing during two phases of the pandemic. The modified TWBS (Appendix 1) consisted of 32 questions. The first 16 of these related to teacher perceptions during the second national lockdown in January and February 2021, and the second set of 16 questions related to wellbeing at the point the survey was administered in May 2021. The first 16 questions were prefaced by the phrase "During the lockdown in January and February 2021, how did the following aspects of being a teacher affect your wellbeing?". The second 16 questions were prefaced by the phrase "Currently, how do the following aspects of being a teacher affect your wellbeing?". Examples of individual items are "Relationships with students in my classes" and "Student motivation". All TWBS questions were presented on a 7-point Likert scale that ranged from "very negatively" to "very positively".

In addition, prior to the modified TWBS, the survey included questions relating to the teacher's age group, gender, subjects taught, and years of teaching experience, plus the proportion of students in the teacher's school who were in receipt of free school meals. At the end of the survey, the teachers were also asked what single thing would most improve their wellbeing as a teacher.

Teachers and senior leaders were recruited to take part in the survey through an article and associated blog post on the Cambridge CEM website (www.cem.org). The survey was delivered using SmartSurvey (www.smartsurvey.com). Responses to the TWBS items were analysed in R (R Core Team, 2021) and the single open-ended question was analysed using MAXQDA (VERBI Software, 2019).

Results

Characteristics of responding teachers

Fifty-four responses to the survey were received from nine schools, and the characteristics of responding teachers are summarised in Appendix 2. For the purposes of interpreting the results, it is important to note that three schools dominated the survey responses (one academy and two independent schools). To avoid comparing groups of responses from the same school with those from individuals, responses were classified into four school groups of similar size for certain parts of the analysis: Schools A, B and C each formed their own group, and other responses formed an "Other" group. Respondents included more female teachers than male, and comparison with published teacher workforce data suggests that the survey respondents represented a slightly more experienced group of teachers than average. Responses were received from teachers of all ages, and represented Arts, Humanities, STEM and other subjects.

Wellbeing during and after lockdown

As explained previously, the survey was designed to map to the three main constructs of the TWBS: (i) organisational wellbeing; (ii) workload wellbeing; and (iii) student interaction wellbeing. Respondents' scores for these wellbeing factors were calculated as the mean score for all items mapping to that factor, for both

"during lockdown" and "after lockdown" responses.

Figure 1 summarises the score distributions for the three teacher wellbeing factors by time period. The TWBS uses a scale from 1 to 7 where 1 indicates a strongly negative effect on the teacher, 4 is neutral, and 7 indicates a strongly positive impact on the teacher (Collie et al., 2015). Figure 1 shows that the median levels of both organisational wellbeing and student interaction wellbeing among respondents were positive both during and after lockdown. For both factors, reported wellbeing was slightly higher after lockdown. By contrast, reported workload wellbeing was overall slightly negative. The median levels of workload wellbeing were very slightly higher during lockdown than after lockdown.

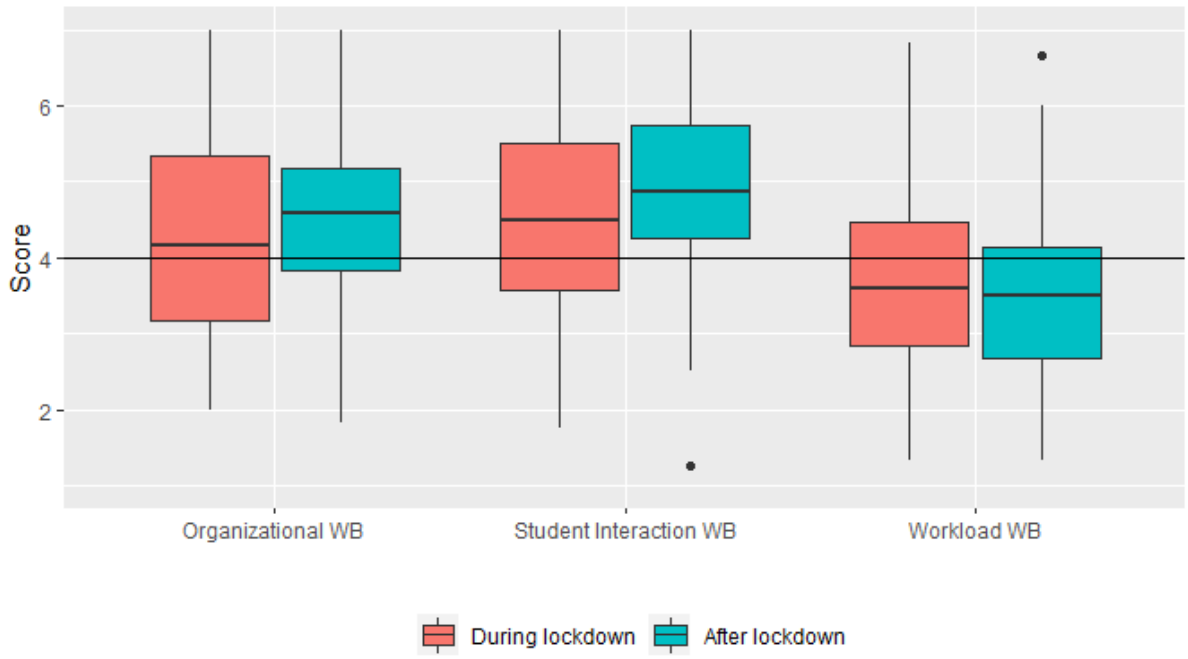


Figure 1: Distributions of wellbeing (WB) scores, by time period.

As Figure 2 shows, there was some striking variation in teachers' reported wellbeing by years of teaching experience. This could reflect changes in the nature of responsibilities and status within the school workforce that are associated with years of teaching experience. In particular, the increasing levels of organisational wellbeing reported by teachers with more years of teaching experience (a phenomenon that has been reported in the literature) could reflect more experienced teachers being more likely than less experienced colleagues to hold leadership positions (e.g., DfE, 2018). It should also be noted that teachers with more years of teaching experience reflect a sample of teachers that is to some extent self-selecting, as many teachers do not persevere in a teaching career for this long. While student interaction wellbeing also tended to increase with years of teaching experience, workload wellbeing appeared to show a non-linear relationship with teaching experience: firstly decreasing, then rising again for the most experienced teachers.

Figure 2: Responses by years of experience.

Figure 3 indicates that there was some variation in teacher wellbeing by gender. As stated previously, some authors (Gloria et al., 2013) found that positive affect was more common among more experienced teachers and stress was more common among female teachers, whereas others (Collie et al., 2015) found higher levels of teacher wellbeing for older and *less* experienced teachers compared with younger and more experienced teachers, and no affect by gender. The results may also reflect the fact that male respondents were skewed towards more years of teaching experience: 38 per cent of male respondents had 21 years or more of experience compared with 27 per cent of female respondents (see Table 4, Appendix 2).

Figure 3: Responses by gender.

Overall changes in participants' wellbeing scores were fairly modest in size (both in absolute terms and viewed in terms of standard deviation). Table 1 shows that the largest difference occurred for student interaction wellbeing, where the mean wellbeing score increased by 0.3 from 4.5 to 4.8.

Table 1: Differences in wellbeing measures.

Measure	Lockdown mean	Lockdown SD	After lockdown mean	After lockdown SD	Change in mean
Organisational WB	4.3	1.4	4.5	1.4	0.2
Student Interaction WB	4.5	1.2	4.8	1.3	0.3
Workload WB	3.6	1.2	3.5	1.1	-0.1

Statistical modelling[3] confirmed that the changes in mean of 0.2 and 0.3 shown in Table 1 for Organisational wellbeing and Student Interaction wellbeing were statistically significantly different from zero, and also estimated a statistically significant increase of 0.4 in Organisational wellbeing score per level of teaching experience (corresponding to an additional five years of teaching experience—see left hand panel of Figure 2). As noted previously, years of teaching experience may serve as a proxy for seniority and the nature of participants' role within a school. It may also reflect a degree of self-selection among those with more years of teaching experience, if teachers with lower teacher wellbeing leave the

3 Details not shown to save space. Available from the authors on request.

profession at higher rates earlier in their careers.

For workload wellbeing, a slightly different model structure was necessary. The results showed that there was no statistically significant effect of time period (during vs. after lockdown) on workload wellbeing, once other factors were accounted for.

Relationships between teacher wellbeing factors

There were moderate correlations between the different wellbeing measures, both during lockdown and after lockdown (Table 2). The original TWBS (Collie et al., 2015) reported a correlation of 0.47 between workload and organisational wellbeing; 0.57 between workload and student interaction wellbeing; and 0.45 between organisational and student interaction wellbeing.

The correlations found in the survey results were broadly in line with these, with two areas of slight difference: firstly, the correlations between workload wellbeing and student interaction wellbeing (0.34 during lockdown, and 0.39 after lockdown) were lower than the value reported by Collie et al. (2015), and secondly, after lockdown, the correlations of both student interaction and workload wellbeing with organisational wellbeing were higher than the values reported by Collie et al. (2015).

In terms of comparisons between the time periods, the organisational wellbeing measure from lockdown was correlated highly with the organisational wellbeing measure post-lockdown; the correlations of workload and student interaction measures between the two time points were lower.

Table 2: Pearson correlations between wellbeing measures.

		During lockdown			After lockdown		
		Workload WB	**Organisational WB**	**Student**	**Workload WB**	**Organisational WB**	**Student**
During lockdown	**Workload WB**	1.00	0.41	0.34	0.65	0.41	0.15
			1.00	0.58	0.51	0.89	0.56
	Student Interaction WB			1.00	0.34	0.51	0.64
After lockdown	**Workload WB**				1.00	0.62	0.39
						1.00	0.65
	Student Interaction WB						1.00

Did lockdown change which parts of the workload affect teacher wellbeing?

Each of the three main wellbeing measures was based on a series of questions linked to aspects of teaching work that made up the measure. In some cases, an aggregated figure (such as an overall wellbeing measure) can mask subtler changes at the question level. In order to investigate this, the results for the individual questions were compared.

Comparing responses to these questions during and post lockdown (Figure 4) showed that generally all aspects of teaching work contributing to the workload wellbeing factor were considered to have marginally negative effects on wellbeing, but that these tended to be smaller during lockdown.

Figure 4: Item means for workload wellbeing.

Figure 5 shows that aspects of work contributing to the organisational wellbeing factor were generally rated neutral or slightly positive, but improved a little post lockdown. The largest difference was for relationships with administrators which was perceived to have improved post lockdown.

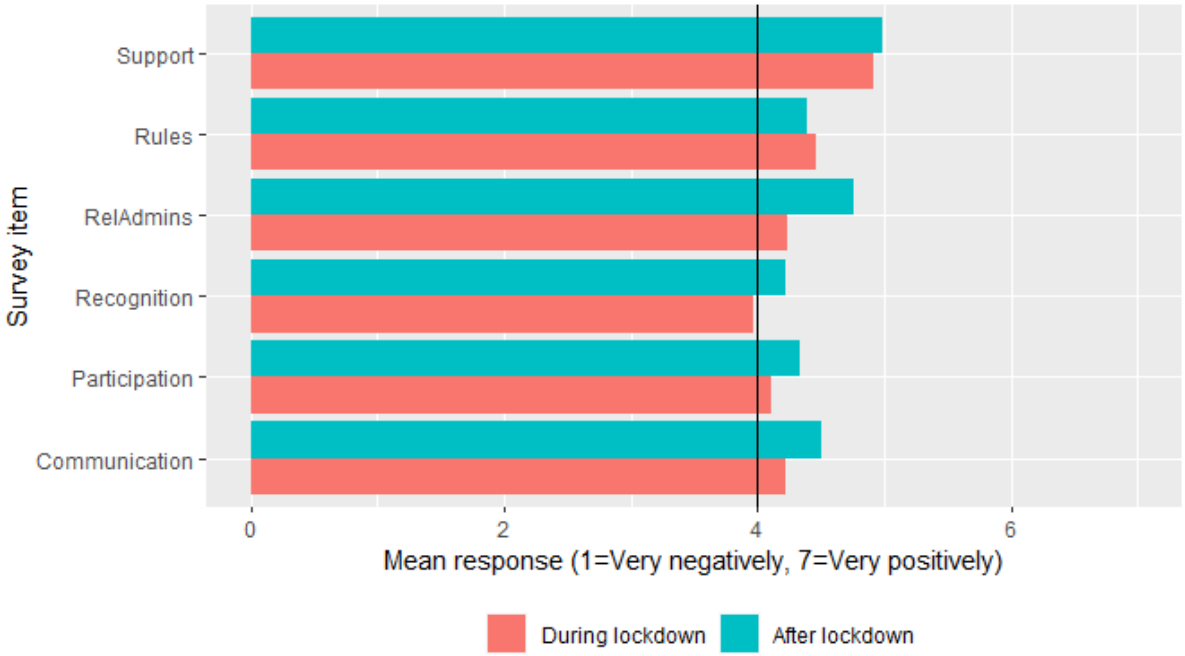


Figure 5: Item means for organisational wellbeing.

The responses to the student interaction wellbeing questions displayed a mixed message (Figure 6). Teachers felt that students' motivation during lockdown affected teacher wellbeing far more negatively than post lockdown, when its impact on teacher wellbeing was overall positive. Conversely, student behaviour was judged to affect their teachers' wellbeing more positively during lockdown—although post lockdown, it was still perceived as a slightly positive influence. Student relationships with teachers also tended to positively influence teachers' perceived wellbeing both during and after lockdown, but teachers reported a more strongly positive impact on wellbeing after lockdown.

Figure 6: Item means for student interaction wellbeing.

In terms of individual factors affecting teacher wellbeing, the largest changes between the two time periods were seen in aspects directly relating to human interactions ("Relationships with administrators at my school" and "Relationships with students in my classes") and student motivation.

Views on improving wellbeing

The survey ended with the following open-ended question, which was analysed qualitatively: "Going forward, what single thing would most improve your wellbeing as a teacher?"

The responses were entered into the qualitative software package MAXQDA and a conventional content analysis was conducted. With this grounded theory (Glaser & Strauss, 1967) approach, themes and thereby coding categories are derived directly from the text data, and each response is then coded using one or more of the codes.

A number of themes emerged during the analysis and these were further refined, by both merging some themes and creating new ones. A list of themes emerged and these were arranged into groups.

A simple Venn diagram (Figure 7) shows two main areas of concern, one associated with school management and issues involved with running a school, the other with government action and the public recognition of teachers. There is an overlap between these two in terms of the time taken to do the job of a teacher, the associated administrative work and this year, the extra work required to produce teacher grades for GCSE and A Level. Outside of these two main areas is "Anything Else", a collection of diverse and eclectic comments that could not be neatly categorised.

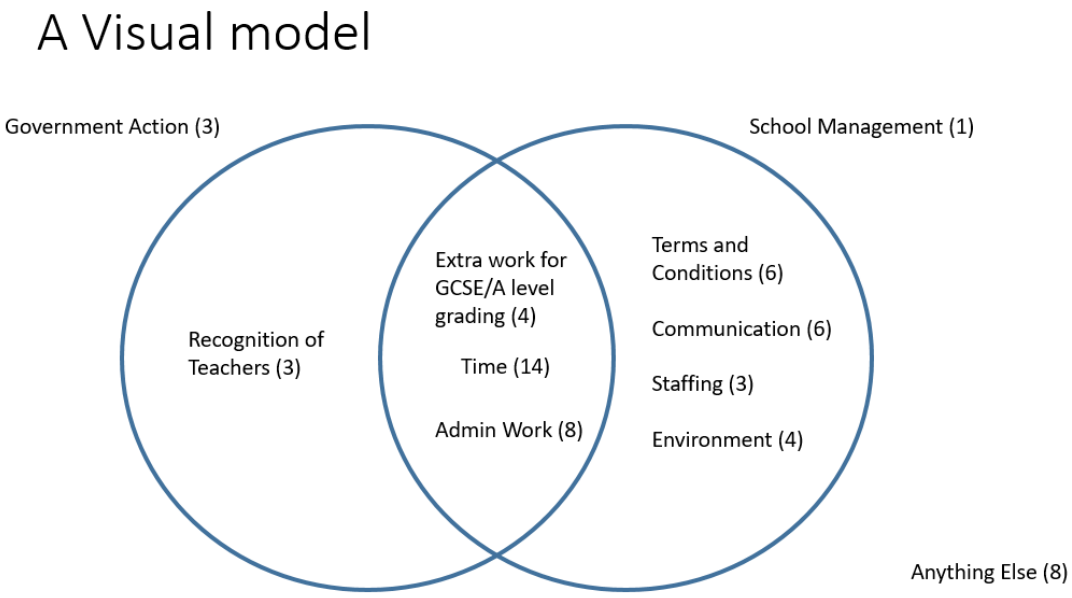


Figure 7: Venn diagram showing the relationship between teachers' areas of concern (the number of times each theme arose is given in brackets).

The advantage of asking an open-ended question is that it allows those being surveyed to answer the question based on the things that are directly affecting them. The question asked in this survey was concerned with the single thing that would improve their wellbeing as a teacher. Since the rest of the survey concerned experiences between teaching during the lockdown period and subsequently when back at school, we expected that the travails of teaching remotely and the issues with planning lessons in ways that had not been used before would be the main focus of the responses. That however, was not the case. It appeared that teachers took lockdown in their stride, and the things that always occupy them—time, administrative work, and general school life—had a far greater effect on their wellbeing. The extracts below give a flavour of the types of things that concerned the teachers in our survey with regard to wellbeing.

Time

Of the various themes that emerged, the one mentioned most frequently was time.

> "Time to deal with emails and communication. Time to action these. More time to really develop personal support for individual students. Time to work with my colleagues to develop teaching and learning."

> "Giving teachers time to teach with support and without interference would help every teacher's wellbeing."

> "Less teaching time and more preparation and marking time."

> "Not having to work until midnight every day to complete the majority of the work expected of me."

Administrative work

Another key area that teachers commented on was the amount of administrative work they were expected to do.

> "Less admin and time to actually just teach students and them to enjoy the subject."
>
> "Less meetings and admin tasks".
>
> "Much less administrative work".

Recognition

Teachers also felt that their profession did not have appropriate recognition.

> "There is very little recognition of the job that teachers do from government or the DfE. Paradoxically schools are increasingly passed on initiatives and requirements in loco parentis all of which take resources that are not provided by government to independent schools."
>
> "Positive media coverage of the profession - it is really wearing to be berated so frequently by politicians."
>
> "Appreciation."

Extra work for GCSE / A Level grading

Although there was no mention of the effects of lockdown, there was a groundswell of concern about having to spend time providing grades for GCSE and A Levels. In the UK, national examinations were cancelled and teacher grades were used instead. Each school was required to submit a set of grades for their students to the awarding bodies. This was perceived as work that teachers would not receive payment for.

Typical comments were:

> "This year, not doing the job of a GCSE examination marker who gets paid to do this. I am doing their job about 4 times over trying to get the evidence together."
>
> "Not having to do the work of the exam boards."
>
> "Recognition for the fact that we are now marking all of the assessments used for Year 11 and Year 13, in our own time, for no extra pay instead of the exam boards, who the schools pay to do this. There is no extra recognition at all for this."

Communication

A number of responses concerned communication within schools.

> "Better communication from the College."
>
> "Better communication between senior leaders and staff".
>
> "Co-ordination between different senior levels about what is due when.

> At the moment, it seems everything gets chucked at us with VERY narrow deadlines from many sides."
>
> "Better communication throughout school and longer deadlines."

Terms and conditions

Some teachers told us that their wellbeing would be improved most by their school addressing their terms and conditions of service. It seems that for some, having worked from home during the lockdown successfully, meant that they could do administrative work from home productively under normal circumstances too.

> "Ability to work from home during PPA [planning, preparation and assessment] time if appropriate."
>
> "A work from home day every now and then on a lighter day perhaps to catch up on admin related activities. Particularly since teaching can still be to a good standard."
>
> "Better pay."
>
> "Salary."
>
> "Reduced workload."

The pandemic

Interestingly, only two teachers mentioned the COVID-19 pandemic.

> "Having more staff. We are continually understaffed for no discernible reason and this means we are all stretched to miss breaks, struggle to mark and set work, teach our own classes to more than an adequate level and miss planning lessons. Although COVID restrictions have been lifted, it is being used as an excuse not to bring in supply teachers. However, this never happened before COVID and therefore is not valid."
>
> "An end to all virus related restrictions. Normality restored."

Anything else

Some teachers commented on things that could improve their wellbeing that could not be categorised within the structure proposed. Although not directly related, they do provide a window onto the issues that teachers face.

> "Improved support for the more demanding pupils' behaviour / SEN [special educational needs] requirements, especially where no LSA [learning support assistant] has been allocated."
>
> "Allowing Business A Level students to take their exams online and to stop using paper assessments."
>
> "Continuous assessment of key aspects would make teaching less onerous and stressful for me to complete the syllabus on time."

Were comments linked with any particular group?

It is possible that all the concerns listed above were from a particular age group of teacher. The issue of time was mentioned by most age groups, although not by the 60+ group. Every age group included someone who considered administrative work to be a barrier to their wellbeing. Terms and conditions were mentioned more by the three youngest age groups and communication by the 60+ group more than any other.

Did one particular school have a particular issue that all the teachers reported back as being a problem? No single issue compromised teachers' wellbeing within any particular school. For the schools with many responses (schools A, B, and C), the comments cover the wide range of themes that developed from this analysis.

Although there was a wide spread of respondents with differing levels of experience, the small sample size precluded any clear conclusions relating level of experience to particular themes. Additionally, the question asked "what *single* [emphasis added] thing would most improve your wellbeing as a teacher?" and the majority of the responses rightly gave a single answer as requested. With hindsight, it might have been better to ask for two or three things, perhaps with a ranking. This might have provided a broader view of the issues affecting teacher wellbeing.

Conclusions

This analysis concerned itself with teachers' perceived wellbeing and the differences between teaching during lockdown with the situation post lockdown. It was a relatively small survey, and the majority of the responses came from three schools (two independent schools and one academy). As such, it cannot be said to be representative of the bigger picture, nevertheless, it is a very interesting reflection on the situation by a small number of respondents and provides a useful way of opening up dialogue about and future research into the issue. The experiences of teachers during lockdown have raised issues such as availability, skills and understanding of technology, teaching remotely and changes in working patterns. None of these are directly addressed by this survey. Our survey was particularly focused on teacher wellbeing, as that area had not been investigated widely and we felt that any profound changes to the working conditions and methods that teachers were being expected to use might manifest themselves in changes to their perceived wellbeing.

The survey addressed three main areas: (i) organisational wellbeing; (ii) workload wellbeing; and (iii) student interaction wellbeing, and of the three, the effects of the organisation and teacher workload on wellbeing were the most prominent in the comments that teachers made. These comments clearly linked to the concept of agency, or ability to make one's own decisions, which was described earlier in this article as being associated with higher levels of wellbeing. In terms of the impact of lockdown on teacher wellbeing, the picture that emerges is that, for the teachers surveyed, there was not a large change in any area that they felt affected their wellbeing. It might be hypothesised that teaching remotely would

be more stressful and would therefore affect wellbeing more negatively than teaching in the classroom. That, however, did not seem to be the case: the results showed that teachers' organisational wellbeing and student interaction wellbeing were only slightly lower during lockdown than after lockdown, and there was no statistically significant change in workload wellbeing. It appeared that the impact of student behaviour on teacher wellbeing was more positive during lockdown, possibly because students were not with their peers, and were instead in the presence or vicinity of their parents / carers. Conversely, the impact of student motivation on wellbeing during lockdown was negative—suggesting that teachers perceived student motivation to be lower than usual—and the impact of teachers' relationships with students was less positive than after lockdown, though still overall a positive impact on wellbeing. These findings are broadly in line with those of Allen et al. (2020) who found no difference in teachers' psychological wellbeing before and during the first national lockdown, but did find students' perceived motivation to be lower. In a Norwegian study by Bubb and Jones (2020), teachers found that classroom management was slightly easier during lockdown, but our findings showed no change in the impact of classroom management on teacher wellbeing.

From the qualitative data on how teachers felt wellbeing could be improved, we found that the issues teachers perceived to most affect their wellbeing were the issues that affected teacher wellbeing regardless of lockdown. Teachers were concerned about the time available to do their jobs, closely followed by the amount of administrative work they were expected to do. Some found that these issues were exacerbated by decisions made by the school leaders. What was interesting is that some of the teachers that had taught for the longest time were among those finding that time pressure and administrative work was affecting their wellbeing. It might be assumed that among these more experienced teachers, many would be school leaders themselves and therefore be able to make changes within the school environment to address these issues. However, it also corroborates the findings of Allen et al. (2020), cited previously, who also found greater stress among senior and head teachers.

To conclude, despite the challenges posed by teaching through the pandemic, teachers' wellbeing during lockdown was measured to be only slightly lower than their wellbeing post lockdown. The issues that teachers reported as strongly affecting teacher wellbeing were those present more generally, such as workload. Ensuring wellbeing needs are met in 'normal' times may, therefore, help to increase resilience when novel challenges arise.

Acknowledgements

We would like to acknowledge Matthew Carroll's input into the design of the study. Hannah North, Antje Diestelhorst and Kayleigh Lauder provided administrative support. Mark Frazer and John Little reviewed a draft of the article.

Ethical considerations

The research was conducted in full accordance with the principles stated in the research team's institutional *Research Ethics Guidance* document. This included obtaining informed consent from participants, protecting their privacy, and obtaining permission for use of anonymised quotes.

References

Allen, R., Jerrim, J., & Sims, S. (2020). *How did the early stages of the COVID-19 pandemic affect teacher wellbeing?* Centre for Education Policy and Equalising Opportunities (CEPEO) Working Paper, 20–15. https://repec-cepeo.ucl.ac.uk/cepeow/cepeowp20-15.pdf

Bandura, A. (1977). Self-efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bubb, S., & Jones, M. A. (2020). Learning from the COVID-19 home-schooling experience: Listening to pupils, parents/carers and teachers. *Improving Schools, 23*(3), 209–222. https://www.doi.org/10.1177/1365480220958797

Children's Commissioner. (2020). *School return: Covid-19 and school attendance* [Briefing]. The Children's Commissioner's Office. https://www.childrenscommissioner.gov.uk/report/school-attendance-since-september/

Coleman, V. (2021). Has Covid-19 highlighted a digital divide in UK education? *Cambridge Assessment Website Blog*. https://www.cambridgeassessment.org.uk/blogs/has-covid-19-highlighted-a-digital-divide-in-uk-education/

Collie, R. J. (2014). *Understanding teacher well-being and motivation: measurement, theory, and change over time* [Thesis submitted for the degree of Doctor of Philosophy]. University of British Columbia. https://doi.org/10.14288/1.0165878

Collie, R. J. (2021). COVID-19 and teachers' somatic burden, stress, and emotional exhaustion: examining the role of principal leadership and workplace buoyancy. *AERA Open, 7*, https://doi.org/10.1177/23328584209861872

Collie, R. J., Shapka, J. D., Perry, N. E., & Martin, A. J. (2015). Teacher Well-Being. *Journal of Psychoeducational Assessment, 33*(8), 744–756. https://doi.org/10.1177/0734282915587990

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour.* Plenum Press.

DfE. (2018). *School leadership in England 2010 to 2016: characteristics and trends* (DFE-RR812). Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/725118/Leadership_Analysis_2018.pdf

DfE. (2021), *Understanding progress in the 2020/21 academic year.* Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/994364/Understanding_Progress_in_the_2020_21_Academic_Year_Initial_Report_3_.pdf

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist,* 55(1), 34.

Dörnyei, Z. (2001). *Teaching and Researching Motivation.* Pearson Education Ltd.

Fox, H. B., Tuckwiller, E. D., Kutscher, E. L., & Walter, H. L. (2020). What Makes Teachers Well? A Mixed-Methods Study of Special Education Teacher Well-Being. *Journal of Interdisciplinary Studies in Education, 9*(2), 223–248. https://files.eric.ed.gov/fulltext/EJ1294633.pdf

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory.* Aldine.

Gloria, C., Faulk, K., & Steinhardt, M. (2013). Positive affectivity predicts successful and unsuccessful adaptation to stress. *Motivation and Emotion, 37*, 185–193. https://doi.org/10.1007/s11031-012-9291-8

Gordon, C., & Debus, R. (2002). Developing Deep Learning Approaches and Personal Teaching Efficacy within a Preservice Teacher Education Context. *British Journal of Educational Psychology, 72*, 483–511.

Henson, R. (2001). *Teacher Self-Efficacy: Substantive Implications and Measurement Dilemmas*. Invited keynote address given at the annual meeting of the Educational Research Exchange, January 26, 2001, Texas A&M University, College Station, Texas. https://files.eric.ed.gov/fulltext/ED452208.pdf

Kim, L., Oxley, L., & Asbury, K. (2021). *"My brain feels like a browser with 100 tabs open": A longitudinal study of teachers' mental health and wellbeing during the COVID-19 pandemic in 2020.* https://psyarxiv.com/cjpdx/download?format=pdf

Kuhfeld, M, & Tarasawa, B. (2020). *The COVID-19 slide: What summer learning loss can tell us about the potential impact of school closures on student academic achievement.* NWEA. https://files.eric.ed.gov/fulltext/ED609141.pdf

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher Efficacy: Capturing an Elusive Construct. *Teaching and Teacher Education, 17,* 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1

Van Petegem, K., Creemers, B. P., Rossel, Y., & Aelterman, A. (2005). Relationships between teacher characteristics, interpersonal teacher behaviour and teacher wellbeing. *The Journal of Classroom Interaction, 40*(2), 34–43. https://www.jstor.

org/stable/23870662

VERBI Software. (2019). *MAXQDA* 2020 [computer software]. VERBI Software. Available from maxqda.com

Weidmann, B., Allen, R., Bibby, D., Coe, R., James, L., Plaister, N, & Thomson, D. (2021). *Covid-19 disruptions: Attainment gaps and primary school responses.* Education Endowment Foundation. https://dera.ioe.ac.uk/37913/1/Covid-19_disruptions_attainment_gaps_and_primary_school_responses_-_May_2021.pdf

Williamson, J., Suto, I., Little, J., Jellis, C., & Carroll, M. (2021). Learning during lockdown: How socially interactive were secondary school students in England? *Research Matters: A Cambridge Assessment Publication, 32*, 22–47.

Yeo, B. (2021). *Caring for teachers: Exploring pre-service teacher well-being, self-efficacy, and vicarious trauma.* https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=10321&context=etd

Appendix 1: Questions in the survey of teachers' wellbeing

School Name

Please indicate your age group [20–29, 30–39, 40–49, 50–59, 60+, Prefer not to say]

Please state your gender [Male, Female, Other, Prefer not to say]

What subject(s) do you currently teach?

For how many years have you been teaching?
[0–5, 6–10, 11–15, 16–20, 21 years or more]

Approximately what proportion of students in your school are eligible for free school meals?
[0–20%, 20–40%, 40–60%, 60–80%, 80–100%, Unsure, Not applicable]

During the lockdown in January and February 2021, how did the following aspects of being a teacher affect your wellbeing? Wellbeing refers to open, engaged and healthy functioning as a teacher.

Currently, how do the following aspects of being a teacher affect your wellbeing? Wellbeing refers to open, engaged and healthy functioning as a teacher.

1. Marking work (Workload WB)[4]

2. Student behaviour (Student interaction WB)

3. Fitting everything into the allocated time (Workload WB)

4. Support offered by school leadership (Organisational WB)

5. Relationships with students in my classes (Student interaction WB)

6. Administrative work related to teaching (Workload WB)

7. Recognition for my teaching (Organisational WB)

8. Student motivation (Student interaction WB)

9. Teaching work that I completed outside of school hours (Workload WB)

10. School rules and procedures that were in place (Organisational WB)

11. Working to finish my teaching preparation tasks (Workload WB)

12. Communication between staff members of the school (Organisational WB)

13. Relationships with administrators at my school (Organisational WB)

14. Class management (Student interaction WB)

15. Working late to attend meetings and activities (Workload WB)

16. Participation in school-level decision-making (Organisational WB)

4 We have included the wellbeing factor that corresponds to each item for readers' information. It should be noted that this information was not visible to respondents during the survey.

Appendix 2: Characteristics of responding teachers

The first question of the survey asked for the name of the respondent's school. Fifty-four responses to the survey were received from nine schools (Table 3). As may be clearly seen, three schools dominated the survey responses (one academy and two independent schools). To avoid comparing groups of responses from the same school with those from individuals, responses were classified into four school groups of similar size for certain parts of the analysis: Schools A, B and C each formed their own group, and other responses formed an "Other" group.

Table 3: Description of respondents' schools.

School	Type	Age range of students in the school	% Free School Meals (self reported)	
A	Independent	2 to 18	0–20%	16
B	Academy	11 to 18	0–20%	15
C	Independent	11 to 16	0–20%	11
Other	FE College	16+	Unsure	3
Other	Independent	3 to 16	0–20%	3
Other	Academy	11 to 18	40–60%	3
Other	Independent	2 to 18	0–20%	1
Other	State	11 to 18	20–40%	1
Other	Sixth Form	16+	0–20%	1

Table 4: Respondents' age, gender, subject area and teaching experience (total N=54)

		Frequency
	20–29	4
	30–39	17
Age Group	40–49	18
	50–59	9
	60+	6
Gender	Female	41
	Male	13
	Arts	7
Subject Area	Humanities	20
	STEM	14
	Other	13
	0–5	9 (7 / 2)
Years' Teaching Experience	6–10	8 (6 / 2)
(No. of females / No. of males)	11–15	11 (9 / 2)
	16–20	10 (8 / 2)
	21+	16 (11 / 5)

What do we mean by question paper error? An analysis of criteria and working definitions

Nicky Rushton, Sylvia Vitello (Research Division) and Irenka Suto (Cambridge CEM)

Introduction

Every year, exam boards produce thousands of question papers for GCSE and A Levels. The majority of these question papers are error free, but a small number of errors are found. In 2019, the last year in which there was a summer series of exams, there were 56 errors in 6,304 papers, suggesting that approximately 98 per cent of papers were free from errors (Ofqual, 2019). One of the reasons that the rate of errors is so low is that papers go through a series of checks before and after they are printed that are intended to eliminate errors. These checks may identify multiple problems within a question paper, but it is not always clear whether these problems constitute an error. Some problems, such as an incorrect number in a mathematics question that makes it unsolvable, or a multiple-choice question with no correct response options, are undoubtedly errors. Other problems, such as a missing "Oxford comma" in a question, or a lack of strict adherence to some of the more obscure rules of grammar, can fall into a grey area. Occasionally, there may not technically be an error in what the candidate sees, but the question paper could function sub-optimally. Examples of this may include items that are awkwardly worded but still answerable, and items with inconvenient layouts that require candidates to flip back and forth between pages.

It is important to be able to define what an error in a question paper is so that there is a common understanding amongst the people taking part in the question paper production process. Otherwise, people's own conceptions could impact upon the way in which they write or check question papers. These personal conceptions could also impact upon the errors that are recorded in the error logs that are used to inform systems improvement and minimise the chance of errors appearing in future papers.

In this article, we will discuss the ways in which error has been conceptualised in the existing literature, considering categorisations that are used in other industries and those used for errors in assessments. We will then report on the

findings from our study to investigate the personal conceptions of error that are used within Cambridge University Press & Assessment, and use these to propose a way of defining errors in assessment materials.

Some industries, such as aviation, medicine and nuclear, have an extensive literature concerning error, and therefore a common understanding of error. In assessment, the literature is less well established. This led Suto and Ireland (2021) to draw upon the work of James Reason, an eminent researcher working on error, to describe error in the assessment context. Suto and Ireland state that system-level failure (that is, imperfect working conditions during the question paper construction process) can lead to human failures (that is, failures among assessment authors and checkers). These human failures can introduce defects into draft assessment instruments (e.g., failing to spell a word correctly), and can prevent those errors from being detected during subsequent stages of the construction process (e.g., failing to detect a spelling error). The consequence of these human failures is that final versions of assessments can contain errors.

According to Suto et al. (2021), this model uses the term *error* in two distinct ways. First, it can relate to a particular type of human failure, an action or inaction, also known as a "human error". Secondly, it can describe the consequences of human failure, resulting from either an action or an inaction; this is known as a "question paper error". Assessment is unusual in using the term *error* in both ways. Most of the error literature focuses on the first of these, the "human error" (e.g., Schubert et al., 2012; Reason, 2013), and does not consider the events or products that result from them as errors per se. For example, in a chocolate factory, a possible consequence of human failure would be described as "chocolate that is too milky" rather than a "chocolate error".

Turning to the assessment industry, candidate impact has been the main focus for classifications of question paper errors and it forms the basis for the categorisations of question paper errors used by the Office of Qualifications and Examinations Regulation (Ofqual) (see Table 1). Similar distinctions can be found in the nuclear and aviation industries. Both industries use the seriousness of impact as one of the ways of distinguishing between an accident and an incident (see European Union, 2010; and International Atomic Energy Agency, n.d.-a and -b).

Table 1: Ofqual's classification of question paper errors

Category 1	Category 2	Category 3
Errors which could or do make it impossible for learners to generate a meaningful response to a question /task.	Errors which could or do cause unintentional difficulties for learners to generate a meaningful response to a question / task.	Errors which will not affect a learner's ability to generate a meaningful response to a question / task.

Although candidate impact is a critical part of defining errors, assessment organisations, like many other industries, also collect data on the manifestation of errors (i.e., what the errors look like). Suto et al. (2021) used this error data to develop a taxonomy of manifested error types. During its development, they

drew upon six different constructs that are critical to validity and reliability in educational assessment:

- accuracy
- clarity (including ease and uniformity of interpretation);
- consistency
- alignment with design intentions (as specified in a syllabus, blueprint or other "source of truth")
- offensiveness (including cultural sensitivity);
- equality of difficulty amongst candidate sub-populations.

These constructs are all aspects or qualities of assessment materials that can be imperfect, and which could provide the basis for describing question paper errors. Although instructions for question paper checks may not use these exact terms, the checkers will be making subjective professional judgements about these constructs while they check for errors. These judgements are rarely "all or nothing" and instead may be placed upon a scale. For example, judgements of clarity may range from being extremely clear to extremely unclear.

Judgements of constructs such as accuracy and clarity are made routinely and may be explicitly targeted by particular question paper checks (see Suto et al., 2021). Judgements may also be made implicitly and unconsciously by colleagues who create and check assessment materials. These implicit judgements could lead to individuals creating their own criteria or thresholds that must be met, in order to decide whether an error (or other problem) in a paper is serious enough to be corrected. However, it is likely that communities of practice evolve (see Wenger, 1998, for discussion), with shared understandings of what constitutes an error and what does not. Different communities may adjust these criteria to meet their own needs, thus creating stricter and looser definitions of error. The criteria may also depend on the context and intended uses; for example, different criteria may be used for identifying errors within the checking processes and for identifying errors with the purpose of logging them.

In this introduction, we have described some of the terms that are used to describe error, and why it is important to have a common understanding of these terms. In the rest of this article, we describe our research investigating whether our colleagues think that a problem in an assessment material constitutes an error. This research was conducted in the context of a larger interview study exploring colleagues' experience of question paper errors and the culture surrounding the discovery of such errors.

Method

We carried out semi-structured interviews with 36 colleagues from across Cambridge University Press & Assessment's assessment production teams. Ten of the participants were senior managers who were involved in system-level decisions about question paper production. Thirteen participants were question

paper managers, who had responsibility for the question papers in a particular subject and / or qualification and oversaw the day-to-day question paper production processes associated with these papers. Finally, 13 participants were checkers who carried out one (or more) of the checks towards the end of the question paper production process. Although the participants were interviewed about a particular role, many participants had worked in a variety of roles and were able to provide insight beyond the role that they were recruited for.

The question paper managers and checkers were mathematics, science, history, or English as an additional language specialists, working on I/GCSEs, AS/A Levels, IELTS, Cambridge English main qualifications or BMAT. These subjects and qualifications were chosen because they represented a range of question paper types, with different question formats and different numbers of errors.

Each participant was interviewed for approximately an hour, either in person or over the telephone. We devoted a section of the interview schedule to investigating participants' definitions of errors. All the participants were asked about what they considered to be an error, and the difference between errors, issues and initial revisions. The senior managers and question paper managers were also asked to identify the stage of the question paper production process they thought a problem should be classified as an error, and to explain what they considered to be a "near miss" for errors.

We used transcriptions of the interviews to identify data relating to definitions of error, issues and near misses. The examples of issues and errors were matched to both Suto et al.'s (2021) taxonomy of manifested error types and the six validity and reliability constructs that Suto et al. used to construct their taxonomy of errors.

Results

We analysed the data according to three aspects of error. The first was the distinction between errors and issues, specifically which types of assessment problems were viewed as "errors" and which were viewed as "issues". The second was the stage of the question paper construction process when problems were considered to be errors. The third aspect was participants' use of the term *near miss*. We chose the first two aspects as we thought they would help us to understand how errors should be defined. Near miss is a term used in other industries such as error and medicine, which include near misses in their error logs. We thought that it was important to understand how it was defined in our industry so that we could ensure that near misses are included in our own error logs.

Problems as errors and / or issues

All the participants gave examples of problems with question papers that they considered to be errors and problems they considered as issues, and many described ways in which the two terms differed. Their responses seemed to centre on the impact of the error on candidates and / or its manifestation.

Candidate impact

When describing errors, several of the participants focused upon candidate impact; they either referred to the Ofqual classifications of error, which distinguishes different types of errors based on candidate impact, or described effects of the error upon the candidates.

> "Ofqual have got three broad categories of error and we've sort of adopted that." (Senior manager)

> "An error is something that would cause confusion to candidates, whatever causes it." (Senior manager)

Some participants described the impact of errors on candidates in more detail. For these participants, problems were errors if they prevented candidates from answering a question and / or could confuse them. One of the participants qualified this by stating that a problem would only be an error if it actually appeared in front of candidates or if it led to an erratum notice being issued.

> Respondent: "I think we should only ever use the word error for something that hits the candidate's desk in a question paper that would cause them difficulty in answering the question. I would say it's not an error if we catch it at any point before the candidate sees the paper."
>
> Interviewer: "Would you consider something that needs an erratum note as [an error]?"
>
> Respondent: "Yes. Because the candidate sees it." (Senior manager)

Manifestation

While candidate impact was clearly an important part of some participants' definitions of error, it was more common for participants to focus on the manifestation of the error when asked what an error was. As we stated in the method section, we mapped the manifestations onto both Suto et al.'s (2021) taxonomy of manifested error types and the six validity and reliability constructs that were used to design it. Although we were able to use the taxonomy's categories to consider the distinctions between errors and issues, there were too many categories for them to be useful when defining question paper errors, so the validity and reliability constructs provided a better organising structure.

In the remainder of this section of the results, we will use the six validity and reliability constructs to examine in more detail the examples of errors and issues that participants provided, and to consider whether there were particular factors that affected whether problems were described as errors rather than issues.

Accuracy

The two most commonly provided examples of errors associated with the accuracy construct were spelling, punctuation and grammar errors (SPAG) and factual inaccuracies. SPAG errors were mentioned by almost all of the participants at some point during their interview, generally without any further explanation. Where examples were given, they generally concerned omitted punctuation or the wrong type of punctuation being used.

> "A question mark that should be a full stop, or a full stop that should be a question mark. Some of those errors, again, they might not actually mislead a candidate, but they just look so awful that we might choose to change them." (Senior manager)

Participants also gave examples of SPAG problems that they thought were issues rather than errors. These included commas, particularly where they were considered to be negotiable, and a question mark used instead of a full stop. Participants stated that punctuation problems such as these were not errors because they had little or no effect on candidates, although in one case they still reported it to Ofqual.

> "Error ... means this is something that somebody's done wrong, and, therefore, a comma that I think should go in isn't an error...." (Checker)

Two participants talked about variations in spellings, such as place names or old (rather than modern) spellings of words, as an example of an issue. Neither participant considered their example to be a spelling error because the word was spelt correctly for its context; however, they did identify it as something that should potentially be changed.

> "If it was SPAG it would be an error. But again, on the history papers, they're primary sources. So what I would consider in the 21st century to be a SPAG error, may be, in the 17th century, if they're trying to use 17th century forms of writing rather than modernising them all—they sometimes do, sometimes don't—it's difficult sometimes to make definitive ... I'd raise it as an issue, certainly, though." (Checker)

Factual inaccuracies were mentioned by nearly two-thirds of the participants. Unsurprisingly, all the participants thought of these as errors rather than as issues. In history papers, examples included incorrect dates, events, names and titles, and using sources where the original version contained incorrect information. For science papers, incorrect units were often mentioned, as well as equations containing the wrong substance or being wrong in another way. There were also issues with the science, either with oversimplification, or with the science being "flawed".

> "There are other categories of errors. So, for instance, where a person writing the question might not fully understand the subject or be basing their knowledge on an over-simplification, which turns out not to be in line with more widely accepted views of the world." (Senior manager)

The final example of an accuracy error was using an incorrect word that was similar in meaning and spelling to the intended word (e.g., alkane for alkene, or nucleus for nuclide). This could be a simple spelling error, or it could be a factual inaccuracy caused by a conceptual misunderstanding.

> "They might see something that's high impact in terms of a candidate's ability to answer a question—alkane where it should say alkene or a query like that." (Senior manager)

Although both SPAG problems and factual accuracy problems could impact

upon candidates, arguably factual inaccuracies are potentially more serious as they could prevent candidates from answering a question or leave them with an incorrect understanding of the subject. The impact on candidates appeared to be an important consideration when participants decided whether problems associated with the accuracy construct were an error or an issue.

Clarity

There were many examples of problems that concerned the clarity of the assessment materials. Approximately one-third of the participants mentioned ambiguous wording that affected the readability of the question as an error. Their examples included the text not making sense, the wording being unclear or inaccessible, particular candidates struggling to understand the question, or issues for English as a second language candidates.

> "Questions where the wording is perhaps a little bit ambiguous but you can still produce a reasonable response which can be adjusted for through the mark scheme." (Senior manager)

Other participants provided examples of problematic wording that they considered to be an issue (as opposed to an error). Some of these appeared to be similar to the examples that other participants considered to be errors. However, the issues examples were about changes to wording that would improve the question, perhaps because the wording was awkward or unnecessarily complex, rather than changes that were needed because the question was incomprehensible or ambiguous.

> "Where an item is maybe not as clear as it could be, problems with the wording, level difficulties. I wouldn't call those errors but issues with how an item has been written." (Checker)

Another type of error that related to the clarity construct was missing content, although few participants mentioned this. Their examples included missing content in questions, missing information in equations, and answer lines where units were incorrectly omitted.

> "To me, a real error that's going to make a real difference, is if you have got … if you've got an equation, and there's something essential that's missing from the equation, which means that you can't do the question." (Checker)

Although it might be assumed that missing content was always an error, one participant's comment shows that it depended on the context.

> "There can be a missing 'and' that breaks the question, or there can be a missing 'and' that is just irritating." (Checker)

Consistency

A few of the participants' examples concerned the consistency of assessment items. Examples of errors included inconsistency between the stimulus material and other parts of the question, conflicting information or data within questions,

and inconsistent question numbering between the question paper and its associated answer sheet.

> "The writing questions on the answer sheet didn't refer to the questions in the test …. On the answer sheet it was either the question 7 or 8 to choose, but on the question paper it said question 42 or question 43 … They found that it didn't actually adversely affect the candidates because they had to choose one or the other. So, the candidates were either writing 42 or 43 or 7 or 8, and it was clear which one they were answering." (Question paper manager)

Although all the examples identified above were categorised as errors, there were other problems with consistency where participants, particularly the checkers, deliberated as to whether they were errors or issues. This seemed to be particularly challenging where the inconsistency was not obvious to the candidate. Examples of this included inconsistencies in spelling between the question paper and the syllabus, or inconsistencies in the place names used on maps. One of the question paper managers stated that "a consistency [problem] can be an error if it stops the candidate from answering a question". This implies that inconsistencies that do not prevent candidates from answering should be considered to be issues instead of errors.

> "There are inconsistencies in question papers which in themselves aren't errors because they're not spelling mistakes or grammatical errors or errors of data, but they're inconsistent, which also could affect the candidate's ability to answer the question." (Question paper manager)

Alignment with design intentions

There are many ways in which a question paper can fail to align with the design intentions. The most common example of errors of this kind that were mentioned in the interviews was incorrect formatting of the papers (e.g., incorrect font, date format or layout on the page). Participants also gave examples of questions that were not on the specification, had inappropriate levels of demand for the paper, or were not original enough.

> "If the proofers fix something up and say, 'Okay, this is an error because of commas, font size, spacing etc.', then it usually is an error because it doesn't reflect the standard set of the paper." (Question paper manager)

Other participants identified formatting problems as examples of issues. In many cases, the distinction between error and issue appeared to be whether participants considered that they would affect or even be noticed by candidates. This was particularly true for problems with fonts, such as an incorrect font or the lack of bold font.

> "To give you an example, we reported an error on a question paper to Ofqual where we had to tell them that we used a character in a different font. It looked almost identical. You needed a magnifying glass to see the difference. But we had spotted it and it was wrong and we treated it as an error, even though it would have absolutely no impact on candidates." (Senior manager)

Some participants suggested that questions with incorrect levels of demand or that were hard to answer were an issue (as opposed to an error) because the problem was how the item had been written or the way that it could be answered rather than something that was incorrect.

> "It's not until it goes in front of candidates that we discover it's really hard to write an overview, but we wouldn't necessarily treat that as an error. If, for example, it had been live and 50 candidates couldn't write an overview, we might think maybe we should pull the task. We wouldn't say this is an error, we need to go through the error procedure." (Question paper manager)

Offensiveness

Only three participants gave examples of problems with offensiveness, perhaps because it is very unusual for question papers to contain this sort of problem. They considered inappropriate language, including emotive words, to be an error but suggested that this was a bigger problem in some subjects (e.g., history or geography) than others (e.g., mathematics). Two of these participants also talked about inappropriate contexts, although one gave it as an example of an error while the other thought that it was an issue because there was nothing wrong with the question except cultural sensitivity surrounding the context.

> Interviewer: "Are there issues to do with the question paper that you'd say are issues and queries rather than being errors, is there a distinction in that sense?
>
> Respondent: Cultural sensitivity is quite a big problem for us, so it may be something is factually correct but it's just not toned in a correct way, or it's on a topic that we really ought not to be assessing, or it has a viewpoint which wouldn't be appropriate for a certain group."
> (Question paper manager)

Equality of difficulty amongst candidate sub-populations

Very few participants gave examples of problems associated with this construct. The only examples of errors mentioned were cultural sensitivity that led to bias against a particular group of students, and language in the questions that would be difficult for students to access if their first language was not English.

> "If there's anything in there, either cultural sensitivity or linguistic barriers, that might affect a group of people, that could be an [reputational] issue. Because then candidates will respond in different ways and there will be bias introduced, which obviously we want to avoid." (Question paper manager)

This construct was also infrequent among the problems that were classified as issues (as opposed to errors). Three examples were given: the accessibility of language for students whose first language was not English, the inclusivity of papers, and a question on an untiered paper that was not accessible to the whole ability range.

> "Issues, for example, could be if the paper is not as inclusive as we would like it. There may be an issue, for example, for young learners which is highly visual with lots of artwork in it. There may be an issue where every

single person in the artwork could be white, which wouldn't be an error. It would be completely us lacking in looking after our candidates at that point. So we do have policies to try to make the papers be as inclusive and diverse as possible." (Question paper manager)

Stage when problems are detected

Problems with papers are detected throughout the question paper construction process. We asked the senior managers and question paper managers to identify the stage of the production process when they would consider a problem to be an error. There was no consensus in participants' answers. Almost every stage of the question paper construction process was mentioned by at least one participant. The most common response was once a paper was signed off as ready to print, or the equivalent point for on-screen tests. Four of the participants thought that it occurred later than this, either once the paper was printed (or live for on-screen tests), or that it should only be an error if a candidate had seen it.

> "I think we should only ever use the word 'error' for something that hits the candidate's desk in a question paper that would cause them difficulty in answering the question. I would say it's not an error if we catch it at any point before the candidate sees the paper." (Senior manager)

Some of the participants did not identify a stage. Two of the participants thought that problems should always be classified as an error, although they did not think that those earlier errors should necessarily be logged and investigated.

> "If there's something wrong, it's an error at any point; however, I wouldn't regard it as an error that needed to be reported or anything until it's basically been printed and then it would be. So an error is usually at the end and I need to reprint it or do an erratum. However, if there's something wrong, that is still an error but it's not reported as such." (Question paper manager)

Two others said that the stage depended on the type of error that was identified, although they identified different stages for the same example. One thought that SPAG problems were always errors whilst the other thought that they only became an error if they had not been noticed by a proof-reader. One of these participants distinguished between the "tweaks and improvements" that are made during the editing process and major problems with the question.

Near misses

The senior managers and question paper managers were also asked whether they used the term *near miss* in association with errors, as this is a term that is used by some industries such as medicine and nuclear. None of the participants said they used it, but most gave examples of errors that they considered to be near misses. Many were errors that had been found at late stages in the question paper production process, either before the paper was printed or before it reached candidates.

> "If you sign something off, so in a sense, technically, it's an error. But then you catch it before it goes out. That'd be a near miss, in my view." (Question paper manager)

Another common interpretation was to use the term *near miss* for errors that appeared in papers but that apparently went unnoticed by candidates. Examples included errors spotted during marking or after papers had been released.

> "We do get some errors that actually don't get identified in the actual sitting of the exam paper, so candidates have all missed it. They've all got on with the paper quite happily. Then it's only when that paper's been dissected in a staffroom for example that somebody will contact us and say, 'Did you know?' I suppose you could designate that as a near miss because it's on there and no one else has spotted it." (Senior manager)

Participants did not seem to agree about whether errors corrected by erratum notices or reprints were near misses, or just errors.

Discussion

In our introduction we described two main uses of the term *error*: (i) as a human action or inaction; and (ii) as a consequence of an action or inaction. Our analysis of interview data focused on the second of these—individuals' conceptualisations of errors that can arise in question papers and related assessment materials. The data revealed that within Cambridge University Press & Assessment there is no single accepted definition of a question paper error. Although several participants provided clear and succinct definitions, most participants were only able to articulate their understanding of error by describing examples of problems that they considered to be an error and those that they did not. Analysis of these responses suggests that there were three interacting aspects that participants considered when deciding whether a problem should be an error: the manifestation of the error, the impact (or potential impact) upon candidates, and the stage at which it was discovered.

The six validity and reliability constructs that Suto et al. (2021) used to develop their taxonomy of question paper errors (accuracy, clarity, consistency, alignment with design intentions, offensiveness, and equality of difficulty amongst candidate sub-populations) provided a comprehensive way of mapping and analysing the manifestations that participants gave as examples. We found examples that were associated with each of the constructs, but some constructs were associated with more examples than others. There could be many reasons why errors associated with some constructs were mentioned more frequently: errors in those constructs could have been more salient, easier to describe, or participants may have been happier for these errors to go "on-record".

The perceived distinction between errors and issues was an interesting one, and it was here that the importance of the interaction between manifestation, candidate impact and production stage could be observed. Some participants gave examples of errors and issues that appeared to be very similar, and on occasions, identical. An example of the latter was full stops appearing instead of question marks and vice versa. Participants who considered these problems to be issues instead of errors often referred to the impact on candidates, stating that the candidate was either unlikely to notice or unlikely to be affected by

the thing that was incorrect. This influence of candidate impact upon personal conceptualisations relates back to the categories in the error classifications developed by Ofqual. Ofqual states that errors in the least serious category do not affect students' ability to answer the question. However, as Suto and Ireland (2021) state, it is difficult to truly understand the consequences of an error upon candidates, as some candidates will be affected by things that others barely notice.

A less common distinction between issues and errors concerned the correctness of what was written. For some participants, problems were only considered to be errors when there was something that was incorrect. A good illustration of this was the example of inappropriate contexts that was mentioned in the results section. It would be possible to have a question where everything was correct, but that would not be suitable for a particular country or group of candidates because of cultural sensitivities, hence its classification as an issue rather than an error. Similarly, some participants thought that other problems with the wording of questions, such as awkward or complex wording, were an issue if there was no incorrect information within the question. For these participants, the impact on candidates was irrelevant in making the distinction between errors and issues, although both the examples of issues described above were likely to have had an effect upon candidates.

The final aspect of error that seemed to impact upon personal definitions of error and issues was the stage at which the problem was discovered. The results showed that there was no consensus with regard to the stage at which problems should be classified as errors. For many participants, the stage at which problems became errors matched the point at which they had to record them in error logs; however, several other stages were also identified. These ranged from considering a problem to be an error at any stage of question paper production (i.e., from the first draft of a question) at one extreme, to only viewing problems as errors if they appeared in front of candidates without any mitigations (e.g., without an erratum notice) at the other extreme. There are several implications arising from this finding. If authors only consider problems to be errors when they appear in front of candidates, they may not check their questions as thoroughly as an author who considers any problem to be an error at any stage. The same may be true of a checker who does not believe that problems should be classified as errors at the stage at which they were checking the paper. This attitude to checking question papers at earlier points in the process could lead to errors being less likely to be spotted, or lower quality papers. Considering problems at any stage to be an error could improve the quality of question papers, particularly if it helps the question paper writers to see all errors as something that they are responsible for. However, if it led to additional checks being instigated, it could also overload the early checking processes and risk duplication of checks at multiple stages. Similarly, a requirement to log and investigate errors discovered at earlier stages would increase workload for question paper managers and could leave less time for them to carry out their own checks.

In addition to participants' personal definitions of error, they were also asked about their use of the term *near miss*. The results showed that this term is not

commonly used within Cambridge University Press & Assessment and that participants interpreted it in very different ways. The most common interpretation was that it referred to errors that are detected very late in the question paper production process. Participants gave two other interpretations: errors that were not spotted until after candidates had finished sitting the papers (i.e., errors that were spotted at marking or when papers were published), and errors that had been corrected by erratum notices or reprints. In their interpretations of *near miss*, participants were clearly influenced by the stage at which the problem was discovered and the impact it had upon candidates, two of the three aspects that also influenced their definition of errors and issues. Two of the types of near misses identified by participants would have been entered into the error log but it is unclear whether the errors discovered at marking or beyond would be.

Conclusion

To facilitate the identification and analysis of question paper errors, and the efforts to minimise their occurrence, there is a need for consensuses on: (i) the definition of question paper errors; and (ii) how they are distinct from less serious question paper issues. This research reveals that we are currently far from achieving this.

It is not necessarily very helpful for everyone to adopt an all-encompassing definition of an error as anything that is wrong or incorrect within a paper at any stage of the question paper production process. This could lead to error logs becoming unmanageable or to people requesting unnecessary edits to questions at late stages in the question paper production process that increase the risk of another error being introduced. The existing definition used by Ofqual is also not sufficient. Its focus upon the impact of the error on candidates does not necessarily provide enough information to decide whether something is an error or not, as people may not reach the same conclusion about impact. Moreover, impact tells us nothing about where in the production process human failures occurred—essential clues to improving processes in the future. Instead, we argue that the most helpful way to define whether a problem was an error is to use a combination of a description of the manifestation of error, the candidate impact of error, and the stage at which the error was discovered (see Figure 1). We propose that the six validity and reliability constructs (accuracy, clarity, consistency, alignment with design intentions, offensiveness, and equality of difficulty amongst candidate sub-populations) should be used to describe the manifestation of the error.

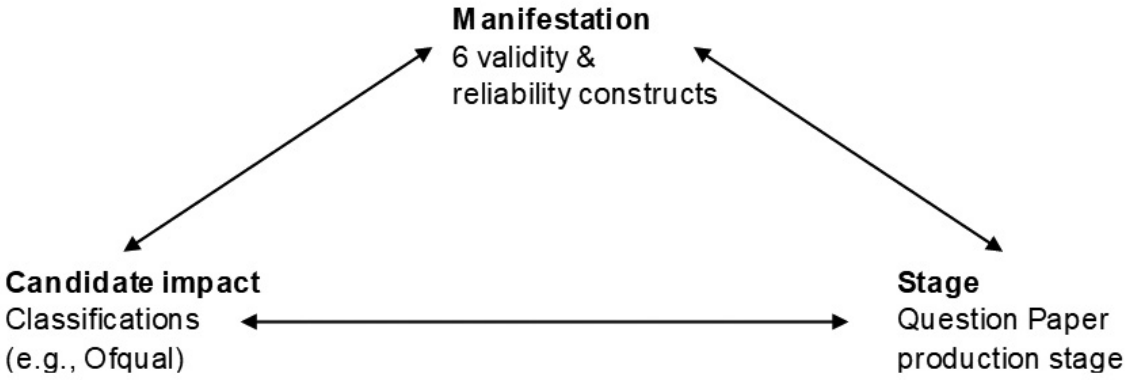


Figure 1: The interaction of manifestation, candidate impact and stage when defining question paper errors.

Any definition should align with, but may not necessarily be the same as, the criteria used to decide whether errors should be included in error logs. For example, the stage at which a problem is considered to be an error for the purposes of the definition may occur before the stage at which an error is added to an error log, but not after it.

Finally, the lack of consensus over the term *near miss* suggests that there is also a need for this term to be clearly defined. We argued that it should be used for the sub-set of question paper errors that are detected late in the checking processes, but are caught just in time (i.e., after printing but before they reach candidates). Such a definition, used in conjunction with a near miss variable in the error logs, would allow investigation into the proportion of errors that are found in time to be corrected, and provide insight into how well the question paper construction and checking processes were working.

References

European Union. (2010). *On the investigation and prevention of accidents and incidents in civil aviation* (Regulation 996/2010). https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ%3AL%3A2010%3A295%3A0035%3A0050%3AEN%3APDF

International Atomic Energy Agency. (n.d.-a). Accident. In *IAEA Safety Glossary*. Retrieved August 27, 2020, from https://kos.iaea.org/iaea-safety-glossary/322.html

International Atomic Energy Agency. (n.d.-b). Incident. In *IAEA Safety Glossary.* Retrieved August 27, 2020, from https://kos.iaea.org/iaea-safety-glossary/754.html

Ofqual. (2019). *GCSE, AS & A level summer report 2019.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/852440/GQ-Summer-Report-2019-MON1100.pdf

Reason, J. (2013). *A life in error: from little slips to big disasters.* Ashgate.

Schubert, C., Winslow, G., Montgomery, S., & Jadalla, A. (2012). Defining failure: the language, meaning and ethics of medical error. *International Journal of Humanities and Social Science, 2*(22), 30–42.

Suto, I., & Ireland, J. (2021). Principles for minimising errors in examination papers and other educational assessment instruments. *International Journal of Assessment Tools in Education, 8*(2), 310–325. https://doi.org/10.21449/ijate.897874

Suto, I., Williamson, J., Ireland, J, & Macinska, S. (2021). On reducing errors in assessment instruments. *Research Papers in Education* (Advance online publication). http://dx.doi.org/10.1080/02671522.2021.1968940

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity.* Cambridge University Press.

Item response theory, computer adaptive testing and the risk of self-deception

Tom Benton (Research Division)

Introduction

For more than a century, the vast majority of high-stakes exams in England have been paper based. Moreover, aside from occasional differentiation of students into tiers, all students taking an assessment are presented with exactly the same set of questions at the same time.[1] This has obvious advantages in terms of transparency. If one student is ranked ahead of another it is simply because, given the same set of questions, one answered more of them correctly than another. All students answering the same questions within a given exam ensures there can be no argument of one student being given an easier assessment than another.

However, as more and more activities in modern life move from the physical to the online realm it is natural for people to consider what benefits might be achieved if high-stakes exams became computer based.[2] The switch to computer-based testing has already begun in other countries such as Israel, Finland and New Zealand (Meadows, 2021). Among the potential benefits that are considered is whether a computer-based format would make it easier to tailor assessments to each individual student through computer adaptive testing.

Computer adaptive testing involves selecting which items to present to a student on-the-fly as the test progresses. In particular, if a student answers an item (or a set of items) correctly then the next item (or set of items) presented to them will be more difficult. Conversely, if a student is struggling, they will tend to be presented with easier items. As such, as the student progresses through the test, the items they are presented with are tailored to match their ability level.

Clearly, computer adaptive testing cannot rely on simply counting how many items students have answered correctly as, by design, some students have been presented with more difficult items than others. To address this, item response theory (IRT) is used. IRT is an overarching theory describing how students respond

1 Although, in a minority of assessments, students may choose which items they answer.

2 Risks are also considered by some authors. See for example Bramley (2021).

to individual test questions (items). In its most common form, it assumes that the probability that any student will answer any item correctly is defined by just two things: a single number describing the ability of the student (unidimensionality) and a small set of numbers (item parameters) describing the key characteristics of the item such as its difficulty and how discriminating it is. For further details, see Harris (1989).

To use IRT within the context of computer adaptive testing, the parameters of all items, such as how discriminating and (most crucially) how difficult they are, must be estimated. This requires some form of trialling before items are used in a high-stakes setting. Then, after students have taken a test, IRT is used to calculate the score that should be assigned to each student while properly accounting for the difficulties of the items they have been presented with (see Wainer et al., 2000).

In theory, computer adaptive testing should make the test more engaging for students as the items they are presented with are more appropriate for their ability. For example, if a student is struggling, rather than being repeatedly presented with questions that are too hard for them to answer, they will find that the test automatically adapts to present them with items more appropriate to their current performance level. Computer adaptive testing should also allow more accurate assessment of each student's ability. For example, ensuring that high ability students are presented with lots of challenging tasks should make it easier to distinguish their relative abilities than if they also had to answer many easy questions.

The potential improvement in measurement precision that can be achieved by computer adaptive testing is normally presented in terms of the extent to which testing time can be shortened without any loss in reliability. That is, rather than keeping the length of exams the same and reducing the level of uncertainty around the score assigned to each student, the benefit of a computer adaptive test (CAT) is usually realised in terms of reducing the length of time students are required to spend taking an exam. According to Straetmans and Eggen (1998, p.51) "on average CATs require about 60 percent of the number of items needed in traditional paper-based test". Other authors suggest that in their specific contexts CATs allow test lengths to be halved with no loss of measurement precision (Kreiter et al., 1999; Weiss, 1982).

The aim of this article is to explore the potential gains from a switch to adaptive testing in the context of large qualifications such as GCSEs and A Levels in England. This context is potentially different from some typical applications of CATs such as general intelligence testing (e.g., non-verbal reasoning tests) or tests of foreign language fluency. In particular, GCSEs and A Levels require students to learn a range of knowledge and skills from a broad range of topics within a subject. As such, the design of examinations is intended to cover numerous topics and skills rather than tightly focus on a single concept.

More specifically, the aim of this article is to better understand whether apparent gains in reliability coefficients from a switch to CATs are likely to translate into real world improvements of the validity of our assessments. The interest in this topic stems from previous research (described in the next section) showing instances

where improvements in reliability do not translate to concomitant changes in predictive validity. The logic for being concerned about a potential gap between supposed (reliability) and actual (validity) benefits of CATs would run as follows:

- In order to even estimate reliability of scores from a CAT we are forced to use IRT. As such, most existing estimates of the improved efficiency through using CATs are based either directly on the output from IRT models or on simulation studies run on the assumption that they are correct.
- The famous aphorism that "all models are wrong, but some are useful" (George Box) clearly applies to IRT. While IRT models function as a very good approximation to our data in many situations, they are not true in an absolute sense.
- Thus, given that, to some extent, the model we are using to estimate scores must be "wrong", how far should we trust estimated improvements in reliability when these are estimated on the assumption that the model is completely correct?
- The real risk here is one of self-deception—thinking that a move to computer adaptive testing is a bigger improvement (in terms of reliability and validity) than it really is. This article will provide an evaluation of the potential size of this risk using real data, that is, not based purely on simulations that assume IRT models fit perfectly.

Previous examples of self-deception risks

This article is by no means the first to draw attention to the risk of self-deception through an over-reliance on IRT models. Three examples are listed below. The first relates specifically to computer adaptive testing and the following two to large-scale empirical analysis of the impact of relying on IRT in other contexts.

Capitalisation on chance

This issue was explored by Veldkamp (2013) and van der Linden & Glas (2000). The issue is that in order to work, computer adaptive testing requires an initial estimate of the difficulty and discrimination of each item. These initial estimates are usually based on relatively small samples of students (perhaps a few hundred), and hence have non-negligible levels of uncertainty attached to them. The result is that, when a CAT selects the next item for a student, it may believe it is selecting a highly discriminating item targeted at just the right ability level, when in fact it is not. Furthermore, since some CATs are designed to try and pick the most discriminating items more frequently, they are liable to tend to select items where the discrimination has been overestimated. As a result, according to Veldkamp and Verschoor (2019, p.293), "the measurement precision of the CATs might be vastly over-estimated".

In other words, the reliability measures generated by a CAT may produce an over-optimistic picture of test quality that is not reflected in reality.

Rescoring using item weights or IRT

Benton (2018) compared various alternatives to simply summing item scores to create overall test scores. The alternatives were intended to help optimise reliability and included both classical methods such as the one suggested by Guilford (1941), and IRT methods such as using a graded response model to produce pupil ability estimates. On average, across analyses of more than 500 assessments, these methods increased the reliability indices (on scales from 0 to 1) from about 0.88 to about 0.89. This may appear a very minor improvement but is actually equivalent to the increase in reliability we might get from lengthening our assessments by 10 per cent[3] (without any of the associated cost). In these terms, it is also very similar to the reported improvement in reliability from transferring the reading literacy tasks in the Programme for International Student Assessment (PISA) to a multi-stage adaptive format in 2018 (OECD, 2019, p.27).

According to IRT, increases in reliability should relate to a reduction in the influence of random error on test scores. This should reasonably be expected to in turn lead to increased correlations with other measures of student ability. However, for Benton's 2018 study the supposed increases in reliability were associated with absolutely no improvement in the predictive value of test scores.

This example illustrates how, if we were entirely reliant on the numbers coming out of an IRT analysis, we might convince ourselves that reweighting items provides an easy way of improving the reliability of test scores at no cost. In fact, the lack of any concomitant improvement in predictive value suggests that, as has been suggested many times in previous research, reweighting items is "futile" (Wang & Stanley, 1970, p.688).

Optimal (fixed) test construction using IRT

Similarly, Benton (2018) compared various approaches to optimal construction of fixed tests. Again, this analysis was based upon real data from more than 500 separate assessments. The research compared the predictive value of half-length tests constructed out of real full-length tests so as to optimise various classical and IRT measures of test reliability. Unlike the research on simply rescoring tests, optimised approaches to item selection did indeed lead to improvements in predictive value when compared to simply selecting items at random. However, the scale of improvement was not as high as might be expected based upon the associated reliability values.

This example again reinforces the possible risks of self-deception from relying entirely upon reliability statistics from IRT analyses. That is, gains in reliability may not necessarily translate into validity. However, the example also accentuates the fact that, while like all models they are "wrong", IRT models are nonetheless "useful". The application of IRT in the study did indeed identify selections of items with greater predictive value on average–just not to the extent that might be hoped given the reliability coefficients.

3 This is easily seen using the Spearman-Brown formula. 0.89≈1.10*0.88/(1+0.10*0.88).

The present study

The current article builds on the item selection research in Benton (2018). In particular, it extends the research to include an examination of the possible gains from allowing the items to be assigned to each student to be selected in an adaptive way, rather than using the same set of items for all students.

The present study makes use of real data throughout. This includes using responses of real students to real items to mimic how they might perform in a CAT, as well as making use of assessment data beyond the tests being studied to give some idea of how different approaches to assessment affect validity. The use of real data for evaluation is crucial. While it is easy to estimate the likely impact of using CATs through simulation, such simulations tend to rely on the assumption that the underlying IRT model is absolutely correct. As such, the use of simulations would entirely undermine the purpose of the research. In order to make some inferences about validity we will look at the correlation of test scores (derived in various ways) with external measures of academic achievement. We will refer to these correlations as "predictive value".

Having said this, the use of real data does have some limitations. All of the data used in the present study is drawn from tests that were originally delivered in a fixed (paper-based) format. This means that the analysis (presented next) cannot entirely mimic the way in which a genuine CAT would operate. In particular, a real CAT would start with a large bank of items that could be presented to students. By carefully selecting which items are presented to each student, the idea would be to either improve test reliability while maintaining test length or to maintain reliability relative to a fixed format while reducing testing time. Neither of these two aims can be tested directly using our real data from fixed format tests. In particular, our methodology will necessarily involve imagining a CAT that assigns each student a subset of items from the original full-length test. Since the imagined CAT is only a subset of the original full-length test it is likely that we will lose rather than gain reliability. As such, the focus will be on which approaches to selecting the subset of items (CAT or fixed form) lead to the smallest losses in terms of reliability and validity. That is, although our real interest is in whether CATs improve test quality, with our data we can only test whether they lead to smaller reductions in reliability and predictive power than other approaches.

A second drawback of using real data from fixed format tests is that they tend to be presented in terms of question stems with a number of subsequent sub-questions. Although, for the purposes of analysis, it is necessary to treat sub-questions as separate items (or else there are too few items to work with) they may not be quite as independent of one another as would generally be the case for distinct items within an item bank underlying a CAT. Although some effort has been made to mitigate the impact of this issue (particularly through checking data for unidimensionality—see below), it remains a caveat against the results presented here.

Method

The data for the present study comes from 159 assessments that were completed as part of GCSEs, A Levels or equivalent international qualifications between 2013 and 2017. The assessments were chosen to meet the following criteria:

- Taken by at least 5000 students. This ensured the accuracy of any item parameters estimated via IRT thereby avoiding issues of capitalisation on chance (see earlier discussion). The median entry size for selected assessments was just under 9000.
- No optional questions. In other words, all questions were compulsory so that whichever items were selected for retention for each student, an item score would be available.
- At least 20 items. This criterion ensured that there would be a reasonable number of items to choose for each student. The median number of items in selected assessments was 32.
- No items worth more than five marks and at least one item worth only one mark. Since the focus of this article is on computer adaptive tests, and such tests rarely (if ever) incorporate items with long mark scales, it seemed reasonable to restrict attention to assessments consisting of relatively low tariff items. Having said that, none of the assessments included in analysis consisted entirely of one-mark items.
- The assessment was deemed to be unidimensional. Unidimensionality was important for the analysis as the intention was to focus upon CATs based on unidimensional IRT models. Unidimensionality was confirmed for each of the assessments using Velicer's MAP criterion (Velicer, 1976) as evaluated by the R package *psych* (Revelle, 2020).[4]

The principle of analysis is as follows. For each assessment we apply some method to select items for each student, calculate a score for each of them based only on data from the selected items, and then calculate the correlation[5] between the resulting scores and a measure of the students' achievement more widely. We label these correlations "predictive value". We also calculate estimates of the reliability of scores from the item selection method. Finally, we compare both predictive value and reliability from the selected items against the original value based on retaining the whole full-length test. The idea is that methods that are more effective at selecting the most appropriate items for each student will retain a greater amount of reliability and predictive value from the full-length tests.

For the purposes of calculating predictive value, the wider achievement of each student was measured via each candidate's external ISAWG[6] (Benton, 2017). The

4 Note that this criterion led to the removal of several hundred assessments from those available for inclusion in the study.

5 To avoid potential issues with outliers, and also the impact of the scales used for different scoring systems, Spearman's rank-order correlations were used.

6 ISAWG stands for Instant Summary of Achievement Without Grades.

external ISAWG is a measure of each candidate's performance across all of the tests that they have taken in a particular examination session, excluding the one being analysed. It is derived using a form of principal components analysis and can be interpreted as a very general form of ability across different subjects. It was used in this analysis as it was easily available for nearly all the candidates included in analysis.

Analysis focused on three methods that selected a single, optimal set of items for all students and two CAT-like approaches where the selected items could vary across students. The three single-form methods were: to select items at random; to select items that maximise expected test information (a concept from IRT relating to the likely reliability of a test) based on a Rasch partial-credit model (PCM); and to select items that maximise expected test information based upon a graded response model (GRM). The CAT-like approaches each attempted to maximise the expected test information for each student individually using either a PCM or a GRM.

The difference between the PCM and the GRM is that the former requires estimation of item difficulty only whereas the latter also estimates the discrimination of each item. In theory, the GRM approach should be superior in that it can ensure that the most discriminating items are selected in addition to ensuring that they are at the most appropriate level of difficulty for the students. In contrast, the PCM model assumes that items worth the same number of marks have the same discrimination parameters and focuses purely on ensuring that items of the most appropriate difficulty are selected. Evaluating whether extra focus of the GRM on how well each item discriminates between students of different abilities actually translates into improvements in predictive value was a key question within this research.

Each item selection method was designed to select items worth half the total number of marks available on the original full-length test.[7] Furthermore, the selected items were intended to reflect as closely as possible the distribution of item tariffs (i.e., the maximum available marks on each item) in the original test.

To further illustrate the procedure that was applied for each assessment, we consider a 40-mark Biology test that was included in analysis. The test consisted of two 4-mark items, two 3-mark items, eight 2-mark items and ten 1-mark items. In this particular instance, each method was designed to select one 4-mark item, one 3-mark item, four 2-marks items and five 1-mark items. Further details on each method are below:

- **Fixed test with random selection of items.** The required number of items with each tariff were simply selected at random. The scores on these same items were retained for all students.
- **Fixed test with item selection relying on the GRM.** First, we fitted a GRM model to the full data set and calculated the item information functions for

7 Real CATs may use more complex stopping criteria such as whether the estimated error of measurement for each student is below some threshold.

each item. These provide an estimate of how much information each item is expected to provide about students at each ability level. For estimation of the GRM the distribution of ability was assumed to follow a normal distribution with a mean of 0 and a standard deviation of 1. Using this fact, we then calculated the expected information we expect each item to supply across students (i.e., averaging the item information functions across the ability distribution). For each item tariff we then selected the items with the highest expected information for retention. The scores on these same items were retained for all students.

- **Fixed test with item selection relying on the Rasch PCM.** The same process as for selecting a fixed test using the GRM was followed. The only difference was that a different IRT model was fitted as a starting point. For estimation of this model, it was assumed that the ability distribution was normal with a mean of 0. However, because, in contrast to the GRM, discrimination parameters are fixed, the model estimates the standard deviation of abilities and this estimate was used in the subsequent calculation of the expected information from each item.
- **CAT-like test with item selection relying on the GRM.** The initial steps for this approach were the same as for the fixed test based on GRM in terms of model fitting and calculation of item information functions. After this, the following procedure was followed separately for each individual student.

    1. Initially set the distribution of the student's ability to be normal with a mean of 0 and a standard deviation of 1.
    2. From the items with the highest tariff still required, select an item with the highest expected information given the individual student's ability distribution.[8] That is, if we still need a 4-mark item we select from among these, if we have already selected sufficient 4-mark items we select from among 3-mark items and so on. Starting with items with the highest tariff makes sense as these are most likely to provide useful information about candidates across a range of different abilities. Choosing items with the highest expected information close to each student's estimated ability will tend to mean more difficult items are assigned to high performing students and easier ones are assigned to lower achievers.
    3. Based on the student's (known) response to the item, update their ability distribution. For example, if they have answered an item fully correctly the mean of their ability distribution will be adjusted upwards whereas if they have answered incorrectly, it will adjusted downwards. The uncertainty around their ability estimate (i.e., the standard error) will also be adjusted.
    4. Unless we have selected all of the items we require of the various tariffs return to step 2 until complete.

8 The first item selected for each student is the same. Subsequent questions will differ across students.

5. The final IRT ability estimate of each student based on their individually selected items is used as their final score. Note that these ability estimates will adjust for the difficulty of the items that were assigned to each student and also give more weight to performance on items estimated to have a higher discrimination.

- Note that simulating a CAT using the above procedure assumes that we would be able to automatically mark all items regardless of their tariff or format. That is, we are assuming that all technological barriers to computer-based testing and auto-marking have been overcome so that we could run a CAT using the same style of items currently used in qualifications in England. This is a fairly large assumption but is used here to allow us to explore the potential of computer adaptive testing in a best-case scenario. From a technical perspective note that all ability estimations made use of expected a posteriori (EAP) estimation and that the item selection approach reflects the posterior-weighted information criterion described by van der Linden and Pashley (2010).
- **CAT-like test with item selection relying on the Rasch PCM**. The procedure was exactly the same as above but with all calculations, including calculating information function and assigning ability estimates (including final scores) to students, based upon the Rasch PCM model. Crucially, the Rasch PCM model assumes the same discrimination parameters for items with the same tariff. This means that the model will not give additional weight to performance on items estimated to be highly discriminating.

Having calculated the scores that would be assigned to each student by each method all that remained was to calculate predictive value and reliability. Predictive value was calculated as the Spearman correlation between final scores and the external ISAWG (i.e., performance more widely beyond the assessment of interest). Note that for fixed form tests, final scores were always simply the sum of the item scores on the selected items. For CAT-like tests, the final scores were based on EAP ability estimates as described above.

There are many ways to calculate test reliability. However, in order to enable the best possible comparability between different techniques, an IRT method of estimating test reliability was calculated for each test score. For the CAT-like methods, this was simply provided by the reliability indices associated with their final set of IRT ability estimates. As noted earlier, for the fixed form methods, each student's score was simply a sum of scores on the selected items. In order to allow comparability with the other methods, these sum scores were converted to equivalent values on the IRT ability scale using the EAP approach of Thissen et al. (1995). The reliabilities of the EAP ability estimates (derived from sum scores) were then calculated. The same approach was used to estimate the reliability of the original full-length test. All model fitting and calculations relating to IRT were performed using the R package mirt (Chalmers, 2012). If we denote the estimate of each student's IRT ability estimate as $\hat{\theta}_i$ and the uncertainty around this estimate as $SE(\hat{\theta}_i)$ then the formula to estimate reliability is:

$$Reliability = \frac{Var(\hat{\theta}_i)}{Var(\hat{\theta}_i) + Mean(SE(\hat{\theta}_i)^2)}$$

Note that for the two separate IRT approaches (GRM and PCM), reliability was calculated on each model's own terms. That is, if the CAT or fixed form was derived from the GRM, then the reliability index was also calculated using this model. If the CAT or fixed form was derived using the PCM, then reliability was also estimated using this model. For this reason, the reliability indices from the different models are not directly compared.[9] For both the full-length test and the random selection of items, both reliability types of index were calculated. Note that, although calculated differently, both reliability coefficients (measured on scales from 0 to 1) can be interpreted in a similar way to more familiar indices such as Cronbach's alpha.

Results

To begin with, we examine the results relating to reliability. These are shown in Figure 1 in two panels relating to the two separate IRT models that can be used to estimate reliability. Each point on the chart represents the reliability of a full-length assessment (the x-axis) and the extent to which this reliability changes (the y-axis) under various approaches to selecting only half the items for each student. Thus, for each assessment the chart includes three points in each panel (one relating to each method) and these are positioned in a vertical line. For example, the leftmost set of points relate to an assessment with an original full-length reliability just above 0.65. Selecting half the items using a CAT-like approach based on a GRM barely reduced the reported reliability. In contrast, in this instance, a fixed form based on the GRM reduced reported reliability by about 0.03 and selecting half the items at random reduced the reliability by about 0.12.

The overall pattern of results in Figure 1 is as expected. CAT-like approaches led to lower reductions in reliability relative to the full-length test than selecting a single fixed form for all students. Although care is needed with the comparison, when judged on their own terms, the extra emphasis on selecting highly discriminating items based on the GRM (and giving more weight to them in scoring) led to smaller reductions in reliability than the CAT-like approach based on the Rasch PCM. Indeed, in one case, through giving more weight to scores on highly discriminating items, the CAT-like approach appears to lead to improved reliability relative to the original full-length test despite consisting of only a subset of the items for each

9 Although it is possible to estimate the reliability of scores derived using one model based upon another model, it is not particularly straightforward. It is also not something I have ever seen done in practice. For these reasons it is avoided in this article.

student. Among the two fixed form approaches in each panel of Figure 1, selecting items in an optimal manner based on an IRT model led to higher reliabilities than selecting items at random.

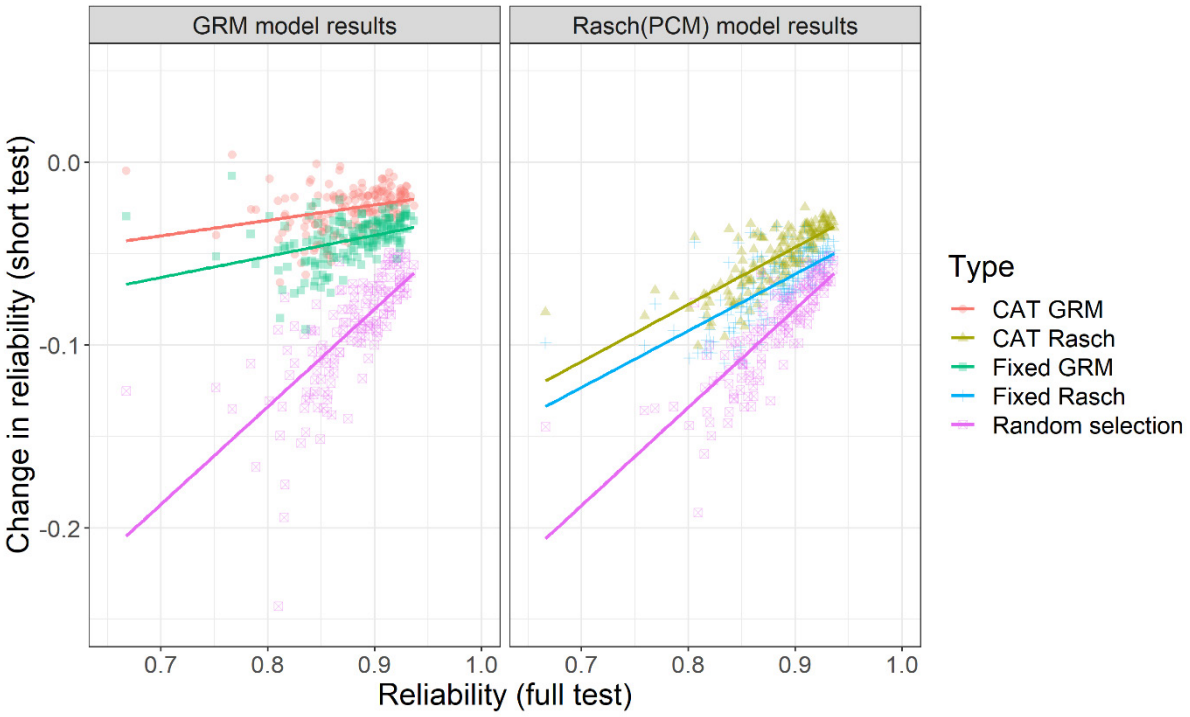


Figure 1: Original full-length test reliabilities and changes in reliability under each of the methods for selecting a subset of items for use with each student. Regression lines have been added to aid interpretation. Results are split by the IRT model used to calculate reliability.

Of most interest in this research is the extent to which the results relating to the superior reliability coefficients of CAT-like approaches in Figure 1 translate into superior predictive value. This is explored in Figure 2. Figure 2 is designed to follow the same pattern as Figure 1 but plots predictive values for the full-length test and changes in predictive values rather than reliabilities. Note that although predictive value can be directly compared across all methods (i.e., between PCM and GRM approaches), for consistency with Figure 1 the split by IRT model is retained.

As can be seen, Figure 2 creates a rather different impression to Figure 1. The advantages of the CAT-like approaches over other methods are reduced relative to the gaps shown in Figure 1. Most surprisingly, the gap between the CAT-like approach based on the Rasch PCM and fixed item selection based on the same model has vanished. On the other hand, the gaps between choosing optimal fixed form tests (using either the PCM or GRM) and selecting fixed form tests at random remain strongly evident.

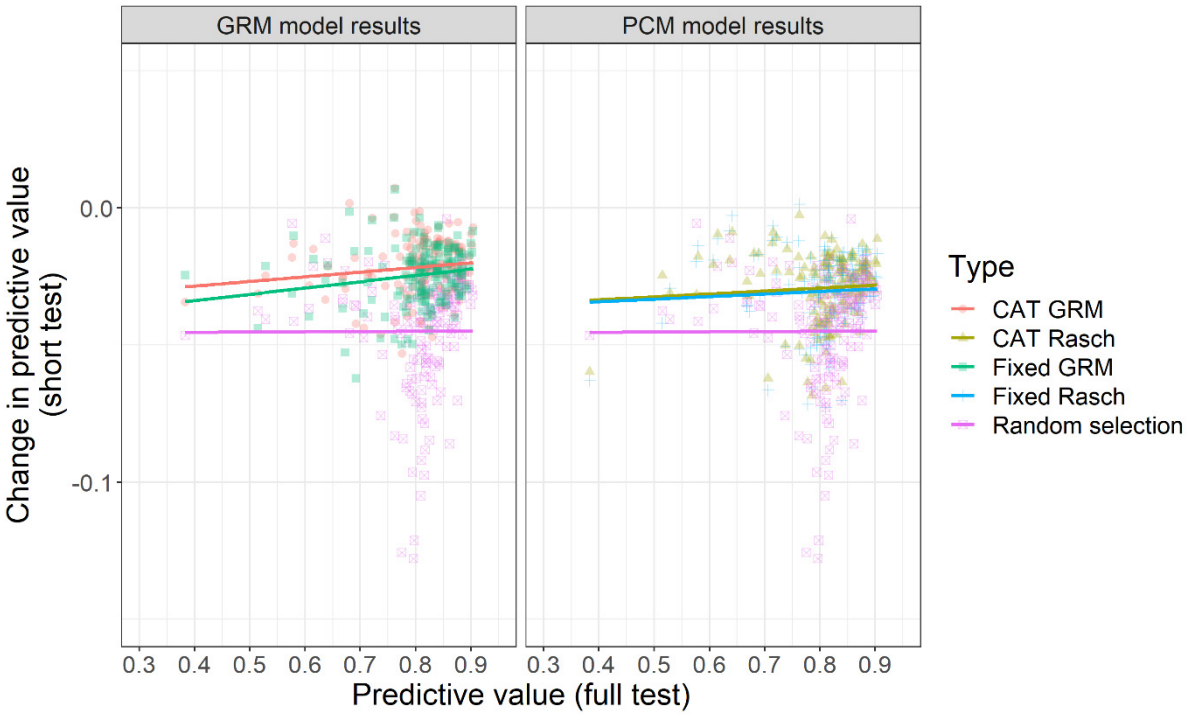


Figure 2: Original full-length test predictive values and changes in predictive value under each of four methods for selecting a subset of items for use with each student. Regression lines have been added to aid interpretation (the lines for CAT Rasch and Fixed Rasch are almost identical).

In order to interpret the two figures, it is helpful to have some idea of how much we would expect changes in reliability to impact upon predictive value. This can be calculated using the following simple formula based upon classical test theory:

$$Expected\ change\ in\ predictive\ value = Original\ predictive\ value\left(\frac{\sqrt{New\ reliability}}{\sqrt{Original\ reliability}} - 1\right)$$

Changes in reliability relative to the full-length test for the CAT-like GRM approach are compared to changes in predictive value in Figure 3. Each point in the chart represents an assessment. The jagged red line represents the expected change in predictive value based on the change in reliability using the formula above. The line is jagged as the change in predictive value depends not only upon the change in reliability but also upon the original values of predictive value and reliability. As can be seen, although there are exceptions, for the majority of assessments the change in predictive value is much worse than would be expected given the reported changes in reliability coefficients.

If the analysis in Figure 3 is reproduced using simulation, then changes in predictive value are far closer to the predicted values based on the above formula. In other words, the failure of changes in estimated reliabilities to be reflected in changes in predictive value must relate to some form of lack of fit in the underlying IRT model. This will be discussed more later. A particularly striking

feature of Figure 3 is the weak relationship between changes in estimated reliability and changes in predictive value. A possible explanation for this is that, in reality, shortening a test (whether using a CAT or otherwise) not only alters reliability but also has some slight impact upon the construct being measured. The changes may either strengthen or weaken the relationship with external measures of achievement. This could lead to the noisy pattern we see in Figure 3.

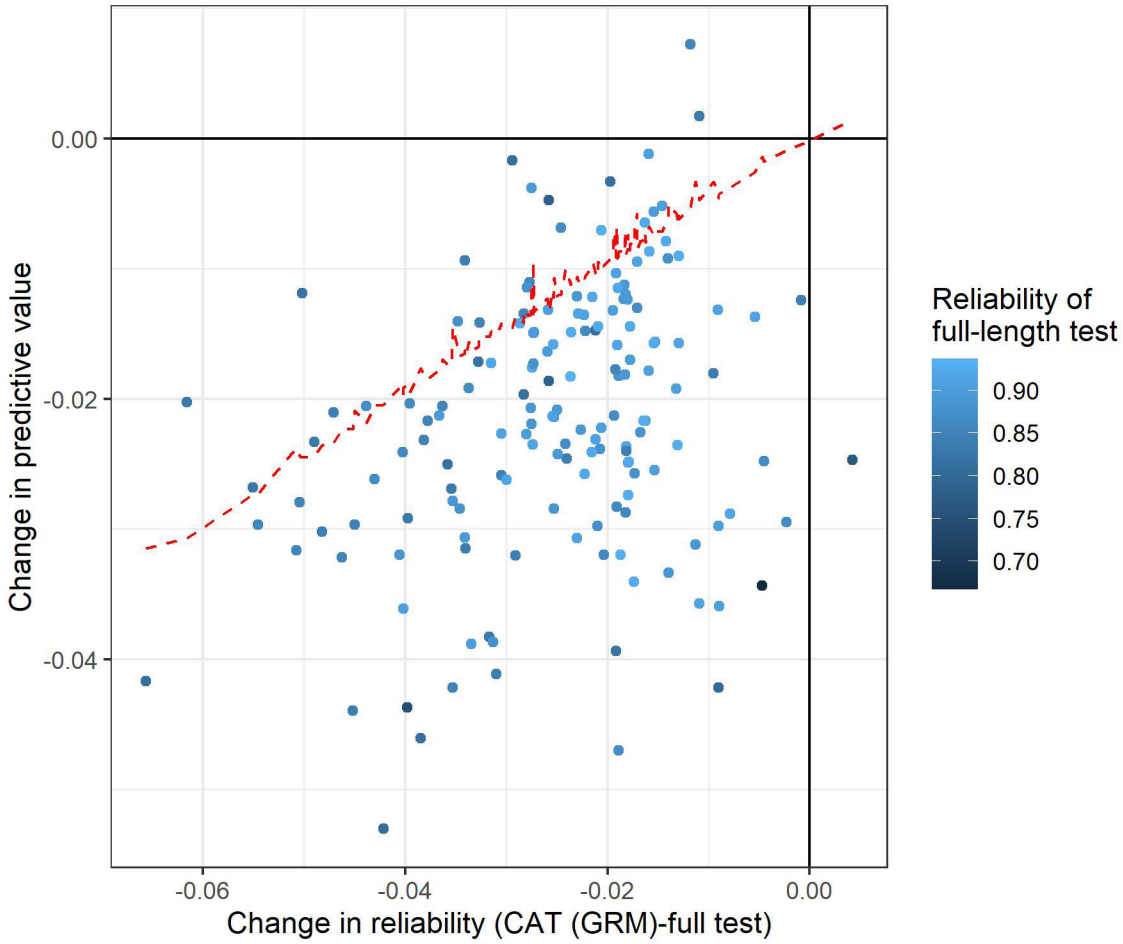


Figure 3: Changes in reliability against changes in predictive value from applying a CAT-like approach based on the GRM. The red line indicates the expected change in predictive value based on a formula from classical test theory.

Table 1 shows the mean predictive value and (relevant) reliabilities across all 159 assessments from each approach as well as the full-length assessments. Note the need to report results to three decimal places in order to properly reveal findings. The highlighting in Table 1 is used to group methods using the same model for item selection.

Table 1 repeats many of the findings described above in a different way. For example, the gap in reliability between a CAT-like and fixed test based on the PCM (of 0.015) does not translate into any meaningful difference in average predictive values (0.001). Similarly, the gap between CAT-like and fixed tests where item selection is based on the GRM is also much smaller in terms of predictive value (0.003) than in terms of reliability (0.018).

The final three columns attempt to convert the mean reliabilities and predictive values into equivalent test lengths relative to a full-length test. The columns based on reliabilities use the Spearman-Brown formula (Spearman, 1910) to convert the mean reliabilities into an equivalent test length compared to the full-length test. The use of the Spearman-Brown formula in this way effectively assumes that items are selected at random. Reassuringly, Table 1 shows that the mean reliabilities of half-length tests selected at random are indeed, according to the Spearman-Brown formula, equivalent to a randomly selected test of about half the length of the full test. The various optimal approaches to item selection for both CAT-like and fixed form tests perform better in terms of reliability. Despite only requiring half the items from the full-length test they achieve an average reliability equivalent to a randomly selected test of between 59 and 81 per cent of the length.

A similar process can be used to generate equivalent test lengths based on predictive value. First, the formula provided earlier is used to convert mean predictive values into equivalent reliabilities. These are then converted into equivalent test lengths using the Spearman-Brown formula. These test lengths are generally lower than those based on reliability coefficients—especially for the CAT-like tests. For example, while the reliability coefficients might lead us to believe that a half-length CAT (based on the GRM) was worth a randomly selected test of 81 per cent length, predictive value suggests it may only be as good as a randomly selected test of 69 per cent length. The only item selection method (besides random) where the equivalent relative length is just as high whether it is based on predictive value rather than reliability is the creation of a fixed form test based on the Rasch PCM. This may be because the rather conservative nature of this approach (essentially just picking items of about the right difficulty for the average student) has less scope for over-optimism about reliability. Also, being a fixed form test, it avoids the need to provide comparable scores for students that have taken different items and the associated additional reliance on assumptions from a given model.

Table 1: Reliabilities, predictive values and associated equivalent test lengths for various approaches to test construction.

		Mean across 159 assessments of...			**Equivalent relative random length based on mean...**		
Method	**Scoring method**	GRM reliability	PCM reliability	Predictive value	GRM reliability	PCM reliability	Predictive value
Full-length test	**Sum score**	0.878	0.881	0.806	100%	100%	100%
CAT (GRM)	IRT	0.853	-	0.785	81%	-	69%
Fixed (GRM)	Sum score	0.835	-	0.782	71%	-	66%
CAT (Rasch PCM)	IRT	-	0.829	0.777	-	65%	62%
Fixed (Rasch PCM)	Sum score	-	0.814	0.776	-	59%	60%
Random	Sum score	0.786	0.791	0.761	51%	51%	50%

Model fits

As mentioned above, repeating the entire exercise using simulated rather than real data leads to much closer agreement between changes in reliability and changes in predictive value. As such, the fact that for our CAT-like approaches higher reliabilities hardly translate into higher predictive values must in some way mean that the model assumptions are not correct. This section discusses the various ways in which real data may not conform to an IRT model and the extent to which this is practically detectable.

The first thing to note is that, in terms of the indices of model fit typically used in IRT, our data did not reveal any obvious problems. Firstly, we consider the fit of the Rasch PCM model. The fit of each item in each data set to the Rasch model was evaluated using inlier-sensitive or information-weighted fit (INFIT) and outlier-sensitive fit (OUTFIT) (Linacre, 2002). Of the 4970 items in the analysis (across all data sets) only 70 (1.4 per cent) had values for these fit indices outside of the range between 0.5 and 1.5 which, according to Linacre (2002), is required to ensure items are "productive for measurement". Only 11 items in total (0.2 per cent) had values of either INFIT or OUTFIT in excess of 2 indicating severe lack of fit. In other words, the vast majority of items displayed a level of fit with the Rasch model that would be deemed acceptable in most operational contexts. Nonetheless, even the relatively small amount of lack of fit in the data appeared to be enough so that apparent gains in reliability may not translate into improvements in predictive value.

We next consider the fit of the GRM models to the data. To check this, overall goodness of fit statistics (root mean square error of approximation RMSEA and Standardized Root Mean Square Residual SRMSR, see Maydeu-Olivares, 2013) were calculated for each of the real data sets. Using these metrics, it was determined that 152 out of 159 of the data sets had values for RMSEA below the level of 0.05 which was recommended by Maydeu-Olivares (2013) as indicating adequate fit. The very largest value of RMSEA across all data sets was only slightly above this threshold at 0.07. Similarly, for 154 of 159 data sets, the value of SRMSR (an easier to understand metric that simply calculates how far pairwise item correlations in each data set are from their predicted values based on GRM on average) was below 0.05—a "substantively negligible amount of misfit" (Maydeu-Olivares, 2013, p.84). The largest value of SRMSR was also 0.07. In other words, by any normal operational definition, the GRM had a very good fit to all of the data sets in the analysis.

Despite the relatively good fit of the data to the various IRT models described above, it is possible that even the small amounts of lack of fit were sufficient to mean that differences in reliability between different techniques did not translate into differences in predictive value. This indicates that the issues shown in the above analysis are not easily detectable simply by looking at the outputs of IRT analyses.

The above measures of model fit are internal in the sense that they look at the extent to which relationships between items within the same test adhere to expectations. However, they do not reflect all of the assumptions of the IRT

model. Perhaps the most crucial assumption of IRT in our context is the definition of measurement error. Usually, and certainly in nearly all simulation studies, measurement error is thought of as entirely random and, thus, unrelated to any external variables. However, this highly simplified conception of measurement error may not reflect reality. In particular, improvements in reliability indices via optimal item selection may not simply mean the removal of purely random measurement error. Rather, they may represent a change in emphasis regarding which specific pieces of knowledge are regarded as particularly pertinent to the construct and which are not. In other words, different approaches to item selection may lead to changes to the construct being assessed. Such changes may or may not be desirable dependent upon the purpose of the assessments. However, we need to remain aware of this potentially unintended consequence of switching to a CAT format and not assume that results from reliability coefficients and simulation studies tell the full story.

Discussion

In many ways, the analysis in this article supports the "useful" nature of IRT models and, in particular, their value developing CATs. On average, test scores derived from a simulated CAT process had higher predictive value than any single fixed test across students. Similarly, there were no instances where using a CAT and the associated algorithm for producing student scores led to markedly lower predictive value than using a random selection (and in most cases it was better). Thus, the article is not a criticism of the use of CATs in themselves. What is at stake here is the rather more technical, but nonetheless important, topic of whether we are able to accurately evaluate test quality based on the output from IRT analyses alone, or whether we risk deceiving ourselves that changes are leading to improved validity when in fact they do not. That is, whether a focus on reliability indices risks overselling the advantages of CATs.

The results of analysis show that, relative to fixed form tests, expected advantages in test quality (based on reliability indices) may not always necessarily translate into verifiably higher predictive values. Having said this, the differences between expectations based on reliability and actual predictive values were often quite small in real terms.

It is worth admitting that very few people are likely to care about the levels of difference in reliability (or predictive value) described in this article. For example, how many people would really care about whether an assessment's correlation with achievement more widely is 0.78 or 0.77? However, the point is that the results here form part of a wider body of work questioning whether computer adaptive testing will necessarily result in improved test quality in every context. For example, previous research (Veldkamp, 2013) has already demonstrated how the uncertainty in the estimated parameters of items used in a CAT may mean that they are less effective than thought. More generally, the issue is that in a CAT we are highly reliant on the accuracy of an IRT model for correctly scaling the scores of students who have taken different sets of items against one another. If the underlying IRT model is not correct in every respect, this may lead to some degree of error in this process.

With these risks in mind, and in the context of high-stakes examinations covering a broad range of content such as GCSEs and A Levels, it is worth considering whether the potential benefits of computer adaptive testing are sufficient relative to the added difficulty in ensuring comparability between scores from different students. It is interesting to note that, in practice, CATs do not always lead to the level of improvement in reliability that might be hoped for. For example, ETS researcher Martha Stocking once quipped that real tests often had so many additional constraints such as ensuring content coverage and avoiding overexposure of individual items that most CATs were actually BATs (barely adaptive tests) (Chuah et al., 2006). Given the likely requirement to ensure that examinations continue to cover the majority of the taught curriculum for each student, this would be a particular risk in the context of qualifications such as GCSEs and A Levels.

In considering the value of CATs, it is worth noting that many of their benefits relate to the application of computer-based testing more generally rather than the adaptive nature of the tests. For example, van der Linden and Glas note advantages such as "the possibility for examinees to schedule tests at their convenience; tests are taken in a more comfortable setting and with fewer people around than in large scale paper-and-pencil administrations; electronic processing of test data and reporting of scores are faster; and wider ranges of questions and test content can be put to use" (van der Linden & Glas, 2010, page vi). All of these advantages are good reasons to explore the possibility of extending the use of computer-based testing in England. Chasing high reliability coefficients through CATs should very firmly stay in second place.

References

Bramley, T. (2021, March 31). Online assessment - the robustness and resilience of the exam system (part 1). *Cambridge Assessment Website blog*. https://www.cambridgeassessment.org.uk/blogs/the-robustness-and-resilience-of-the-exam-system-part-1/

Benton, T. (2017, November). *Pooling the totality of our data resources to maintain standards in the face of changing cohorts.* Paper presented at the 18th annual AEA-Europe conference, Prague, Czech Republic. http://www.cambridgeassessment.org.uk/Images/429428-pooling-the-totality-of-our-data-resources-to-maintain-standards-in-the-face-of-changing-cohorts.pdf.

Benton, T. (2018, November). *Exploring the relationship between optimal methods of item scoring and selection and predictive validity.* Paper presented at the Association for Educational Assessment – Europe conference, Arnhem/Nijmegen, The Netherlands. https://www.cambridgeassessment.org.uk/Images/525258-exploring-the-relationship-between-optimal-methods-of-item-scoring-and-selection-and-predictive-validity.pdf.

Chalmers, R. P. (2012). *mirt:* A Multidimensional Item Response Theory Package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. http://www.jstatsoft.org/v48/i06/.

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education, 19*(3), 241–255. https://doi.org/10.1207/s15324818ame1903_5.

Guilford, J. P. (1941). A simple weight scoring for test items and its reliability. *Psychometrika, 6*(6), 367–374. https://doi.org/10.1007/BF02288593

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35–41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

Kreiter, K. D., Ferguson, K., Gruppen, L. D. (1999). Evaluating the Usefulness of Computerized Adaptive Testing for Medical In-course Assessment. *Academic Medicine, 74*(10), 1125–1128. https://doi.org/10.1097/00001888-199910000-00016.

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement, 11*, 71–101. https://doi.org/10.1080/15366367.2013.831680

Meadows, M. (2021, June 15). Speech at City of London Schools Conference 2021. GOV.UK. https://www.gov.uk/government/speeches/dr-michelle-meadows-speech-at-city-of-london-schools-conference-2021.

OECD. (2019). *PISA 2018 Technical Report*. Chapter 2 - Test Design and Test Development. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-02-Test-Design.pdf./

Revelle, W. (2020) *psych: Procedures for Personality and Psychological Research*. https://CRAN.R-project.org/package=psych. Version = 2.1.3.

Spearman, C. (1910). Correlation Calculated from Faulty Data. *British Journal of Psychology, 3*, 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Straetmans, G. J. J. M., & Eggen, T. J. H. M. (1998). Computerized Adaptive Testing: What It Is and How It Works. *Educational Technology, 38*(1), 45–52. https://www.jstor.org/stable/44428447

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39–49. https://doi.org/10.1177%2F014662169501900105

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of Adaptive Testing*. Springer.

Van der Linden, W. J., & Pashley, P. J. (2010). Item Selection and Ability Estimation in Adaptive Testing. In Van der Linden, W. J., & Glas, C. A. W. (Eds.), *Elements of Adaptive Testing*. Springer.

Veldkamp, B. P. (2013). Ensuring the future of Computerized Adaptive Testing. In Eggen, T. J. H. M. & Veldkamp, B. P. (Eds.), *Psychometrics in practice at RCEC* (pp.137–150). RCEC. https://ris.utwente.nl/ws/files/253342308/Eggen2012psychometrics.pdf#page=43.

Veldkamp, B. P., & Verschoor, A. J. (2019). Robust computerized adaptive testing. In Veldkamp, B. P. & Sluijter, C. (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp.291–305). Springer. https://library.oapen.org/bitstream/handle/20.500.12657/22945/1007216.pdf?sequence=1#page=291.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*(5), 663–705. https://doi.org/10.3102%2F00346543040005663

Weiss, D.J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement, 6*(4), 473–492. https://doi.org/10.1177%2F014662168200600408

Cambridge Assessment Network

Build a CPD programme with experts from Cambridge

Flexible assessment training packages

Whatever your organisation's professional development needs, as assessment leaders we are committed to supporting your people.

Our ready-made options cover most requirements giving you and your team the confidence and knowledge to transform your assessment practice.

Example sessions include:

- The purposes and principles of good assessment
- Writing multiple choice questions for impact
- Writing effective examination assessment questions
- Designing assessment strategies

Just take your pick from our range of training sessions and let us know what you need.

"The session gave participants ideas, strategies and some reflective learning as takeaways as well as some useful aide-memoire handouts. The Network team did an excellent job and covered our brief to the letter."

Dr Andrew Harries, University of Central Lancashire, on an in-house Multiple Choice Questions session

Menu of options: cambridgeassessment.org.uk/your-cpd

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

The future of education, are we any closer to knowing?

The future of education is being widely debated. Cambridge University Press & Assessment is bringing an international lens to the discussion through original research and focused insights.

During 2021, we've actively countered misconceptions about teaching, learning and assessment, and shone light onto the educational reform experiences of many countries around the world which underline the importance of policy formation being driven by evidence.

cambridge.org/future-of-education

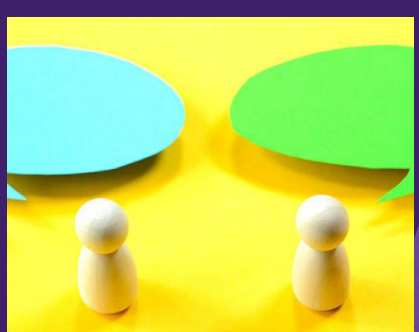
Where exams are taken at age 16 around the world

Curriculum coherence and high-performing education systems

Solving the problem of interrupted learning

Building lifelong skills for the 21st century workplace

Cognitive science in developing teaching practice



Benefits of first and second languages to learners

Research News

Anouk Peigne Research Division

The formation of Cambridge University Press & Assessment

In August, Cambridge University Press and Cambridge Assessment joined together to form a single organisation. With the Press doing far more work on incorporating assessment into learning materials, and Cambridge Assessment dramatically increasing its work on curriculum more widely, this is a welcome development. In the reform and improvement work which we are doing around the world, the link between learning and assessment is becoming stronger and stronger; so the creation of the single organisation makes tremendous sense. It is entirely consistent with perspectives from international comparative research, which highlights the importance of "curriculum coherence" between curriculum aims, learning programmes and approaches, and assessment—something which we have emphasised strongly over the past decade. We will continue to work with a wide range of publishers globally, and continue to innovate in both assessment and learning—not least in providing immediate responses to the disruption to education caused by the pandemic, and shaping the form of post-pandemic arrangements.

Publications

The following reports and articles have been published since *Research Matters*, Issue 31:

Coleman, V. (2021). *Digital divide in UK education during COVID-19 pandemic: Literature review.* Cambridge Assessment Research Report. https://www.cambridgeassessment.org.uk/Images/628843-digital-divide-in-uk-education-during-covid-19-pandemic-literature-review.pdf

Constantinou, F. (2021). How novel can examination questions really be? Exploring the boundaries of creativity in examination question writing. *Research Papers in Education* (Advance online publication). https://doi.org/10.1080/02671522.2021.1961297

Mouthaan, M. (2021). Old Wine in New Bottles? The European Union's Organizational Response to Reforming EU-African Migration Cooperation. *Journal of Common Market Studies* (Advance online publication). https://doi.org/10.1111/jcms.13203

Suto, I., & Ireland, J. (2021). Principles for minimising errors in examination papers and other educational assessment instruments. *International Journal of Assessment Tools in Education, 8*(2), 310–325. https://ijate.net/index.php/ijate/article/view/37

Suto, I., Williamson, J., Ireland, J., & Macinska, S. (2021). On reducing errors in assessment instruments. *Research Papers in Education* (Advance online publication). https://doi.org/10.1080/02671522.2021.1968940

Vidal Rodeiro, C. L., & Vitello, S. (2021). Progression to post-16 education in England: the role of vocational qualifications. *Research Papers in Education* (Advance online publication). https://doi.org/10.1080/02671522.2021.1961295

Conference presentations

The British Educational Research Association's annual conference was held online in September 2021. Our researchers presented two papers:

Vidal Rodeiro, C.L., & Vitello, S. (2021, September 13–16). *Progression to post-16 education: the role of vocational qualifications* [Paper presentation]. British Educational Research Association 2021 Conference, online.

Vidal Rodeiro, C.L., & Macinska, S. (2021, September 13–16). *Equity or unfair advantage? Impact of access arrangements on students' performance* [Paper presentation]. British Educational Research Association 2021 Conference, online.

On 27–28 September 2021, Cambridge University Press & Assessment hosted a 2-day online seminar for researchers at UK awarding organisations and regulators to share their latest thinking. On average over 100 delegates attended the five sessions, which contained 23 presentations in total, grouped into themes of: the future; comparative judgement; accessibility and inclusivity; marking and teacher assessment. AQA, Pearson, No More Marking, CCEA, Ofqual and SQA all contributed.

There were 12 presentations from Cambridge University Press & Assessment:

Effective teaching and learning during the pandemic. Alison Rodrigues and Lynda Bramwell.

How well do we understand wellbeing? Teachers' experiences in an extraordinary educational era. Chris Jellis.

Comparative judgement for moderation: a feasibility study. Lucy Chambers and Carmen Vidal Rodeiro.

Awarding using comparative judgement: do judges attend to construct-irrelevant features? Lucy Chambers.

Robustness of script evidence in comparative judgement awarding activities. Joanna Williamson.

Equality of access to access arrangements and their impact on students' performance. Carmen Vidal Rodeiro.

Working definitions of error used within Cambridge Assessment. Nicky Rushton.

From flying a plane to creating exam papers: how the SHELLO model can help us minimise errors in assessment materials. Sylvia Vitello.

Comparing levels-only marking and comparative judgement. Tom Benton and Emma Walland.

Auto-marking of short free-text responses in science. Gareth Wadge, Tom Sutch and Nick Raikes.

More like Germany's? System and ideological tensions in the UK government attempt to make vocational education and training (VET) in England more like the German model. Tony Leech.

Using educational research evidence in an agile product development. Sarah Hughes.

Blogs

The following blogs have been published since *Research Matters*, Issue 31:

Bramley, T. (2021, March 31). Online assessment - the robustness and resilience of the exam system (part 1). https://www.cambridgeassessment.org.uk/blogs/the-robustness-and-resilience-of-the-exam-system-part-1/

Bramley, T. (2021, April 07). Continuous assessment - the robustness and resilience of the exam system (part 2). https://www.cambridgeassessment.org.uk/blogs/the-robustness-and-resilience-of-the-exam-system-part-2/

Coleman, V., Constantinou, F., Greatorex, J., & Mouthaan, M. (2021, 01 June). Is curriculum coherence a fundamental characteristic of high-performing education systems? https://www.cambridgeassessment.org.uk/blogs/curriculum-coherence/

Coleman, T., Constantinou, F., Greatorex, J., & Mouthaan, M. (2021, 08 July). How should cognitive science be used in developing teaching practice? https://www.cambridgeassessment.org.uk/blogs/cognitive-science/

Coleman, T. (2021, 22 July). Has Covid-19 highlighted a digital divide in UK education? https://www.cambridgeassessment.org.uk/blogs/has-covid-19-highlighted-a-digital-divide-in-uk-education/

Oates, T. (2021, May 13). Assessment - perhaps it's just about good questions... https://www.cambridgeassessment.org.uk/blogs/assessment-its-about-good-questions/

Oates, T. (2021, 09 June). Why should we be talking more about oracy? https://www.cambridgeassessment.org.uk/blogs/oracy/

Oates, T. (2021, 17 June). Education reform is more than a question of what. https://www.cambridgeassessment.org.uk/blogs/reform-timeframe/

Oates, T. (2021, 03 August). Here's how to solve the 'hyper problem' of interrupted learning. https://www.cambridge.org/news-and-insights/blogs/heres-how-to-solve-the-hyper-problem-of-interrupted-learning

Rushton, N. (2021, 09 September). The relationship between marking and grading. https://www.cambridgeassessment.org.uk/blogs/the-relationship-between-marking-and-grading/

Walland, E. (2021, 17 September). Understanding grading, standards and grade inflation in England. https://www.cambridge.org/news-and-insights/insights/grading-standards-blog2

Sharing our research

We aim to make our research as widely available as possible. Listed below are links to the places where you can find our research online:

Journal papers and book chapters: www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Research Matters (in full and as PDFs of individual articles): https://www. cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/ Conference papers: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

Research reports: www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Data Bytes: www.cambridgeassessment.org.uk/our-research/data-bytes

Statistics reports: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

Blogs: www.cambridgeassessment.org.uk/blogs/

Insights (a platform for sharing our views and research on the big education topics that impact assessment around the globe): https://www.cambridgeassessment.org.uk/insights/

Our YouTube channel: https://www.youtube.com/channel/UCNnk0pi7n4Amd_2afMUoKGw contains Research Bytes (short presentations and commentary based on recent conference presentations), our online live debates #CamEdLive, and podcasts.

You can also learn more about our recent activities from Facebook, Instagram, LinkedIn and Twitter.

Contents / Issue 32 / Autumn 2021

Cambridge University Press & Assessment
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

researchprogrammes@cambridgeassessment.org.uk
www.cambridge.org