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.

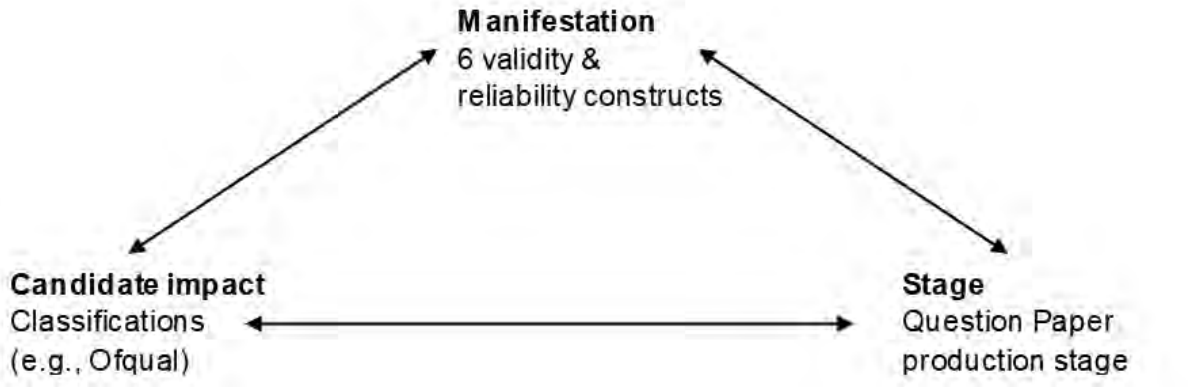

Figure 1: The interaction of manifestation, candidate impact and stage when defining question paper errors.

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.