International Journal of Assessment Tools in Education
2021, Vol. 8, No. 2, 310–325

https://doi.org/10.21449/ijate.897874

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Review Article**

Principles for Minimizing Errors in Examination Papers and Other Educational Assessment Instruments

Irenka Suto[1,*], **Jo Ireland**[1]

[1]Cambridge Assessment, UK

ARTICLE HISTORY

Received: Dec. 03, 2020
Revised: Mar. 11, 2020
Accepted: Mar. 15, 2021

Keywords:
Error,
Fairness,
Instrument design,
Examination design.

Abstract: Errors in examination papers and other assessment instruments can compromise fairness. For example, a history question containing an incorrect historical date could be impossible for students to answer. Incorrect instructions at the start of an examination could lead students to answer the wrong number of questions. As there is little research on this issue within the educational assessment community, we reviewed the literature on minimizing errors in other industries and domains, including aviation, energy, and medicine. We identified generalizable principles and applied them to our context of educational assessment. We argue that since assessment instrument construction is a complex system comprising numerous interacting components, a holistic approach to system improvement is required. Assessment instrument errors stem primarily from human failure. When human failure occurs, it is not good enough to suggest that 'to err is simply human'. Instead it is necessary to look deeper, evaluating the latent working conditions that underpin the efficacy of procedures, making the human failure more or less likely. Drawing from the aviation industry's ergonomic SHELLO model, we articulate and explore three of the most critical working conditions that relate to our context: (i) time pressure, (ii) workload and stress, and (iii) wider organizational culture, including good error data collection. We conclude with recommendations for best practice in minimizing errors in assessment instruments. A 'good' error culture should be promoted, which avoids blaming individuals. Errors should be acknowledged readily by all, and system owners should take a scientific approach to understanding and learning from them.

> "Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth."
> (Jules Verne, 1864)

1. INTRODUCTION

As well as motivating students to pursue their ambitions, fair assessments can build trust and confidence in education systems within society at large (Nisbet & Shaw, 2020). To date, much research on improving assessment instruments has focussed upon the key topics of validity and reliability. An additional but oft-overlooked aspect of fairness is the presence of errors in final

*CONTACT: Irenka Suto ✉ suto.i@cambridgeassessment.org.uk 🖃 Cambridge Assessment, The Triangle Building, Shaftesbury Road, Cambridge, CB2 8EA.

ISSN-e: 2148-7456 /© IJATE 2021

or near-final versions of assessment instruments[†] (Baranowski, 2006; Rhoades & Madaus, 2003; Rodriguez, 2015). For example, a simple typographical error could make an examination question unanswerable. A multiple-choice question could have two correct response options, confusing students, or even no correct response options. Some missing information could make a question harder to answer than intended. Faulty instructions at the start of an assessment could lead students to answer the wrong number of questions.

In many assessment contexts, due to high levels of professionalism, errors of this kind are an infrequent albeit longstanding problem. In England and Wales, for example, the vast majority of high-stakes examinations for secondary school students are error-free; the national regulator reported that in 2019, just 71 errors were identified across more than 6300 question papers, non-exam assessments and materials administered that year (Ofqual, 2019). However, occasionally errors are reported in the UK's national media, (for example, Richardson, 2017; Meredith, 2019), and as in South Korea and New Zealand (New Straits Times, 2015; BBC, 2017) some error incidents have led to public outcries. This is because even rare errors can have wide-ranging and unpredictable impacts on students. Their anxiety levels can be affected, as can time management, and therefore their general performance during the assessment. Ultimately, students' life chances can be damaged. It is clear that whether instruments are summative or formative, paper-based or computer-based, innovative or traditional, and whether they are created by teachers, teams within assessment organizations, national experts, or others, the assessment community should strive to make them free from errors.

In this paper, we argue that each assessment instrument error should be viewed not merely as the result of human error about which little can be done, but as a symptom of a deeper and more complex problem which spans the international assessment community. As Dekker (2002a) points out:

> "Although it is a forgiving stance to take, organizations that suggest that 'to err is simply human' may normalise error to the point where it is no longer interpreted as a sign of deeper trouble." (Dekker, 2002a, p. 145).

We take some of the first steps in understanding why errors occasionally occur in assessment instruments, and why the detection of errors can be slow despite the numerous checks included in most construction processes. We draw upon the wealth of research literature on error reduction that exists in complex sectors such as medicine, manufacturing, the nuclear industry, and aviation, extracting those principles that generalise across contexts. In recent decades, greater understanding of how and why errors occur in these domains has been credited with significant improvements in safety as well as quality, saving countless lives. There is a clear opportunity for educational assessment professionals to utilise this considerable body of knowledge too.

In general, there are three main strategies for addressing the problem of errors. First, make fewer errors in the first place. Secondly, detect more of the errors that do arise, and do so rapidly; that is, make fewer errors in detecting errors. Thirdly, improve methods of negating any undesirable consequences of errors. In our context, a critical limitation of the third approach is the impossibility of mitigating the impact of an error at the level of the individual. Whilst one student might be badly confused, upset and/or delayed by a particular error, another might not even notice it. Giving all students in a cohort full marks (or no marks) for a question containing

[†] It is the norm for errors to arise and to be corrected rapidly during the (typically iterative) early development phase of instrument creation, during which teams of professionals work to maximise the quality of items and instructions. This article focuses *not* on these 'early' errors, but on the small minority of errors that evade detection during revision and checking procedures, to reach or almost reach students.

an error is therefore too crude a remedy, and even the most sophisticated statistical methods cannot identify how particular individuals have been affected. We therefore focus on the first two of these strategies. Our overarching goal has been to identify some key principles for best practice in minimizing errors in assessment instruments.

2. COMPLEX SYSTEMS

Many systems through which errors arise and are detected (or not) are complex. Oates (2017) draws from Mitleton-Kelly (2003) to explain an important distinction between complex systems and complicated systems in education:

> "Complicated systems have many parts and many interactions, but give predictable outcomes. A chronograph is complicated, but gives a highly regulated and consistent output: a measurement of time. By contrast, complex systems possess a large number of interacting components, with outcomes which are not a simple function of the interaction of the parts." (Oates, 2017, p. 9)

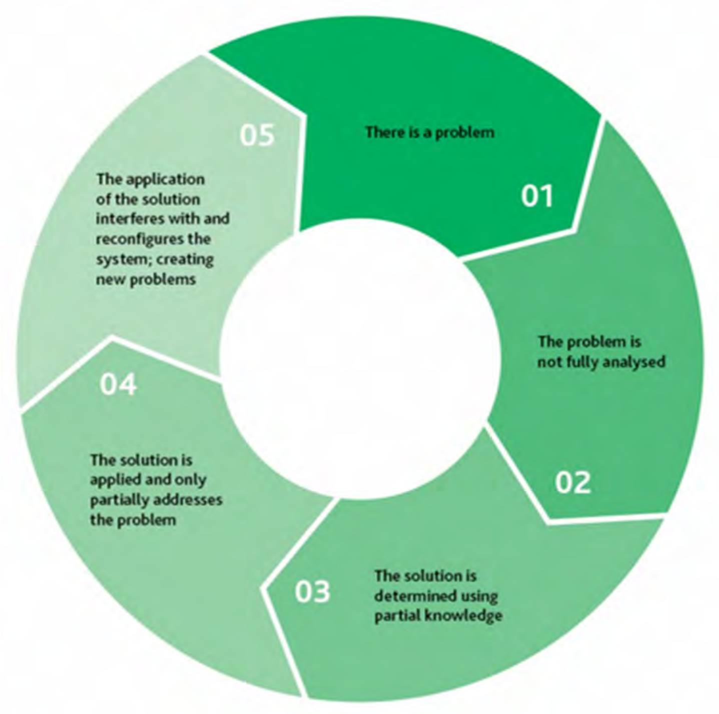
Oates (2017) goes on to argue that educational systems in all countries are complex and there is no single aspect of innovation which will secure a perfect system. Instead of cherry-picking initiatives from other contexts, a holistic approach to system improvement is required in which all of the components of the system are identified and included in its initial analysis.

To our knowledge, comprehensive analysis of systems in which assessment instruments are constructed and checked for errors has yet to happen. Instead, systems have evolved via the addition of extra checks, often in direct response to errors reaching students (for example, Harrison, 2011). This is partly because poor performance, when noticed, frequently calls for a rapid response. In the UK, assessment organizations must be seen by everyone, from students and teachers, to the national regulator, to the general public, to be doing something tangible and immediate to address the problem in the system. This is likely to be the case in other assessment contexts too. Also, it is usually easier to focus on one or two components of a system than to attempt its complete review, which risks the potential consequence of a complete overhaul being recommended.

One cumulative effect of multiple 'add-an-extra-check' initiatives is the diffusion of responsibility that each checker experiences. A second cumulative effect can be a cumbersome, costly, and overstretched construction process in which deadlines are hard to meet. When the process eventually approaches breaking point, consequent initiatives then reverse the direction of travel by focusing on streamlining and reducing the activities within the process. It is easy to envisage the overall state of affairs as a pendulum of change swinging slowly back and forth in response to external pressures.

In complex systems, this initiative-based approach to innovation and reform is risky. Without rigorous experimental trialling and evaluation (the gold standard of which is widely considered to be the Randomised Controlled Trial (RCT) approach used in medicine), it is impossible to conclude whether or not any particular action results in fewer errors. What is of greater concern, however, is that if action is taken prior to full analysis, then it can result in an inadequate response to the real causes of poor performance. Moreover, through its implementation, premature action can affect the system in unforeseen ways, creating new problems rather than remedying existing ones (Oates, 2017). To give a simple example, suppose an assessment organization relies upon outdated computer software which staff find laborious and unintuitive to use. Instigating yet another check of an already much-checked assessment instrument could lead to administrative overload for whoever is organizing the process and inadequate time to complete all checks, resulting in other errors. Achtenhagen (1994) developed an influential 'cycle of planned failure' to describe this kind of problem (Figure 1).

Figure 1. *Achtenhagen's (1994) cycle of planned failure, as reported in Oates (2017).*


05
There is a problem
01
The application of the solution interferes with and reconfigures the system; creating new problems
The problem is not fully analysed
04
The solution is applied and only partially addresses the problem
02
03
The solution is determined using partial knowledge

Arguably, the key to breaking this cycle of planned failure is a comprehensive analysis of the problem, which includes the identification of less visible components of the system. That is, it is crucial to consider the covert contributors to poor performance and not just the most obvious ones. Complete comprehensiveness is an enormous and elusive research ambition. However, the general approach of gaining a better overview of the system is one that we have sought to apply here to the problem of assessment instrument errors. We have begun to identify and articulate components and the more covert ones in particular, in an attempt to look more widely than has been done in previous efforts to minimize errors.

3. THE SWISS CHEESE MODEL

The pre-eminent name in research on the causes and detection of error is James Reason (e.g. 1990, 2013) whose theories have been applied in a range of hazardous industries including aviation, medicine and off-shore engineering. His work has led to considerable reductions in errors and their negative consequences and thereby to marked improvements in industrial safety. Analysing many contrasting disasters in the 1970s, 1980s, and 1990s, Reason identified three shared characteristics:

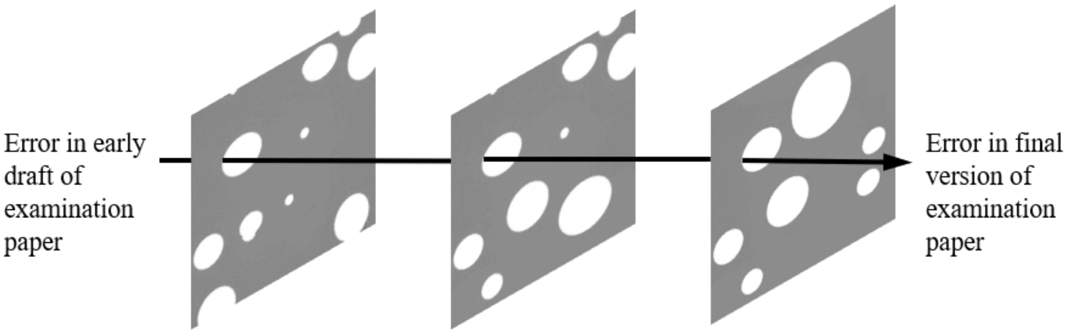
(1) Contributory factors which were present within the system prior to the occurrence of the disaster. All complex systems contain these 'resident pathogens'.

(2) Numerous defences, checks, and safeguards which were already in place within the system. These were designed to prevent known hazards from damaging people or assets.

(3) An unanticipated concatenation of human unsafe acts and local triggers, which defeated the numerous defences, creating a trajectory of opportunity for accidents to occur.

This analysis led Reason to create his most famous contribution to the field of error research: the Swiss Cheese model (Reason, 1990, discussed at length in Reason, 2008). In this model, which we have adapted to our context in Figure 2, the system defences that an organization or community puts in place are represented as slices of cheese. In an ideal world these defensive layers would be intact. In reality, however, they resemble Emmental cheese in having numerous ‘holes’. In contrast to holes in Swiss cheese however, ‘holes’ in systems are continually in flux, opening, closing, and moving around. The existence of holes in any particular defensive layer is not usually a problem. It is only when holes in successive layers align that a pathway of opportunity for disaster is created.

Figure 2. *Swiss Cheese Model (adapted from Reason, 1990).*


Error in early draft of examination paper
Error in final version of examination paper

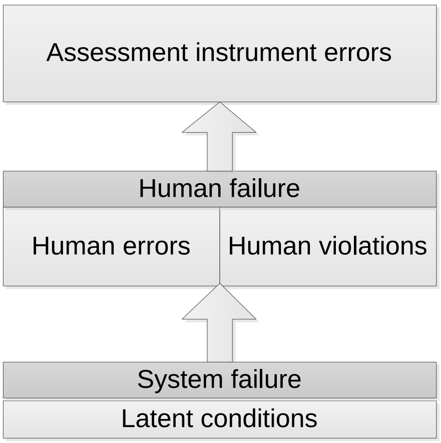
It is crucial to stress that in *any* of the components within an industrial system, there is the potential for practices to occur which engender holes in system defences. It is usually relatively easy to identify the concrete stages or activities within a production process, and then to focus investigations on the human errors (unintended) and procedural violations (intended) that could occur within them. These two kinds of human failure give rise to holes in the defensive ‘Swiss cheese’ layers that open and shut only briefly; their effects are short-lived.

Broadly speaking, within large assessment organizations, the production of examination papers (and other assessment instruments) entails an initial construction phase in which questions are drafted. This is followed by an iterative and often lengthy phase of reviewing and editing, in which questions are checked, re-checked, refined, and combined into examination papers. This phase may include pre-testing the questions with students, or ‘working’ the items or paper as a student proxy. In the next phase, questions and papers are then modified for students with particular needs or for a different mode of administration (for example, an on-screen version of a paper-based examination may be created). Finally, checks are made at a senior level prior to formal sign-off for printing and distribution to candidates. Human errors and violations of procedure could potentially occur during any of these activities, giving rise to errors in examination papers.

Unfortunately, the components of assessment instrument construction and other complex systems (such as aviation, e.g. Wiegmann & Shappell, 2003) are actually far more wide-ranging than this. Reason (2013) argues that system designers, builders, and managers, and procedure writers, inadvertently create ‘latent conditions’ (also called ‘resident pathogens’) which give rise to much larger and longer-lasting holes in the defensive layers. Latent conditions may lie dormant and undiscovered for years until one day they combine with human failures (errors and violations) and local triggers to create an accident trajectory.

We applied Reason’s work to our context to create a simple model of assessment instrument errors (Figure 3). That is, we adopted his theoretical position that system-level failure engenders human failure, which in turn gives rise to manifested errors (Reason, 2013) such as those that appear in assessment instruments.

Figure 3. *Model of assessment instrument errors.*



It follows that the question of what causes errors in assessment instruments can be addressed at two levels: at a psychological level of explanation, and at a system level of explanation. Battmann and Klumb (1993) and more recently Reason (2013) have explored the occurrence of violations, and there is an even richer psychological literature on when and why different types of human errors occur. Common explanatory psychological phenomena include inattentional blindness (Bruner & Postman, 1949; Aimola Davies et al., 2013), inadequate situational awareness (Endsley, 1995; Wickens, 2008), strong habit intrusions (Reason, 2013), and various limitations to working memory (Baddeley, 2010; Reason, 2013). These phenomena have been applied extensively to explain errors in industries such as aviation (Jones & Endsley, 1996), construction (Akinci, 2014), and medicine (Gawande, 2011; Pronovost & Vohr, 2011). In this paper, however, we focus on the latent working conditions that can contribute to system failure and underpin these human failures. These are often known as the root causes of errors. According to Reason (2013), whilst human failures take specific forms which can be hard to predict, latent conditions can be identified before a negative event takes place. A proactive form of system management is therefore needed, which entails regularly monitoring the system's vital signs.

4. THE SHELLO MODEL OF LATENT WORKING CONDITIONS

The aviation industry has made huge improvements to its safety record by identifying and addressing problems with latent working conditions within its systems. It has accepted for some time that human factors play a critical role in every aviation activity, from flight training to airline management (International Civil Aviation Organization, 1993). In a seminal paper, Edward (1972) argued that four types of interacting resources contribute to aviation accidents: software, hardware, environment, and liveware (people). He suggested that the source of every accident can be categorised as liveware, or as a combination of three major relationships: liveware-software, liveware-hardware, and liveware-environment. Edward's model is known as SHEL. It spawned a field of study described by Wiegmann and Shappell (2003) as the ergonomic perspective on human error because it emphasises human-machine-environment interactions. Over time it has been modified by multiple authors.

Chang and Wang (2010), for example, extended SHEL to become SHELLO. They identified the following factors as significant in accidents:

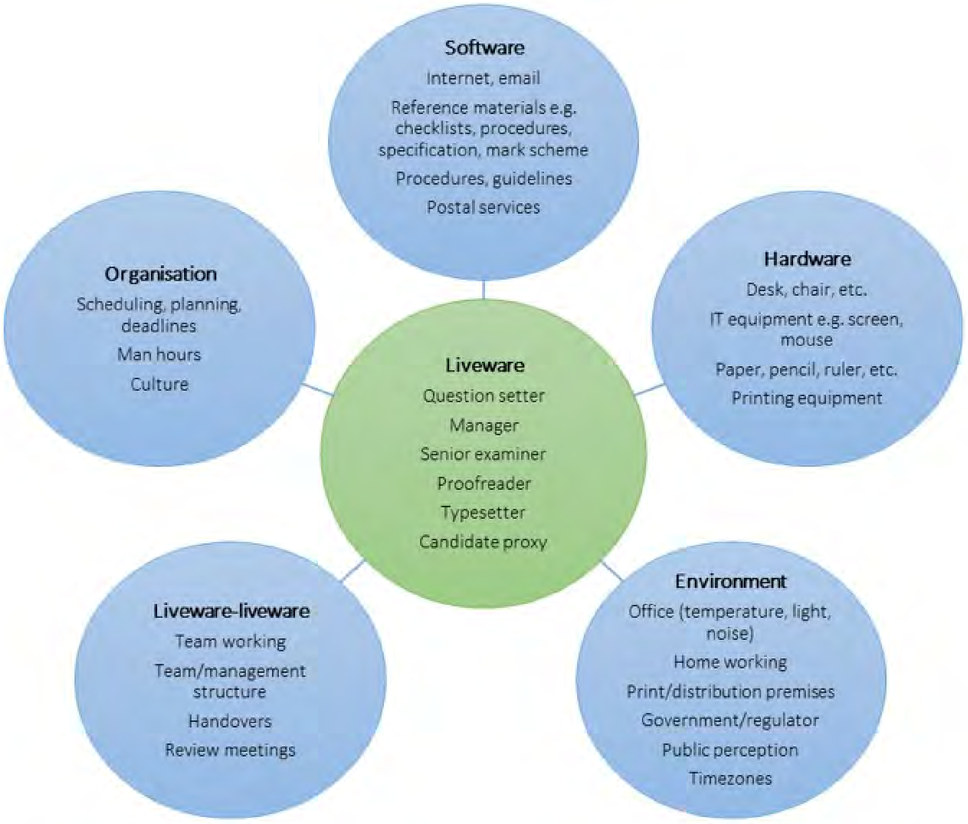
- **S**oftware (procedures, manuals, checklists)
- **H**ardware (tools, equipment, physical structure)
- **E**nvironment (physical environment, work patterns, management structures, public perception of industry)

- **L**iveware (people, managers)
- **L**iveware-liveware (person-to-person communication)
- **O**rganization (managerial model, decision-making patterns, culture).

Human operators feature in all interactions in this model (liveware to software, liveware to hardware, and so on) and carry risks of committing errors and violations. For those working within the system, the SHELLO model can contribute to an awareness of the context and the need for the factors to dovetail with one another to prevent breakdowns which might result in human errors. To use an example from air traffic control, the cause of an error might be cited as 'operator fatigue' which is contained within the liveware category. However, further investigation might show that the organization operated a culture of working long shifts, or that some aspect of the office environment had a part to play. SHELLO has been used successfully to develop numerous risk management strategies, for example, to help airline pilots to reduce runway excursions (Chang et al., 2016).

We used SHELLO as a basis for understanding the latent working conditions affecting educational assessment instrument construction. To do this, we populated its template with relevant factors (Figure 4). Taken as a broad suggestion of the factors which might be involved without being tied to any particular construction process, Figure 4 shows an interactively complex system.

Figure 4. *A SHELLO model of the factors affecting assessment instrument construction.*



Although the model is simple, it shows the potential impact that a system can have on error and how defences can be breached. Suppose, for example, that while undertaking a new check, a flickering light causes the checker to lose concentration and phone a janitor for help. On hold to the janitor and mindful of time, the checker continues with the check without paying full attention to the task, and makes a slip. This human error results in an error in the examination.

Does the examination error stem ultimately from problems with liveware, the environment, the organization, or all three? It is clear that both resources and working culture play an important role in determining the quality of the examination.

4. TIME PRESSURE

In the SHELLO model, time pressure lies within the 'organization' category of latent conditions. Although it would be easy to assume that time pressure always has a negative effect on task accuracy, the issue is more complex than a straightforward speed/accuracy trade-off. Drawing upon regulatory focus theory (Higgins, 1997), Förster et al. (2003) describe two types of goal pursuit among workers: *promotion* focus, and *prevention* focus (Table 1). In proof-reading and similar tasks, colleagues with a promotion focus adopt a risky processing style that is concerned with getting 'hits', that is, spotting lots of errors in the text quickly. Colleagues with a prevention focus, on the other hand, adopt a more careful processing style. They are concerned with avoiding making errors in spotting errors in the text. The focus that someone adopts can present at personality or task level, and can be chronic or momentary. For example, Förster et al. (2003) suggest that a promotion focus increases with a colleague's proximity to goal completion.

Table 1. *Types of goal pursuit.*

	Promotion focus	Prevention focus
Concern of colleague	Accomplishments and aspirations Gains and non-gains	Safety and responsibility Losses and non-losses
Behaviour of colleague	Strategic eagerness Risk taking	Strategic vigilance Risk averse
Task speed	High	Low

In one of the experiments described by Förster et al. (2003), two groups of participants were asked to complete a proofreading task as quickly and accurately as possible, identifying errors which had been deliberately created in a passage of text, within a fixed time period. One group was given a promotion focus: they would receive more money for a good speed/accuracy score. The other group was given a prevention focus: they would lose money if they didn't achieve a high enough score.

The promotion focus was found to enable faster proofreading and the identification of more of the easy-to-spot errors compared with the prevention focus. In contrast, the prevention focus led to higher accuracy in finding more difficult errors than the promotion focus did. Through speed and searching for easy errors, the promotion focus maximised proofreading performance overall, as measured by the total number of errors detected in the fixed time period. Förster et al. (2003) concluded that speed/accuracy trade-offs are a function of both regulatory focus and task difficulty. Whereas easy errors are found quickly with a promotion focus which enhances speed, difficult errors are more accurately found with a prevention focus.

The findings indicate that system designers and managers should think carefully about whether to encourage checkers to adopt a promotion focus or a prevention focus, through remuneration and performance management strategies, for example. They should also use caution when making decisions about how much time is allocated to tasks. Scrutiny of the data on the types of error arising in the assessment instrument construction process may indicate whether a particular focus is likely to yield better outcomes at a particular stage of the process. It may be that both focuses could be used in successive checks of an instrument, to improve the detection of both easy-to-detect and difficult-to-detect errors. It is worth noting that the goal of proof-reading and similar checks is usually to detect *every* error, and a prevention focus seems particularly appropriate. Given that a prevention focus requires plenty of task time, however, encouraging a prevention focus in combination with rapid working to tight deadlines could be a recipe for disaster.

5. WORKLOAD AND STRESS

Time pressure is usually linked to workload. When both are *too* high, colleagues become stressed. According to Dekker (2002a), tunnelling and regression are both reactions to stressful situations which can result in a loss of situational awareness. Tunnelling describes a fixation on an increasingly narrow portion of one's operating environment (Dekker, 2002a). There are many well-recounted examples, particularly in aviation. For example, a flight crew becomes distracted by an anomaly in the cockpit and fails to notice a loss of altitude. Other examples come from medicine: a doctor becomes so focussed on attempting and re-attempting a difficult surgical procedure that she fails to notice time passing and the patient's condition deteriorating (Syed, 2015). Regression occurs when actors revert to previously learned routines and fail to notice the critical differences of the current situation. Tunnelling and regression could have implications for the assessment instrument construction process if, for example, a question writer has in mind a previous oversight and consequently focuses on that aspect of the task while failing to notice other potential errors.

Dekker (2002a) argued that one way to think about workload and stress is to identify the type of demand-resource mismatch. The problem may not always be time pressure. For example, when working to a tight schedule, the coping resource that a particular colleague requires might be professional skills in workload management. If these skills are insufficiently developed then this could contribute to human errors or violations. For a different colleague working to exactly the same schedule, the problem may be slow computer software. If this resource is inadequate, then the colleague may become stressed and make errors despite excellent workload management skills.

Other related factors shaping the impact of workload can include multiple, competing goals, and not only in terms of the balance across elements such as minimizing costs, maximising accuracy, and adhering to deadlines. Colleagues with different roles within the same team working towards a common, larger goal may not feel pressure from, or responsibility for the same smaller goals set in order to reach the larger one (Dekker, 2002a). There may also be mismatches between team members' knowledge and/or an assumption that others possess the same knowledge, which can result in a lack of coordination.

6. THE WIDER ORGANIZATIONAL CULTURE SURROUNDING ERRORS

Over the past three decades, leaders in industries ranging from transportation and aviation to off-shore energy and nuclear power have identified serious issues with their culture surrounding errors. Their acknowledgements of problems, coupled with concerted efforts to rectify them, have been credited widely for significant reductions in major incidents and for higher safety and quality standards in general. This has not been the case in every workplace, however. The idea of organizational culture can seem so vague and elusive that some senior managers simply pay lip service to it by sending their staff on generic courses, or by checking periodically that they have a formal procedure in place for everything.

In the context of assessment instrument construction, failure to engage with this issue deeply would be a huge oversight. It is worth thinking about what a poor culture surrounding errors looks like in practice, and contrasting it with a good culture. This is what Matthew Syed has done for other industries (Syed, 2015). The principles that he draws together can be applied to assessment instrument construction. The essence of Syed's argument is that a good culture is one in which we readily acknowledge errors and take a highly systematic and scientific approach to understanding and learning from them.

7. GOOD DATA COLLECTION

First, the systematic collection of good data on errors and violations is at the heart of a good organizational culture. It is important to understand what types of errors and violations are occurring, and where in the system they arise. In order to spot recurring or 'signature' errors (those with a subtle pattern) the data needs to be detailed and comprehensive. Syed (2015) offers criminal justice as an example of a system in which detailed error data is *not* usually collected, although wrongful convictions have long been established unequivocally (Borchard, 2013). As in many other countries, the jury system in England and Wales is a secretive one. As jurors' activities (and errors) cannot be scrutinised, there is no opportunity to learn from them.

8. COGNITIVE DISSONANCE

Moreover, Syed (2015) argues that rather than admitting to failure and using it as a learning opportunity, members of many police forces and prosecution services experience 'cognitive dissonance'. This is the inner tension we feel when our beliefs are contradicted by evidence (Festinger, 1957). We do not like to perceive ourselves as irrational or foolish, or to have wasted a lot of time pursuing a cause in vain, as it threatens our self-esteem. Rather than accept that our original judgements were faulty, denial is a more comfortable option psychologically. It is much easier to reframe the evidence, spin it, filter it, or ignore it altogether. If anything we tend to become even more entrenched in our beliefs.

There are many examples of police officers and legal prosecutors refusing to accept DNA exonerations in cases they worked hard on (Innocence Project, 2021). Another example of cognitive dissonance is of people standing by decisions to abandon their families and possessions to join cults, even after prophecies of the world's end have turned out to be wrong. A typical post-prophecy argument would be that the cult leader and his/her followers are praying so hard and behaving so well that God has shown mercy, which is an even stronger reason to stick with the cult.

Cognitive dissonance could play a part in assessment instrument construction when managers (who may be conscientious workers with considerable expertise) face the discomfort of an error passing undetected through a carefully designed system. They may try to resolve conflicting beliefs by placing the blame elsewhere (e.g., a checker failed in their role) or they may argue that despite everyone's best intentions, such failures cannot be avoided. According to Dekker (2002b), when failure results in cognitive dissonance, it is usually easiest to place the blame on an individual. He suggests: 'Faced with a bad, surprising event, people seem more willing to change the individuals in the event, along with their reputations, rather than amend their basic beliefs about the system that made the event possible' (Dekker, 2002b). Reverting to the 'bad apple' theory provides (unfounded) reassurance that the system is essentially safe and errors arise from unpredictable humans working within the system.

9. AVOIDING A BLAME CULTURE

According to Reason (2013) when an accident occurs, the key question is not who blundered, but how and why the system defences failed. Enquiries into mishaps frequently reveal errors and violations committed by those at the coalface. At this point it is easy for senior managers to conclude that 'to err is simply human' and that all processes worked as they were designed to. As discussed previously, however, the crucial next step is then to investigate the workplace factors (latent conditions) contributing to these errors and violations. These will be factors in the SHELLO model, such as work pressure, inadequate training or briefing, under-staffing, inappropriate tools and equipment, and so on. These provocative factors are probably the consequence of decisions made by senior management. Reason (2013) points out that such decisions may turn out to be mistaken, but not necessarily so. Almost all high-level decisions simultaneously have positive consequences for some colleagues and negative consequences for others elsewhere in the system. It is rarely possible to please/help everyone all of the time.

As causal factors in the workplace are systemic in nature, blame at the individual level is unhelpful. Some form of 'no blame' culture in which colleagues feel able to report every failure, however big or small, is crucial to obtaining comprehensive error data for analysis and system improvement. However, Dekker (2017) argues that many organizations today have a retributive just culture instead. This approach asks: which rule is broken? Who did it? How bad was the breach and what should the consequences be? Who gets to decide this? Where staff are penalised for every error through performance management strategies, remuneration or public shaming, they will be unwilling to own up to slips, mistakes and violations. They will be more likely to hide them, risking problems further down the line. Moreover, at the level of process development, the managers responsible for them will be prone to cognitive dissonance. That is, they will find endless justifications and work-arounds for the decisions that they have made and the procedures that they have implemented. According to Dekker (2017), retributive justice rarely promotes honesty, openness, learning and prevention. Instead, he argues for a culture of restorative justice. This approach asks: Who is affected? What do they need? Whose obligation is it to meet that need? How do you involve the community in this conversation? Edmundson (1999) describes the 'psychological safety' that is needed for restorative justice to take hold within teams.

10. A SCIENTIFIC APPROACH TO UTILISING ERRORS

In contrast to the criminal justice system, the aviation industry has a very positive culture surrounding error. Syed (2015) termed this 'black box' thinking, after the indestructible box with which every aircraft is equipped. During a flight the box records all instructions sent to the onboard electronic systems, as well as the conversations and sounds in the cockpit. When an accident occurs, the data in the box is analysed and the causes of the accident are identified. Rather than concealing, ignoring or stigmatising failure, aviation culture treats every incident as a data rich learning opportunity. Independent investigators are given carte blanche to interrogate all the data. Since any information provided by interested parties is inadmissible in court, their openness and full disclosure is probable. Afterwards, the report is made available to the public and airlines are legally obliged to implement the recommendations. As everyone can access the data, everyone can learn from the errors. Procedures can then be improved, to avoid any repeat of the accident.

Learning from failure is also at the heart of the modern scientific method. The philosopher Karl Popper (1963) argued that science advances through its vigilant response to its own errors. Scientific theories make predictions that can always be tested and this is a huge strength. Unlike in astrology or psychoanalysis, hypotheses are made which can be refuted definitively. When this happens, new ones are developed, the field of enquiry progresses and our body of scientific knowledge grows.

As mentioned previously, the gold standard in scientific method, at least in some circumstances, is arguably the RCT. It has revolutionised pharmacology, for example. Without RCTs, there is a risk that closed loop thinking is perpetuated through skewed interpretations of evidence. That is, those who feel they have benefited from a new treatment may be highly vocal about it whilst those who did not benefit may slip under the evaluative radar. This leads to a false perception of efficacy, potentially perpetuating the use of the treatment on very shaky grounds. Although RCTs are widely used to test the efficacy of medical interventions, they have proven equally successful in other contexts. Syed (2015) cites examples from large-scale manufacturing to British Cycling and the Olympic Team GB, where the performances of both products and people have been optimised. A systematic ‘trial and error’ approach is often taken until success is achieved.

The success of this approach relies upon rapid feedback on the outcome of each trial and a cultural willingness to try again and again, using the feedback to learn about what works and what does not work. To minimize errors in assessment instrument construction, RCTs could in theory be used repeatedly in research using past instruments to establish the relative efficacies of different checks. They could also be used to investigate the skillsets needed to perform particular checks, and aspects of procedure (e.g. time) needed to develop instruments with minimal errors. Such a systematic approach to error would reduce weaknesses within systems which would otherwise persist due a reliance on unjustifiable assumptions that current procedures are optimal.

Although system improvers frequently seek fast one-off elixirs, the ‘slowly but surely’ approach outlined above actually embodies the theory of marginal gains. This is the idea that lots of small improvements add up to a large improvement, so it is worth making each small improvement. Because the search for marginal gains takes time, it should ideally be part of an organization’s usual activity. Given the infrequency of assessment instrument errors, however, it must be recognised that the resource involved in employing RCTs in this search would be huge. The gains to validity that might be achieved by devoting this resource to other areas might actually be larger, and cost-benefit analyses would undoubtedly be needed prior to embarking on this approach. A more realistic and cost-effective approach to minimizing errors might instead be to focus upon the considerable insights that can be obtained through detailed analyses of routinely collected error data, in error logs, for example. Such analyses might ultimately lead to a smaller number of highly targeted RCTs which focus specifically upon the most persistent problems.

Kahneman (2011) stresses that organizations seeking to improve should routinely look for efficiency improvements, and the operative concept is routine. He argues that expertise develops through a growth mindset and continual learning at the organizational level as well as at an individual level. Similarly, Weick et al. (1999) claim that the power of a safe culture lies in instilling an ongoing ‘collective mindfulness’ of the many entities that can compromise a system’s safeguards. Reason (2013) suggests that if there is a phrase that captures the essence of an unsafe culture, it is unwarranted insouciance. His epitaph for a lot of culture-induced organizational accidents would be: ‘There was always something more pressing to do.’ Of course, this approach costs time and money, and ultimately that must be weighed against of costs of serious errors occurring.

11. WORKFORCE ATTITUDES AND INTERPERSONAL SKILLS

A good organizational culture relies heavily upon the attitudes and interpersonal skills engendered in the workforce. Working in the fields of aviation and air traffic control, Kontogiannis and Malakis (2009) identified multiple attitudinal factors which play a part in error detection. These include: vigilance and alertness (including being able to ‘make the familiar strange’); suspicion and curiosity; awareness of vulnerability to errors, and awareness of degradation and disengagement, for example through distraction, fatigue or illness. Being able to cope with frustration from errors was also considered important. The authors also identified relevant team factors. These include: assertiveness, for example, feeling able to question decisions of senior colleagues and those senior staff being open to challenges; the abilities to cross-check and monitor others; the ability to adopt multiple perspectives; and strong communication of intent.

In aviation, these factors are developed in staff through Crew Resource Management (CRM) training. CRM evolved in response to evidence that many aviation accidents did not originate from aircraft technical issues or the crew's lack of knowledge, but from the responses of the crew to the situation in which they found themselves. CRM training aims to develop cognitive and social skills in support of technical training (Civil Aviation Authority, 2014). It is a mandatory component of commercial aircrew training in most countries. Such is the success of CRM that the format has been adapted for use by other industries, such as medicine, nuclear power and the offshore oil industry (Bleetman et al., 2012; Flin et al., 2002). The training content varies between industries, but generally includes teamwork, situation awareness, risk assessment, decision making, communication and workload management.

CRM can play a key role in mitigating cognitive dissonance as a cause of errors. For example, suppose a senior manager asks a checker to carry out an additional check of an assessment instrument. The checker believes that the procedure could increase time pressure further down the chain of checks, but experiences cognitive dissonance because he or she simultaneously believes that (i) the proposed check is wrong and shouldn't be carried out, and (ii) that the manager holds seniority over them and is right. Through CRM training, the checker should feel confident enough to voice concerns to the senior colleague. Equally, having also had CRM training, the senior manager should welcome a query without seeing it as a challenge to authority.

12. CONCLUSION

In this paper we have considered how errors arise in different industries, primarily at a system level of explanation, identifying generalisable principles and applying them to our context of educational assessment instrument construction. The literature we have reviewed is just the tip of the iceberg, given that in some industries, whole books and entire journals have been dedicated to some of the issues explored. It is clear that aviation, energy, and medicine take errors extremely seriously, and this is because errors compromise safety as well as quality; they can be literally a matter of life and death. Although the consequences of errors in educational assessment instruments are rarely so overtly catastrophic, they may nonetheless have life-changing consequences for students.

We have argued that since assessment instrument construction is a complex system comprising numerous interacting components, a holistic approach to system improvement is required. Cherry-picking initiatives from other contexts or introducing yet another examination paper check will not work. Within most assessment construction systems it is relatively easy to identify concrete activities and the human errors and violations that can occur when they are carried out. When human failure occurs, it is not good enough to explain it away by suggesting that all procedures worked as intended but that 'to err is simply human'. It is necessary to look deeper. That is, it is crucial to evaluate the latent working conditions that underpin the efficacy of the procedures, making the human failure more or less likely. This is how organizations in other industries successfully improve their performance.

Latent working conditions which can give rise to human failure and ultimately to errors in question papers are created unwittingly by system designers and procedure writers, and by senior management more generally. These conditions are wide-ranging. They relate to software, hardware, the working environment, the people involved, and organizational culture. Potentially affecting all latent conditions, a good organizational culture is one in which individuals are not blamed for their errors. Instead, errors are acknowledged readily by all but are not trivialised. Senior managers take a highly systematic and scientific approach to understanding and learning from them.

Drawing from these conclusions, we recommend five linked principles for best practice in minimizing errors in assessment instruments. First, a culture of restorative justice should be promoted, in which individuals are not blamed or penalised for errors. This approach asks: Who is affected? What do they need? Whose obligation is it to meet that need? How do you involve the community in this conversation? Secondly, coupled with psychological safety, this will make it possible to collect truly comprehensive data on errors, including data on the latent conditions that engender errors. This will in turn make it possible to identify recurrent 'signature' errors and their potential causes.

Thirdly, there is a need to instil in all authors and checkers of instruments an ongoing collective mindfulness of the many entities that can compromise the system's safeguards. It should be part of routine activity to investigate as many errors as possible - ideally all – to gain an understanding of why things went wrong. Fourthly, there is a need to hypothesise potential solutions to problems and test them scientifically. As a starting point, an RCT approach could be used determine the relative efficacies of different types of checks, using 'seeded' errors in past assessment instruments. Fifthly, this approach of learning from errors needed to be embedded into organizational culture at all levels of staff, so that the necessary resource is made available.

Finally, it is worth noting that the implications of the literature we have reviewed can be extended well beyond our stated context of assessment instrument construction. Awarding organizations and others involved in test construction produce numerous other types of document, including syllabuses, procedural manuals, reports for the regulator, legal contracts, research papers, and so on. All of these documents are prone to errors and the consequences can be serious. Arguably, there is nothing to stop anyone in the educational assessment community from adopting the mindset and approach to improvement advocated here in these related contexts.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship contribution statement

Dr Irenka Suto is a Principal Research Officer. Prior to joining Cambridge Assessment in 2005, she studied at the University of Cambridge and conducted post-doctoral research into financial decision-making processes. She has a long-standing interest in the many human judgements and decisions entailed in educational assessment, as made by students, teachers, examiners and administrators.

Jo Ireland is a Research Officer at Cambridge Assessment. Her research focuses mainly on the comparability and validity of assessments, and has included the development and application of tools to analyse the cognitive demand of examination questions.

ORCID

Irenka Suto https://orcid.org/0000-0001-6871-901X
Jo Ireland https://orcid.org/0000-0003-1237-7860

13. REFERENCES

Achtenhagen, F. (1994, June). Presentation to Third International Conference of Learning at Work, Milan, Italy.

Akinci, B. (2014). Situational Awareness in Construction and Facility Management. *Frontiers of Engineering Management*, *1*(3), 283-289. https://doi.org/10.15302/J-FEM-2014037

Aimola Davies, A., Waterman, S., White, R. & Davies, M. (2013). When you fail to see what you were told to look for: Inattentional blindness and task instructions. *Consciousness and Cognition*, *22*(1), 221-230. https://doi.org/10.1016/j.concog.2012.11.015

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136-R140.

Baranowski, R. (2006). Item editing and editorial review. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 349-357). Lawrence Erlbaum Associates.

Battmann, W., & Klumb, P. (1993). Behavioural economics and compliance with safety regul ations. *Safety Science*, *16*, 35-46. https://www.sciencedirect.com/science/article/pii/092575359390005X

BBC (2017, November 21) *NZ minister orders probe into 'impossible' maths exam*. https://www.bbc.co.uk/news/blogs-news-from-elsewhere-42065574

Bleetman, A., Sanusi, S., Dale, T., & Brace, S. (2012). Human factors and error prevention in emergency medicine. *Emergency Medicine Journal*, *29*, 389-393. http://emj.bmj.com/content/29/5/389.long

Borchard, E. M. (2013). *Convicting the innocent and state indemnity for errors of criminal justice.* The Justice Institute. (Original work published 1932)

Bruner, J. S., & Postman, L. (1949). On the perception of incongruity: a paradigm. *Journal of Personality*, *18*(2), 206-23. https://doi.org/10.1111/j.1467-6494.1949.tb01241.x

Chang, Y.-H., & Wang, Y.-C. (2010). Significant human risk factors in aircraft maintenance technicians. *Safety Science*, *48*(1), 54-62. https://doi.org/10.1016/j.ssci.2009.05.004

Chang, Y.-H., Yang, H.-H., & Hsiao, Y.-J. (2016). Human risk factors associated with pilots in runway excursions. *Accident Analysis & Prevention*, *94*, 227-237. https://doi.org/10.1016/j.aap.2016.06.007

Civil Aviation Authority (2014). Flight-crew human factors handbook. CAA. http://publicapps.caa.co.uk/docs/33/CAP%20737%20DEC16.pdf

Dekker, S. (2002a). *The Field Guide to Human Error Investigations*. Ashgate.

Dekker, S. (2002b). Reconstructing human contributions to accidents: the new view on error and performance. *Journal of Safety Research*, *33*(3), 371-385. https://doi.org/10.1016/S0022-4375(02)00032-4

Dekker, S. (2017) *Just Culture: Restoring trust and accountability in your organization*. CRC Press.

Edmundson, A. (1999). Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly*, *44*(2), 350-383.

Edward, E. (1972). Man and machine: systems for safety. Proceedings of the British Airline Pilots Association Technical Symposium, London.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Flin, R., O'Connor, P., & Mearns, K. (2002). Crew resource management: improving team work in high reliability industries. *Team Performance Management: An International Journal*, *8*(3/4), 68-78. https://doi.org/10.1108/13527590210433366

Förster, J., Higgins, E. T., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organizational Behavior and Human Decision Processes*, *90*(1), 148-164. https://www.sciencedirect.com/science/article/pii/S0749597802005095

Gawande, A. (2011). *The checklist manifesto: How to get things right.* Metropolitan Books.

Harrison, A. (2011, June 9) Students hit by more exam errors. *BBC*. https://www.bbc.co.uk/news/education-13710868

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280-1300. https://pdfs.semanticscholar.org/6b64/5e0418ae70e82cc322dd6fbf0647ae2523e4.pdf

Innocence Project (2021). *The innocence project*. https://www.innocenceproject.org

International Civil Aviation Organization (1993). *Investigation of Human Factors in Accident s and Incidents. Human Factor Digest No.7*. https://skybrary.aero/bookshelf/books/2037.pdf

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, and Environmental Medicine, 67*(6), 507-512.

Kahneman, D. (2011). *Thinking, fast and slow*. Penguin Group.

Kontogiannis, T., & Malakis, S. (2009). A proactive approach to human error detection and identification in aviation and air traffic control. *Safety Science, 47*, 693-706.

Meredith, R. (2019, May 17) AS-level Economics exam error under investigation in NI. *BBC.* https://www.bbc.co.uk/news/uk-northern-ireland-48313904

Mitleton-Kelly, E. (2003). Ten principles of complexity and enabling infrastructures. In E. Mitleton-Kelly (Ed.), *Complex systems and evolutionary perspectives on organisations: the application of complexity theory to organisations*. Elsevier.

New Straits Times (2015) *S. Korea exam chief resigns over errors in high-stakes college test.* https://www.nst.com.my/news/2015/09/s-korea-exam-chief-resignsover-errors-high-stakes-college-test

Nisbet, I., & Shaw, S. (2020). *Is Assessment Fair?* Sage.

Oates, T. (2017). A Cambridge Approach to improving education. *Cambridge Assessment.* http://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-improving-education.pdf

Ofqual (2019). GCSE, AS & A level summer report 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/852440/GQ-Summer-Report-2019-MON1100.pdf

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge.* Routledge and Kegan Paul.

Pronovost, P., & Vohr, E. (2011). *Safe Patients, Smart Hospitals: How One Doctor's Checklist Can Help Us Change Health Care from the Inside Out*. Penguin books.

Reason, J. (1990). *Human error*. Cambridge University Press.

Reason, J. (2008). *The human contribution*. Ashgate.

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

Richardson, H. (2017, May 26) GCSE exam error: Board accidentally rewrites Shakespeare. *BBC.* https://www.bbc.co.uk/news/education-40059967

Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem*. Boston College.

Rodriguez, M. (2015). Selected-response item development. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 259-273). Routledge.

Syed, M. (2015). *Black box thinking. Marginal gains and the secrets of high performance*. John Murray.

Verne, J. (1996). *Journey to the centre of the earth.* Wordsworth Editions Limited. (Original work published 1864.)

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (1999). Organising for high reliability: Processes of collective mindfulness. In R. S. Sutton & B. M. Staw (Eds.). *Research In Organizational Behavior*, *21*, 23-81.

Wickens, C. D. (2008). Situation awareness: Review of Mica Endsley's 1995 articles on situation awareness theory and measurement. *Human Factors*, *50*(3), 397-403. https://journals.sagepub.com/doi/pdf/10.1518/001872008X288420

Wiegmann, D. A., & Shappell, S. A. (2003). *A Human Error Approach to Aviation Accident Analysis*. Ashgate.