Language Teaching (2023), 1–12
doi:10.1017/S0261444823000034

CAMBRIDGE UNIVERSITY PRESS

FIRST PERSON SINGULAR

The ethical turn in writing assessment: How far have we come, and where do we still need to go?

Martin East[1]* and David Slomp[2]

[1]University of Auckland, Auckland, New Zealand and [2]University of Lethbridge, Lethbridge, Canada
*Corresponding author. Email: m.east@auckland.ac.nz

(Received 18 January 2023; accepted 20 January 2023)

1. Introduction

Both of us were drawn into the writing assessment field initially through our lived experiences as schoolteachers. We worked in radically different contexts – Martin was head of a languages department and teacher of French and German in the late 1990s in the UK, and David was a Grade 12 teacher of Academic English in Alberta, Canada, at the turn of the twenty-first century. In both these contexts, the traditional direct test of writing – referred to, for example, as the 'timed impromptu writing test' (Weigle, 2002, p. 59) or the 'snapshot approach' (Hamp-Lyons & Kroll, 1997, p. 18) – featured significantly in our practices, albeit in very different ways. This form of writing assessment still holds considerable sway across the globe. For us, however, it provoked early questions and concerns around the consequential and ethical aspects of writing assessment.

Towards the end of 2017, we took on the co-editorship of *Assessing Writing*. One of our first acts was to convene a Special Issue to celebrate and critically reflect on the 25 years of scholarship in the field since the journal had been launched (Slomp & East, 2019). In a brief editorial, Martin wrote:

> At its heart, writing assessment is about exploring the most appropriate ways in which we can gather and report on evidence of individual language users' writing proficiency. Fundamental to these ways of collecting evidence are, first and foremost, our attempts to define a writing proficiency construct. Then comes a consideration of how we operationalise that construct in assessments that will give us construct valid and reliable evidence of proficiency. These three concepts – construct definition, validity and reliability – are foundational to arguments around what makes a good or useful assessment of writing proficiency. (East, 2019, p. 1)

The timed impromptu writing test is seen by many as fulfilling these arguably timeless guiding principles, by no means perfectly, but certainly sufficiently.

However, Martin went on to argue, 'as we have come to recognise the diversity of both our assessment contexts and the students who take assessments in these contexts, the question of fairness has become more apparent' (East, 2019, p. 1). In the face of a vast and disparate range of audiences, the fitness for purpose of the standardised writing test has come under scrutiny and this testing format comes up against questions around equity and bias. In our view, to fail to take these questions into consideration would be unethical.

For both of us, the ETHICAL dimension of writing assessment comes to the fore when we consider the CONSEQUENCES of the assessment on those who are arguably the primary stakeholders in the assessment endeavour – those who take the assessments. We argue that, as the field moves into the future, ethical drivers must become paramount in our understanding and operationalisation of good or effective writing assessment, and we consider the challenges that the field still faces when seen from an ethical standpoint.

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

Often when we think of ethics, the mantra of 'first do no harm' comes to mind. However, when we consider the damage that can be done through large-scale snapshot writing assessment, we would suggest that an equally important ethical principle to guide our work is the idea that we should 'make good trouble.'[1] In what follows, we seek to challenge and disrupt status quo thinking as we confront the consequential and ethical dimensions of writing assessment.

It is important to acknowledge at the outset that it is not our purpose to argue here for the dismantling of standardised testing (the industry is arguably way too monolithic and entrenched for that). It is, rather, to bring the ETHICAL dimension of this kind of writing assessment to the fore and to ask questions of this testing format.

To contextualise our concern for the ethical dimension, we begin with an overview of the fundamental issue at stake – valid, reliable, and fair writing assessment. We then outline our own journeys into writing assessment as a field of research and where *Assessing Writing* has fitted into the ethical debates. We go on to exemplify some emerging ethical tensions through two examples – technology and translanguaging. We conclude with an overview of what we argue needs to happen as the field moves forward.

2. Valid, reliable, and fair writing assessment – a fundamental issue

Valid and reliable assessment of language users' writing proficiency is a central issue of concern among a range of stakeholders. We acknowledge that timed writing tests have been pervasive in the field precisely because, at least within an orientation that may be labelled 'technocentric' (Huot, 2002), they have generally been regarded as practical and standardised measures of test takers' writing proficiency that are both valid (e.g., Beigman Klebanov et al., 2019) and reliable (e.g., East, 2009). From this perspective, it has also been argued that this form of assessment is fair, and consequently ethical, because the timed assessment of writing is seen as offering 'a uniform context for examinees' performance … so that examinees all start from a basis of equal opportunity … [and the] constructs to be assessed cannot be biased for or against any particular population' (Cumming, 2002, p. 74).

Indeed, when we consider the origins of these forms of assessment in imperial China (dating back several millennia) and the UK and US (beginning in the nineteenth century), it is apparent that they were motivated by a desire for equity and for the promotion of opportunity (Elliot, 2005; Hamp-Lyons, 2002). The underlying rationale for their design, implementation, and use was essentially to ensure that positions of privilege and advancement were awarded based on ability and merit, not on social standing or social connection. This ethic depends principally on validity and reliability, but also includes considerations of fairness.

Differentiating between test takers on the basis of ability and merit may be considered all very well, and indeed ethical, if 'the primary function performed by tests is that of a request for information about the test taker's language ability' (Bachman, 1990, p. 321). We recognise that there will always be those who are more proficient or adept in a particular skill than others. It is not necessarily unethical to discriminate between different levels of test taker proficiency if the basis of our evaluation is valid and reliable.

On the other hand, Bachman (1990) argued that tests should provide 'the greatest opportunity for test takers to exhibit their "best" performance' so that they are 'better and fairer measures of the language abilities of interest' (p. 156). We contend that we owe it to those being assessed to ensure that any test or assessment of writing proficiency is not only valid and reliable as a measure of language ability, but also fair and equitable as a means of capturing 'best performances' – that is, performances that are not unduly influenced by facets of the assessment procedure that potentially bias the test against certain test takers through no fault of their own. From this 'humanistic' perspective, it would be unethical not to consider the issues arising from a test format that might hinder some test takers from displaying in the test what they know and can do.

[1]The former expression is attributed to the ancient Greek physician Hippocrates and (falsely) to the Hippocratic Oath, and the latter to the late US Civic Rights Leader and Congressman John Lewis.

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

The beginning of an ethical stance is to consider the human dimension, and the impact of timed impromptu writing assessments on individuals and populations. Shohamy (2001) highlighted this when she spoke of 'winners and losers, successes and failures, rejections and acceptances' (p. 113). Thus, performing at one's best becomes important (or 'high-stakes') because, as East (2016) put it, 'the results have a wide range of consequences for those who are required to take the assessments … [and] are frequently the only indicators used to make … [m]any important and far-reaching decisions' (p. 9).

When we take the people at the receiving end of the assessment into account, we maintain that timed impromptu assessment designs under-represent key aspects of the construct associated with the skills needed to compose text across the varied settings and circumstances in which real world writing is occurring. Portfolio assessments of writing were proposed as one means of addressing the limitations of timed tests because they enabled those being assessed to complete their writing in more authentic conditions, providing multiple samples of writing across a range of contexts and genres, and having opportunities to receive and incorporate feedback on the writing. Unfortunately, portfolio assessments, while useful as a form of classroom assessment, have rarely been used in large-scale assessment programmes, since the complexity and cost of managing such systems have been difficult to overcome in large-scale settings.

We are left with the reality that, currently, much of the large-scale assessment of writing continues, to this day, to rely on timed impromptu models of assessment. This being the case, we argue that it is incumbent on all of us with a stake in the effective assessment of writing to scrutinise the one-time snapshot assessment method from a consequential/ethical perspective. A concern to balance issues of validity and reliability with the consequential dimensions of the test taking experience has exercised both of us over the years, underpinned and driven by our own early experiences of working with school-aged students who were learning to write, and our own subsequent research.

3. Critical perspectives from the practitioner's desk

3.1 Martin's experiences

Martin's interest in writing assessment emerged over 25 years ago during a time of significant assessment reform in the UK, when the General Certificate of Secondary Education (or GCSE), the first high-stakes examination taken by students of 16+ years of age, was substantially modified. Martin was responsible for preparing students of French and German as additional languages for assessments of their developing language proficiency, including their emerging writing skills. For language assessment, the reforms meant, for example, that a traditional end-of-year summative writing test was complemented by the option, instead, to collect ongoing evidence of writing as coursework. This inevitably included writers' access to a range of resources to support them with their writing. As part of these reforms, bilingual dictionaries were also for the first time allowed into the summative writing test that was available instead of the coursework option. Dictionaries were also allowed in the summative reading and listening elements of the assessment, and in the preparation part of the speaking test.

Seen from a Communicative Language Teaching (CLT) perspective, the introduction of dictionaries was an attempt to bring a greater level of authenticity to the assessments. After all, within a communicative paradigm 'dictionaries may be seen as legitimate "tools of the trade"' because they are 'frequently used for purposes of communication, discourse creation, and negotiation of meaning in real-world contexts' (East, 2008, p. 14). Why not therefore admit them into assessments? Indeed, an important component of one influential theoretical model of communicative competence (Canale & Swain, 1980) was strategic competence, or the strategies used to 'compensate for breakdowns in communication due to insufficient competence or to performance limitations and … enhance the rhetorical effect of utterances' (Canale, 1983, p. 339). Dictionary use arguably had a role to play as one of a range of strategies that students should be equipped to utilise.

Writing with the use of a dictionary changed the landscape for writing assessment significantly. That is, allowing dictionaries in writing tests necessitates a rethink of what the construct of writing

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

proficiency actually is. From a narrow construct perspective, allowing dictionaries leads to construct irrelevant variances in performances, potentially making the test too easy or, conversely, distorting the evidence available from the writing sample. Within a wider construct definition where, for example, compensations for breakdowns in communication or enhancing rhetorical effect are seen as components of the construct of interest, it is potentially unfair to disallow dictionaries because the construct would then be under-represented.

From the teacher perspective, Martin watched his students struggle, not only with time to complete the tests they took, but also with inadequately developed skills in using dictionaries effectively, despite best efforts to prepare them. Dictionary use was perhaps potentially unfairly hindering students from displaying in the tests what they knew and could do. Martin began to question whether the dictionary WAS having any beneficial impact, or, conversely, whether its use was doing more harm than good. This questioning led to an important series of studies into dictionary use in writing examinations (see, e.g., East, 2008). It also raised a range of ethical questions. As Hamp-Lyons (2000) put it, fairness is a difficult concept because 'there is no one standpoint from which a test can be viewed as "fair" or "not fair"' (p. 32).

3.2 David's experiences

Unlike Martin, who was drawn into the field because of the challenges and questions raised by innovations in assessment design and use in the additional language context, David was drawn into the field because of stagnation and tradition. Working as a classroom teacher of Grade 12 students of Academic English (17+ years of age) in Alberta, Canada, David was confronted with the challenge of preparing his students for a government-mandated timed impromptu writing test that had seen little change or innovation since it was first implemented in the 1980s.

David's goal for his students was that they would develop the knowledge, skills, and dispositions needed to thrive in their lives as writers beyond the classroom. He discovered quickly, however, that the qualities that made students successful on a high-stakes, timed test were quite different from those needed to thrive as writers in their lives outside of school. For example, timed examination-based writing did not allow for, value, or measure students' ability to collaborate with others, engage in substantive revision, or create highly polished text. Because the tests had been so stagnant for so long, and because the political and public discourse in Alberta cultivated the idea that standardised examination scores were more trustworthy than grades assigned by classroom teachers, a culture of teaching to the test had become entrenched, with negative washback implications. Writing pedagogy, especially in the Grade 12 context, was built upon the narrow focus of helping students succeed in the examination – a problem that persists in Alberta to this day (Slomp et al., 2020).

David saw a clear disconnect between the assessment and the skills deemed necessary for successful real-world performance. He recognised that success in the examination without the development of a more robust set of skills, understandings, and dispositions would not serve his students well over the long-term. Research demonstrating the gap between student writing ability upon graduating from high school and expectations placed on students in post-secondary and workplace contexts reinforced for him the futility of limiting instruction to the narrow construct that the examination was measuring.

In his own classroom, David ended up teaching two writing programmes: writing for the examination and writing for life. He perceived that it would have been unethical not to do this. His subsequent research agenda focused on the consequences of assessment on writing pedagogy and critically examined the validity of standardised writing assessment designs through a detailed exploration of the effects of the Grade 12 examination on the teaching of writing in Alberta.

3.3 Consequences in real-world contexts

Essentially, from our perspective as teachers we saw first-hand the consequences of assessment design choices on the quality of information collected through writing assessments and on the validity of

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

inferences that could be drawn from assessment scores. We also saw the broader impact of these design choices on our own teaching and assessment practices, and on what our students were learning to value in their development as writers.

For Martin, a struggle emerged between dictionary use as a legitimate, real-world, and helpful resource for students and dictionary use as a hindrance to the writing process and the writing product. For David, the struggle was around a narrowly defined and narrowly measured construct and the need to broaden out the teaching and learning to embrace real-world writing needs. In both cases, the writing assessments had consequences, whether intended or unintended.

Shaped by our early experiences, our programmes of research began to take into consideration the CONSEQUENCES of writing assessment. This consequential dimension is a perspective that we bring to our work as co-editors-in-chief of *Assessing Writing* and they also frame our thinking about what is important for the field into the future.

4. From consequences to fairness, and from fairness to ethicality

Way back in 1994, scholars and practitioners in the field saw the launch of the journal *Assessing Writing*. Its inauguration came at a time of considerable experimentation in the wider field of educational assessment. Stakeholders in assessment were wrestling not only with a broader view of assessment that began to embrace a variety of assessment methods (such as classroom-based assessments, coursework, and portfolios, as well as a range of approaches, including more formative and performance-based assessment tasks), but also with the consequences of assessment design and use (see, e.g., Gipps, 1994; Gipps & Murphy, 1994). *Assessing Writing* became a crucial medium for 'negotiating and compromising among the interest groups involved' (White, 1994, p. 25).

Moss's (1994) contribution to the inaugural issue articulated the beginnings of a principled approach to considering the consequences stemming from the implementation and use of assessment programmes. She observed:

> Often in the past, policy makers have been too quick to implement assessment systems without adequate attention to the potential and actual consequences of their actions. … Those who implement assessment policy need to carefully evaluate both the quality of the information and the intended and unintended consequences of using assessments. Before any assessment is operationalized, policy makers should become informed about the existing research on the consequences of various assessment choices, compare alternative approaches to assessment in light of the differential consequences they might foster, and explicitly evaluate the vision of education implied in those consequences. After the system is implemented, policy-makers should hold themselves accountable through ongoing monitoring of the consequences of their actions … . Nothing should be taken for granted. (Moss, 1994, pp. 124–125)

A consideration of the CONSEQUENTIAL dimension of assessment was also largely being influenced by Messick's (1989) seminal work on validity. Messick admonished the broader field to identify intended, unintended, positive, and negative consequences stemming from an assessment's design and use. Consequential validity became an important component of construct validity. In practice, however, consequential concerns were often limited to a focus on score interpretation and use, and considerations of fairness in test design and use were somewhat embryonic. It seemed that, despite Moss's (1994) assertion, a lot was being taken for granted.

The ethical turn in writing assessment was exemplified in a 2016 special issue of the *Journal of Writing Assessment* (Kelly-Riley & Whithaus, 2016). It was observed, on the one hand, that in the assessment literature, the 'conversation of consequences' had shifted since the 1990s to 'articulations of fairness.' It was noted, on the other, that considerations of fairness were potentially still not sufficient or not sufficiently developed. Indeed, it was suggested that the next phase in the scholarship on writing assessment 'must take up larger questions of involving the ETHICS of assessments' (p. 1, our emphasis).

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

While these theoretical step-changes offer hope for developments in writing assessment design and use, it is very evident that old orientations continue to shape current research. In the Special Issue of *Assessing Writing* we referred to in the opening section of this article, Slomp (2019) asserted that, from 2004 through to 2018, the trend within manuscripts reporting validity evidence published in the journal has been for an increased focus on the scoring inference, and a decreased focus on construct representation. Paralleling this shift, the percentage of manuscripts publishing validity evidence related to the CONSEQUENCES of assessment has steadily declined from almost 60% in the first five years of publication to less than 20% between 2014 and 2018. In the same issue, Poe and Elliot (2019) examined trends in research on fairness published in the journal. They found that while the overall frequency of research on fairness has grown from 1994 to 2018, the primary concern in this body of scholarship is on the ABSENCE of fairness rather than on the broader implications associated with equity of opportunity to learn or to demonstrate learning. Indeed, Poe and Elliot observed that in the 25 years of scholarship covered by their review, only three articles addressed the concept of ethics and fairness. It seems we still have a substantial way to go.

5. The ethical turn in writing assessment

The goal of achieving a more just society through the implementation of valid, reliable, and fair assessments has proven more difficult to achieve in practice than it has in theory or principle. An assessment programme is, after all, one tool within a broader social system. In some cases, external factors such as cheating and corruption have undermined this ethic; in other cases, systemic inequity and its differential impact on opportunities to learn embed social standing in the very fabric of the assessment process; in still other cases, unexamined bias and hidden assumptions – woven into the very processes, structure, items, criteria, data analysis, and reporting that form an assessment programme – differentially impact populations of test-takers.

Furthermore, the two values of validity and reliability, and the conditions that contribute to them, often work at odds with one another (Slomp & Fuite, 2005; Wiggins, 1994). The design options used to achieve higher degrees of reliability (such as standardised one-time snapshot writing tests) tend to lead to reductions in construct representation and, consequently, validity. Conversely, as the target domains sampled in writing assessment broaden to capture as much of the construct as possible, associated design choices open up greater possibilities – not only for valid assessment (e.g., performance outcomes derived from portfolios, and opportunities for redrafting and responses to feedback), but also for random and systematic error to influence scores. This sets up a dynamic tension at the heart of writing assessment scholarship.

Historically, validity and reliability were dealt with as related but separate concerns. This led to situations where one value was played off against the other. More contemporary models of validity have begun to change this calculus. Validation models such as Kane's (2006, 2013) Interpretive Use Argument (IUA) and Bachman and Palmer's (2010) Assessment Use Argument (AUA) both incorporate concern for reliability directly within the chain of inferences needed to support a validity argument. The scoring, generalisation, and extrapolation inferences within an IUA, for example, require the collection of information related both to construct representation and to consistency and replicability of measures. Conceptualised in this way, validity arguments require assessment developers to find ways to maximise both areas of concern.

We contend, however, that if we ignore or sideline a concern for consequences, assessment design choices are simply options. Nonetheless, no programme of assessment is neutral; all programmes have an effect on individuals and the systems to which they are connected. Slomp (2016) argued for the need to take account of several facets of fairness. These include: the limitations of practices deriving from colonialism and capitalism; access to educational structures associated with literacy; opportunities to learn; disaggregation of data so that score interpretation and use can be clearly understood for all groups and for each individual within those groups; justice in the sense that each individual

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

has a fair chance to secure opportunities to succeed; and – arguably centrally – robust and maximum construct representation that is clearly articulated in advance of the assessment.

In this final section, we provide two examples – technology and translanguaging – where the timed impromptu writing test as currently operationalised seems to fail to take adequate account of developments in the ways in which writing may be undertaken in the real world and is therefore arguably missing some important ethical considerations. The increased use of computers for writing purposes creates a useful example of where ethics must arguably confront computer-mediated assessment practices. The issues raised by translanguaging create a useful example of where ethics must arguably confront established assessment practices.

6. Automating the writing process – an ethical challenge?

It is evident that technology is increasingly playing a dominant role when it comes to writing. The use of computers is ubiquitous, and increasing numbers of students across the globe are now regularly turning to computers, not only to seek out information through the internet, but also to craft a range of written outputs for a variety of audiences – from simple emails to complex coursework assignments. The parallel access to support resources and writing tools (such as grammar/spelling/plagiarism checkers and online translators/chatbots/artificial intelligence) has significantly altered the conditions for writing in the real world (see, e.g., Oh, 2022).

If the assessment of writing is to take into account the ways in which language users write via computers in real-world contexts, it must contend with, and factor in, the tools that lie at the disposal of writers. The challenge becomes the collection of evidence that represents a valid and reliable sample of a language user's writing ability. The International English Language Testing System (IELTS) has aimed to address the challenge by insisting that its writing examination, when taken on a computer, must replicate the conditions of the pen-and-paper test. However, this approach fails to take into account that where, for example, support resources are a part of what it means to write via computer, the construct measured by the test is potentially inadequate. Indeed, as the breadth and range of technology-mediated composition tools and processes continue to expand, the construct being measured in the computer-based IELTS test becomes increasingly under-represented, creating downstream negative consequences, both for test takers and for those who educate them.

From an ethical perspective, other problems emerge. What about language users who do not have adequate access to computers or the internet, or whose technological skills lag behind their contemporaries? In writing-by-computer assessment contexts, these students may be unfairly compromised in ways that may be irrelevant to the construct being measured (see, e.g., Guapacha, 2022, for an exploration of some of the tensions for computer-based versus pen-and-paper based writing tests in Colombia).

Problems also emerge as the field shifts toward the automated scoring of written text. From a technocentric perspective, automated scoring is an important advance for writing assessment because it enables greater consistency and reliability in scoring (as well as practicality). From a humanistic perspective, however, the concern continues to be that automated scoring strips away critically important elements of the writing construct in order to achieve that consistency. Such approaches to scoring measure factors such as length of text, syntactic complexity, and diction, but they do not and cannot, for example, measure the impact of a text on a human reader – the capacity of a text to evoke a response. More problematic still from a humanistic perspective is that algorithms cannot construct meaning in the way that humans do. Thus, scoring in this way reduces what is a very human act to something else entirely (at the very least, the involvement of human raters in the process of scoring writing samples provides some potential for human perspectives to be taken into account). Again, the downstream consequences of such shifts in scoring are not just related to the implications of receiving a score based on a very narrow construct, but, more significantly, on what messages writing for an algorithm teaches students about what qualities matter in their own writing.

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

7. Translanguaging – a looming ethical dilemma for testing?

Translanguaging represents the phenomenon where users of language draw on linguistic resources both within and beyond the target language, including dimensions of users' first language, to maintain the communicative effectiveness of utterances. This is more than code-switching between two (or more) languages. In translanguaging, the boundaries between 'named' languages are challenged and languages operate as part of a multilingual and multimodal meaning-negotiation resource.

In and of itself, translanguaging is not a new phenomenon, and is a strategy that has often been drawn on to different extents in contexts that require communication of individuals across and between different language backgrounds (see, e.g., Bonacina-Pugh et al., 2021, for a state-of-the-art review). Academic interest in translanguaging is, however, relatively recent. It has been described as 'an emerging concept' in the field of applied linguistics (Nagashima & Lawrence, 2022, p. 736) and a 'new linguistic reality' (García & Wei, 2014, p. 29). It also arguably represents 'part of the metadiscursive regimes that students in the twenty-first century MUST perform' (García, 2011, p. 147, our emphasis).

In communicative practice, translanguaging can liberate language learners from the limitations of a monolingual or single language focus, creating a plurilingual space in which learners can operate and make sense of their learning experiences (Cenoz & Gorter, 2021; García & Kleifgen, 2020) and can learn to use and live with fuller repertoires for meaning-making (Jones, 2020). Ushioda (2017) argued for a need to move away from second language acquisition frames of reference that are 'concerned with progression toward proficiency in a particular language' and towards 'a linguistic multi-competence framework' (p. 469).

The practice of translanguaging does not have to represent a denial of attempts, or an abrogation of responsibility, to communicate in the target language. As Jones (2020) noted, its purpose 'is not that language users can do anything they want.' Rather, its judicious use means that 'they are able to bring to bear a wider range of resources to respond to the conventions and contingencies of whatever situation they find themselves in' (p. 540). In this sense, translanguaging represents 'the deployment of a speaker's FULL linguistic repertoire WITHOUT REGARD FOR watchful adherence to the socially and politically defined boundaries of named languages' (Otheguy et al., 2015, p. 283, our emphases). Translanguaging is thus finding its way into the discourse around language pedagogy as a means whereby, in the words of Jones (2020), we 'interrogate and challenge dominant understandings of language' (p. 536).

Translanguaging is, however, largely absent in discussions on language testing and assessment (Schissel et al., 2019). This is no doubt because translanguaging represents a challenge to the conventional understanding of a language as a static and 'bounded' system (Otheguy et al., 2015; Toohey, 2019). As Shohamy (2011) acknowledged, language assessments remain principally single language focused because they are based on 'a monolingual, homogenous, and often still native-like construct' (p. 419) that 'forbid[s] any other languages to "smuggle in"' (p. 421). A significant challenge for adopting translanguaging into testing situations thus remains 'reluctance among test developers to engage in ML [multilingual] assessment' (Chalhoub-Deville, 2019, p. 472) since 'standardised assessments are usually administered in one language only' (García & Wei, 2014, p. 133).

Given the proposals that support translanguaging, it would seem unethical to deny students the opportunity to translanguage when completing an assessment of their writing. Arguments in favour are similar to those that can be presented regarding the use of dictionaries in writing assessments – translanguaging is an authentic reflection of what language users might do in real life contexts and (seen from the perspective of communicative competence) translanguaging arguably represents one more strategy whereby language users can compensate for communication breakdowns and enhance rhetorical effect (Canale, 1983). Yes, it would require a broadening out and redefinition of the construct of interest in the assessment, but this is something that, from a consequential or ethical point of view, we should not necessarily be reluctant to entertain.

Nonetheless, one counter-argument among language testers is that, if we admit translanguaging or a multilingual frame into the assessment, we would undermine our ability to gather evidence of TARGET

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

LANGUAGE writing proficiency. If a writing test represents a request for information on how well a given student can handle the target language in written discourse, translanguaging surely undermines that goal. If, however, a writing assessment represents a request for information on how well a given student can enact rhetorical intent within and across a range of real-world writing tasks and contexts, translanguaging surely supports that goal. These are complex tensions to resolve. Not least of these tensions are the social and political imperatives driving translanguaging against the monolithic monolingual orientation of established high-stakes assessment systems. We need to consider where we go next in light of the tensions that are apparent.

8. Construct representation as key to the future of writing assessment

Everything that we have outlined so far leads to the inevitable conclusion that writing ability is an inherently multifaceted construct. Corrigan and Slomp (2021) considered more than 100 articles defining the writing construct and synthesised their propositions into a model that describes six domains of expertise in writing: critical discourse knowledge; discourse community knowledge; rhetorical aim knowledge; genre knowledge; substantive knowledge; and communications task process knowledge. These domains are themselves intrinsically complex sub-constructs. White et al. (2015) situated writing ability within a broader and more general span involving seven dimensions – environmental, multimodal integration, rhetorical conceptualisation, cognitive, interpersonal, and intrapersonal – divided into 28 sub-domains. To make valid inferences and judgments about one's writing ability, we must surely take into account evidence related to the facets of this complex matrix of domains. Not every facet can or will be present in a given construct definition or assessment rubric for a particular testing situation. However, when we narrow the construct in our construct sample, we need to account for the consequences that might stem from that design decision. The key issue is to be open to consider what the construct should look like in a particular testing domain, for a particular set of inferences, and for a particular set of test takers. That said, this approach challenges arguments for standardisation.

Where do we begin to address the challenges? With the technocentric or the humanistic? That is, do we begin with the theories and technologies of assessment, or do we begin with the personal, with the human beings who are impacted by our assessment practices? Can we begin from both positions at once? Does a concern for the humanistic obviate concern for the technocentric? Does the technocentric erase the people at the heart of the enterprise?

Elliot (2005) brought a focus on the ethical dimension of the validity/reliability trade-off when he argued:

> . . . what about all those lives that were being changed, all those students who found themselves on the one side or other of a cut score? What lives have been irrevocably altered under assessment conditions that were unreliable? Who makes the apologies, and who awards the reparations? (p. 346)

Hamp-Lyons (2002) suggested four generations of writing assessment. The first three of these, in her view, have been: (1) direct testing; (2) multiple choice testing; and (3) portfolio-based assessment. She proposed that the fourth generation needs to address the importance of balancing the humanistic with the technological. She argued:

> . . . resisting the potential of computer-based assessment to dehumanise and automatize the testing act, the needs of stakeholders must be the focus of a humanistic 'turn' in attention to writing assessment. The huge variation in stakeholders – particularly learners – and their backgrounds and needs must be acknowledged in testing solutions. (p. 13)

Hamp-Lyons continued:

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

> The ethical dilemmas and challenges we face in balancing society's need for assessments with our determination to do our best for learners are very great. Accepting a shared responsibility for the impact of writing assessment practices will put consideration of our own ethical behaviour at the top of our agenda. (p. 14)

Now more than 20 years later, we should perhaps rephrase Hamp-Lyons' assertion here – acceptance of a shared responsibility MUST put ethics to the top of the assessment agenda as we seek to balance accountability and measurement with doing the best for our learners.

As we argued at the beginning, the concepts of construct definition, validity, reliability, and fairness represent the foundations to assertions around what makes an assessment of writing proficiency good or useful. Perhaps, then, the place to start is with our definitions of the construct. These definitions must be informed by what writing is, or has become, or needs to become, in the real worlds that our students inhabit. We cannot know if we are measuring the right knowledge, skills, and dispositions if we do not first know what knowledge, skills, and dispositions we should be aiming for. Furthermore, our constructs must be clearly understood and articulated in advance of assessment implementation. From that starting point, we can begin to consider what represents valid assessments of that construct, how those assessments may be put into operation, and how the evidence from those assessments may be evaluated to provide meaningful information. Fundamentally, ethical assessment practices must pay systematic attention to the social consequences of assessment design and use at every stage of an assessment's development.

None of this will be an easy task. In essence, each design choice we make carries with it consequences for the information the assessment collects, for the inferences and decisions that we make on the basis of test scores, and for the students, teachers, and systems of education within which these assessments are situated. At the same time, each design choice reflects an underlying struggle for control of educational systems and enterprises. Both innovation and stagnation in assessment practices are reflections of the state of the struggle between a range of stakeholders: teachers, researchers, testing firms, governing bodies, and students – who each bring different assumptions, perspectives, and demands to educational systems and to the act of assessing writing. In the process of insisting on the centrality of ethics in our decision-making, we need to consider what White had proposed in the very first issue of *Assessing Writing*, back in 1994 – the future of writing assessment must lie in negotiation and compromise among the different interest groups involved.

As part of the process of dialogue between stakeholders, we as moral agents, those of us who design and use assessments of writing, need to be critical and reflexive, prepared to question established practices, willing to recognise the consequences and impacts stemming from those practices, and ready to take action to shift practices to ensure that the interests of everyone affected by our assessments are addressed. We present this First Person Singular article as a provocation not to shy away from the complex and potentially messy debates that will ensue.

Acknowledgements. This article is dedicated to the memory of Liz Hamp-Lyons, who passed away on 9 March 2022. Liz was Editor Emeritus of *Assessing Writing*, having edited the journal from 2002 to 2016, and had a passionate concern for the consequences of writing assessments on those who were required to take these assessments.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Beigman Klebanov, B., Ramineni, C., Kaufer, D., & Yeoh, P. (2019). Advancing the validity argument for standardized writing tests using quantitative rhetorical analysis. *Language Testing*, *36*(1), 125–144. https://doi.org/10.1177/0265532217740752

Bonacina-Pugh, F., da Costa Cabral, I., & Huang, J. (2021). Translanguaging in education. *Language Teaching*, *54*(4), 439–447. http://doi.org/10.1017/S0261444821000173

Canale, M. (1983). On some dimensions of language proficiency. In J. W. J. Oller (Ed.), *Issues in language testing research* (pp. 333–342). Newbury House.

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. http://doi-org.ezproxy.uleth.ca/10.1093/applin/I.1.1
Cenoz, J., & Gorter, D. (2021). *Pedagogical translanguaging*. Cambridge University Press. http://doi.org/10.1017/9781009029384
Chalhoub-Deville, M. (2019). Multilingual testing constructs: Theoretical foundations. *Language Assessment Quarterly*, *16* (4-5), 472–480. http://doi.org/10.1080/15434303.2019.1671391
Corrigan, J. A., & Slomp, D. (2021). Articulating a sociocognitive construct of writing expertise for the digital age. *The Journal of Writing Analytics*, *5*, 142–195. http://doi.org/10.37514/JWA-J.2021.5.1.05
Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing*, *8*(2), 73–83. http://doi.org/10.1016/S1075-2935(02)00047-8
East, M. (2008). *Dictionary use in foreign language writing exams: Impact and implications*. John Benjamins.
East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88–115. http://doi.org/10/1016/j.asw.2009.04.001
East, M. (2016). *Assessing foreign language students' spoken proficiency: Stakeholder perspectives on assessment innovation*. Springer.
East, M. (2019). Editorial: 25 years of assessing writing. *Assessing Writing*, *42*, 100422, 1-3. http://doi.org/10.1016/j.asw.2019.100422
Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.
García, O. (2011). Educating New York's bilingual children: Constructing a future from the past. *International Journal of Bilingual Education and Bilingualism*, *14*(2), 133–153. http://doi.org/10.1080.13670050.2010.539670
García, O., & Kleifgen, J. A. (2020). Translanguaging and literacies. *Reading Research Quarterly*, *55*(4), 553–571. http://doi.org/10.1002/rrq.286
García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Mcmillan. http://doi.org/10.1057/9781137385765
Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*. The Falmer Press.
Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Open University Press.
Guapacha, M. (2022). Cognitive validity evidence of computer- and paper-based writing tests and differences in the impact on EFL test-takers in classroom assessment. *Assessing Writing*, *51*, 100594, 1-21. http://doi.org/10.1016/j.asw.2021.100594
Hamp-Lyons, L. (2000). Fairnesses in language testing. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 30–34). Cambridge University Press.
Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, *8*(1), 5–6. http://doi.org/10.1016/S1075-2935(02)00029-6
Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 200 writing: Composition, community, and assessment*. Educational Testing Service.
Huot, B. (2002). *(Re)articulating writing assessment for teaching and learning*. Utah State University Press.
Jones, R. (2020). Creativity in language learning and teaching: Translingual practices and transcultural identities. *Applied Linguistics Review*, *11*(4), 535–550. http://doi.org/10.1515/applirev-2018-0114
Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). American Council on Education.
Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. http://doi.org/10.1111/jedm.12000
Kelly-Riley, D., & Whithaus, C. (2016). Introduction to a special issue on a theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1), 1–4. Retrieved from https://escholarship.org/uc/item/8nq5w3t0
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. (3rd ed., pp. 13–103). Macmillan.
Moss, P. (1994). Validity in high stakes writing assessment: Problems and possibilities. *Assessing Writing*, *1*(1), 109–128. https://doi.org/10.1016/1075-2935(94)90007-8
Nagashima, Y., & Lawrence, L. (2022). To translanguage or not to translanguage: Ideology, practice, and intersectional identities. *Applied Linguistics Review*, *13*(5), 735–754. http://doi.org/10.1515/applirev-2019-0040
Oh, S. (2022). The use of spelling and reference tools in second language writing: Their impact on students' writing performance and process. *Journal of Second Language Writing*, *57*, 100916, 1-12. https://doi.org/10.1016/j.jslw.2022.100916
Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, *6*(3), 281–307. http://doi.org/10.1515/applirec-2015-0014
Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in assessing writing. *Assessing Writing*, *42*, 100418, 1-21. http://doi.org/10.1016/j.asw.2019.100418
Schissel, J., Leung, C., & Chalhoub-Deville, M. (2019). The construct of multilingualism in language testing. *Language Assessment Quarterly*, *16*(4-5), 373–378. http://doi.org/10.1080/15434303.2019.1680679
Shohamy, E. (2001). The social responsibility of the language testers. In R. L. Cooper (Ed.), *New perspectives and issues in educational language policy* (pp. 113–130). John Benjamins Publishing Company. https://doi.org/10.1075/z.104.09sho
Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *Modern Language Journal*, *95*(3), 418–429. http://doi.org/10.1111.j.1540-4781.2011.01210.x

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press

Slomp, D. (2016). Ethical considerations and writing assessment. *Journal of Writing Assessment*, *9*(1), 1–6. Retrieved from https://escholarship.org/uc/item/2k14r1zg

Slomp, D. (2019). Complexity, consequence, and frames: A quarter century of research in assessing writing. *Assessing Writing*, *42*, 100424, 1-17. http://doi.org/10.1016/j.asw.2019.100424

Slomp, D., & East, M. (Eds.). (2019). *Assessing writing special issue: Framing the future of writing assessment*. Elsevier.

Slomp, D., & Fuite, J. (2005). Following Phaedrus: Alternate choices in surmounting the reliability/validity dilemma. *Assessing Writing*, *9*(3), 190–207. http://doi.org/10.1016/j.asw.2004.10.001

Slomp, D., Marynowski, R., Holec, V., & Ratcliffe, B. (2020). Consequences and outcomes of policies governing medium-stakes large-scale exit exams. *Educational Assessment, Evaluation and Accountability*, *32*(4), 431–460. http://doi.org/10.1007/s11092-020-09334-8

Toohey, K. (2019). The onto-epistemologies of new materialism: Implications for applied linguistics pedagogies and research. *Applied Linguistics*, *40*(6), 937–956. http://doi.org/10.1093/applin/amy046

Ushioda, E. (2017). The impact of global English on motivation to learn other languages: Toward an ideal multilingual self. *Modern Language Journal*, *101*(3), 469–482. http://doi.org/10.1111/modl.12413

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

White, E. M. (1994). Issues and problems in writing assessment. *Assessing Writing*, *1*(1), 11–27. https://doi.org/10.1016/1075-2935(94)90003-5

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Utah State University Press.

Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing*, *1*(1), 129–139. https://doi.org/10.1016/1075-2935(94)90008-6

Martin East is Professor of Language Education and current Head of the School of Cultures, Languages and Linguistics, the University of Auckland, New Zealand. Prior to this, he was for ten years a language teacher educator in the University's Faculty of Education and Social Work. His research interests lie principally in innovative pedagogical and assessment practices, with a particular focus on task-based language teaching (TBLT) and task-based language assessment (TBLA).

David Slomp is Professor of Literacy and Assessment and Associate Dean of Graduate Studies and Research in Education in the Faculty of Education, the University of Lethbridge, Alberta, Canada. He is also Board of Governors' Teaching Chair. These roles have provided him with opportunities to focus on research and teaching excellence in post-secondary settings. His research principally examines the ethics of assessment and how assessment practices impact students, teachers, and educational systems.

Cite this article: East, M., & Slomp, D. (2023). The ethical turn in writing assessment: How far have we come, and where do we still need to go? *Language Teaching* 1–12. https://doi.org/10.1017/S0261444823000034

https://doi.org/10.1017/S0261444823000034 Published online by Cambridge University Press