Advances in Health Sciences Education (2020) 25:1003–1018
https://doi.org/10.1007/s10459-019-09939-9

REFLECTIONS

The compatibility principle: on philosophies in the assessment of clinical competence

Walter Tavares[1,2] · **Ayelet Kuper**[1,3,4] · **Kulamakan Kulasegaram**[1,5,6] · **Cynthia Whitehead**[1,5]

Received: 22 March 2019 / Accepted: 25 October 2019 / Published online: 1 November 2019
© Springer Nature B.V. 2019

Abstract

The array of different philosophical positions underlying contemporary views on competence, assessment strategies and justification have led to advances in assessment science. Challenges may arise when these philosophical positions are not considered in assessment design. These can include (a) a logical incompatibility leading to varied or difficult interpretations of assessment results, (b) an "anything goes" approach, and (c) uncertainty regarding when and in what context various philosophical positions are appropriate. We propose a compatibility principle that recognizes that different philosophical positions commit assessors/assessment researchers to particular ideas, assumptions and commitments, and applies ta logic of philosophically-informed, assessment-based inquiry. Assessment is optimized when its underlying philosophical position produces congruent, aligned and coherent views on constructs, assessment strategies, justification and their interpretations. As a way forward we argue that (a) there can and should be variability in the philosophical positions used in assessment, and these should be clearly articulated to promote understanding of assumptions and make sense of justifications; (b) we focus on developing the merits, boundaries and relationships within and/or between philosophical positions in assessment; (c) we examine a core set of principles related to the role and relevance of philosophical positions; (d) we elaborate strategies and criteria to delineate compatible from incompatible; and (f) we articulate a need to broaden knowledge/competencies related to these issues. The broadened use of philosophical positions in assessment in the health professions affect the "state of play" and can undermine assessment programs. This may be overcome with attention to the alignment between underlying assumptions/commitments.

Keywords Assessment · Clinical competence · Educational measurement · Performance based assessment · Philosophy · Rating process · Validity

✉ Walter Tavares
walter.tavares@utoronto.ca

Extended author information available on the last page of the article

Introduction

A range of often implicit philosophical positions have come to underlie perceptions of the nature of competence (i.e., ontologies), how people come to know, understand and assess competence (i.e., epistemologies), and how these can be collectively justified. By philosophical positions we mean sets of recognized assumptions and commitments or intellectual frameworks that provide assessment scholars with lenses for examining assessment problems and solutions. For example, emphasizing reliability as an indicator of the degree of error associated with an assessment model is influenced by a psychometric or measurement model that is closely aligned with positivism or post-positivism. Where philosophical positions could once have been reasonably assumed to be shared implicitly within the assessment community, advances in assessment science have suggested that a wider range be considered (Govaerts et al. 2007; Govaerts and Vleuten 2013; Hodges 2013). This has led to such innovations as the increased use of comments (Brutus 2010; Hanson et al. 2013; Ginsburg et al. 2015) and the treatment of rater variability as signal as well as noise (Gingerich et al. 2014). The importance of defining ontological and epistemological stances and ensuring their alignment has been clearly established within the health professions education research domain. However, such explicit acknowledgement of the range of beliefs about what is real and knowable and about how people come to know and understand has largely been overlooked within assessment. This has the potential to undermine assessment scholarship and practice due to unwitting contradictions and incongruence between assessment designers' beliefs about competence, its assessment, and justifications. While the use of a wider range of assessment strategies drawing on a variety of underlying philosophical positions promises to advance assessment science, this is also raising new logical and empirical challenges related to the compatibility of those strategies in practice.

At least three potential practical problems can emerge when philosophical positions are not carefully considered in the design of assessment programs. These are derived when misalignment between major components of an assessment program exist, specifically between (a) what competence is (ontology), (b) how specific strategies and activities are used to determine it (epistemology) and (c) how claims or inferences derived from the assessment program can be justified. First, the implicit linkages of philosophical positions associated with conflicting assumptions and commitments to these three assessment program components can lead to confusing and poorly supported practices with multiple possible interpretations of results. Second, when issues of compatibility are ignored, an "anything goes" approach can dominate assessment ignoring how philosophical positions matter to the detriment of the work being done. Third, the increasing number of possibilities for ways to think about competence, assessment and justification without a clarification of their underlying philosophical positions leaves educators questioning when, how, why, and in what contexts each of these philosophical positions may be appropriate, and whether practices rooted in different positions can be merged or exist simultaneously.

As a potential way forward, we propose a *compatibility principle* which provides a fundamental logic in the process of assessment and its justification. This notion of compatibility is adapted from an incompatibility (incommensurability) thesis in mixed methods research in which researchers have debated the suitability of merging different philosophically-oriented approaches (Howe 1988; Johnson and Onwuegbuzie 2004; Hathcoat and Meixner 2017). In assessment, we propose that the compatibility principle refers to the obligation to recognize that different philosophical positions can exist between and within assessment plans and that these positions commit assessment designers to particular

ideas, assumptions and commitments. This applies to conceptualizations of competence (i.e., constructs), assessment strategies (i.e., guiding principles informing processes) and activities (e.g., tools, data collection and analyses), as well as justification (e.g., application of validity frameworks) and interpretation. Our primary concern is the logic of a philosophically-informed assessment, its assumptions, and its commitments, rather than with the assessment tools or mechanics being used, as the latter can often be used to produce data within multiple different positions. We argue that it is these underlying philosophical positions and their enactment that determines the appropriateness of an assessment plan/program. Here we are primarily concerned with the state of assessment science (while recognizing that an analogous focus on ontological and epistemological alignment is a rather old concept to those who study and teach about rigor in the research process, including for research about assessment). However, while borrowing from the research process, our framing is different in that the role and relevance of philosophical positions in the context of a growing diversity of positions in assessment and its component parts has yet to be recognized as fundamental when specifying criteria for good assessment (Norcini et al. 2011; Eva et al. 2016; Norcini et al. 2018).

Thus, in this perspective we examine first the plurality of philosophical positions informing components of assessment including (1) competence, (2) assessment strategies and (3) justifications. Second, we propose a philosophical congruence or compatibility principle as a way of overcoming the risk of undermining assessment programs. We provide examples—real and hypothetical—of how philosophical positions alter similar features of assessment and how incongruence among the three components of assessment can lead to different interpretations. Third, we raise competing arguments against a compatibility principle, specifically the role of pragmatism, accepted exemplars, and philosophical pluralism. Fourth, we identify future directions, for the community to consider to advance the role and congruence of philosophical positions in the assessment of clinical competence. Our aim is to optimize assessment work by problematizing and making visible these theoretically informed but practical assessment issues. Our team of writers is intentionally ontologically diverse, having academic contributions in assessment informed by positions that span post-positivism, constructivism/interpretivism, and critical theory. We leveraged these diverse positions in our thinking through of the roles of philosophies in the assessment of competence.

Plurality in the philosophical positions informing components of assessment programs

Competence

Conceptualizations of competence are plentiful in part because of the philosophical positions that inform them. Consider the ontological position that competence exists, can be viewed as latent traits, and can be measured using a limited number of representative dimensions.(Kane 1992; Epstein and Hundert 2002) (see Fletcher et al. 2003; Norcini et al. 2003; Kim et al. 2006; Tavares et al. 2012 as applied examples). This view aligns with philosophical positions that speak of concepts of truth—positivism, post-positivism—even if enacted differently across social contexts. However, these ontological positions are no longer universally-held within health professions education. Some have argued instead that post-positivism may underestimate the complexities of competence, calling for alternative

understandings informed by different philosophical stances (Govaerts et al. 2007; Hodges 2013; Bartels et al. 2017). They suggest there are constructs for which competence is interpersonal, socially situated, co-constructed, and not directly measurable. In this view, a limited number of dimensions fails to represent well all the relevant attributes or factors necessary to understand competence. For instance, Kuper et al. (2007) holding an interpretivist/constructivist ontology, described competence as context-laden, situational, interpersonal and socially constructed (Kuper et al. 2007). Similarly, Hodges and Lingard (2013) described how "competence" is shaped by cultural, political, social, behavioral and economic representations of what it is to be a competent professional at a particular time and place (Hodges and Lingard 2013). Competence can therefore be both as reflective of true, latent attributes with measurable representative dimensions and/or as co-constructions derived from and mediated by social interactions.

Assessment strategies

Different philosophical positions also guide us differently as to how competence should be assessed. This involves a type of epistemic pairing or alignment between the underlying ontological assumptions and commitments informing a construct and its assessment. For instance, depending on the underlying philosophical position(s) we hold, we may disagree on the purpose of multiple sampling, what data sources serve as meaningful, how raters are to engage in the assessment process, and/or how we should interpret the methods or quality of an assessment program. Assessment strategies that may be viewed as generating high levels of measurement error under one ontological and epistemological pairing may be precisely the kinds of activities that would be most informative in another. In other words, if educators/researchers understand competence in fundamentally different ways, it follows that some assessment activities may be better aligned with one or the other. For example, Whitehead et al. (2015) positioned competence ontologically as socially bound and co-constructed (i.e., consistent with constructionism). Their subsequent recommendations to use ethnographic assessment activities therefore meets logical epistemic compatibility because of a need to assess competence as a series of complex social constructs in context (Whitehead et al. 2015). However, some flexibility in epistemic pairings exists when one considers the degree of *neutrality* associated with selected assessment activities. For example, the use of text might be regarded as epistemologically neutral because the functional activity of gathering words intended to capture or reflect notions of competence may be analyzed in more than one way: analytic possibilities range from classical content analysis (i.e., counting how many times a word is used) (Neuendorf 2016) to theory-informed inductive and interpretive critical approaches (Lincoln and Guba 1985), thus allowing the textual data itself a degree of neutrality. The alignment that assessment activities and strategies have with ontological positions on competence and the degree of their "neutrality" therefore both become relevant.

Validity as justification

Another major component of assessment is the establishment of the logical and/or empirical support for inferences, decisions or meanings applied to assessment program data (i.e., justifications) using validity frameworks. While there is consensus on many aspects of validity Cizek 2012, 2016) one key disagreement is in the underlying philosophical position. That is, the inclusion of a validity framework in and of itself is neither inherently

indicative of, nor naturally associated with, a particular philosophical position, making coherence with varied positions on competence and assessment activities flexible but also uncertain. For example, Messick claimed to be describing validity in a "constructive-realist" way, arguing that attributes are real, independent of our attempt to characterize them, but that there may also be constructs for which there is no counterpart in reality (Hood 2009; Messick 1989). Others have argued that Messick's validity model "does not require a commitment to existence of the attribute purportedly being measured" and that the model is ontologically free or possibly aligned with realism (Hood 2009). More recently. Kane (2013) describes validity as a pragmatic, scientific activity (Kane 2013). Concerns over philosophical positions are downplayed in place of the view that validity and validation are simply contingent on the claims being made. Kane (2013) makes no claims about representing a "truth", and instead argues for a justified belief obtained using whatever means necessary, leaving truth in the background. (Kane 2013) Borsboom, however, argues that truth is an important feature of validity frameworks and that 'justified true belief' should replace 'justified belief' because truth and justification can come apart (Borsboom and Markus 2013; van Heerden Gideon and Mellenbergh 2013). Thus validity is increasingly used by researchers leveraging different philosophical perspectives to incorporate many different ways of arguing for a justified belief, but it is unfortunately too often adopted without explication of the relevant philosophical issues and their implications. Evan a recent review exploring validity in health professions education did not identify philosophical positions as a main finding other than suggesting researchers used validation strategies that focused on "psychometric", "mostly psychometric" or "mostly expert judgment". (St-Onge et al. 2017).

In summary, as a result of the variety of contemporary philosophical positions there is both variability and (when left as tacit) indeterminacy in the nature/definition of competence, the strategies and activities used to assess that competence, and their validity/validation: logical coherence between these can no longer be assumed. See Fig. 1 for an illustration of these issues. The philosophical positions held about the major components of assessment should translate into different practical choices as well as different

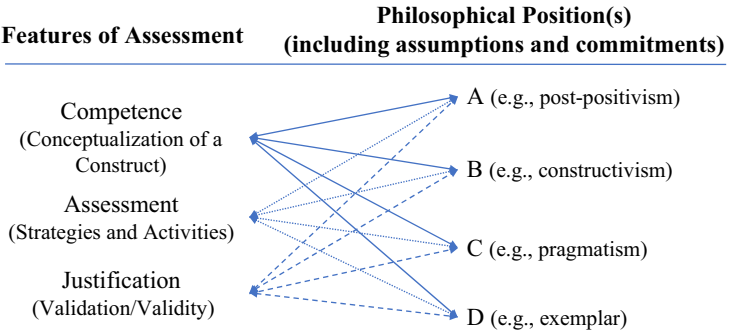


Fig. 1 Illustrating the individual features of assessment, including conceptualization of competence, assessment strategies and activities, and justification and how they can be informed by or aligned with a number of different philosophical positions. Various epistemic combinations are also illustrated

interpretations of the quality and appropriateness of assessment programs. When philosophical positions are ill-defined or poorly considered, including those underlying validity, then those who develop and evaluate assessment programs may end up ascribing meaning in different ways (Wiliam 2017). Drawing on debates in the mixed methods literature, we therefore propose a compatibility principle as a way of mitigating the risk of undermining assessment programs by giving attention to philosophical congruence between assessment components.

The compatibility principle: overcoming the risk of undermining of assessment programs with attention to philosophical congruence

Compatibility (or commensurability) of philosophical positions, introduced initially by Kuhn (1970), became a strong theme as mixed methods research evolved (Howe 1988). In that context researchers debated the incompatibility of different philosophical positions within the same study (Howe 1988). The argument in part was that problems emerge not necessarily at the level of practice (e.g., data collection) or methods—there is no one-to-one correspondence between philosophical positions and methods (Denzin and Lincoln 2011; Ghiara 2019) but at the level of philosophical positions and the assumptions, commitments and interpretations they impose. This debate led to some pragmatic views, suggesting that so-called quantitative studies could include concepts that were more ontologically compatible with qualitative methods (and vice versa), and that the philosophical positions underlying each were logically independent (Johnson and Onwuegbuzie 2004). Some felt that the debate should not be at the level of the methods or of the nature of reality, but at the level of the question and whatever it took to provide reasonable inquiries and assertions (Hathcoat and Meixner 2017). However, others argued that mixing methods depended on maintaining philosophical consistency, as opposed to methods, and that the pragmatic "what works" maxim neglected the unavoidable philosophical aspects of social inquiry (Hathcoat 2013) In this framing, incompatibility can emerge when researchers approach problems with philosophical positions whose assumptions are contradictory or in conflict. This issue of (in)compatibility continues to be argued in the mixed methods research community (Ghiara 2019).

As we have described, within assessment in health professions education different philosophical positions now exist at the levels of competence, assessment strategies, and justification. For each of these components and overall, once a position is adopted, one is committed to the empirical consequences of that perspective. When these are in conflict, a risk of logical and empirical incompatibility emerges. As a result, justification arguments may become flawed, multiple interpretations of the same data can exist, strategies that ignore underlying assumptions may be blindly perpetuated, and the ways to best utilize or integrate new scholarship in assessment become unclear. A compatibility principle in assessment, on the other hand, promotes a logic of philosophically-informed assessment that drives assessment choices and determines the boundaries of acceptable assessment strategies for a given context.

The compatibility principle applied

In the examples that follow, we position the assumptions and commitments associated with positivism/post-positivism in contrast with those of constructivism in the context

of assessment programs. We selected these positions based on recent recommendations in health professions education to leverage the latter in response to the limitations of the former (Govaerts et al. 2013). We recognize that there is nothing new in advancing that these two philosophical positions are incompatible, but their application in assessment is less obvious when considering conceptualization of competence, assessment strategies, justifications, and their interactions. Second, we recognize that there are many more philosophical positions to consider and that lines are likely more blurred than we illustrate here. It is the attention to the assumptions and commitments of any philosophical position (as opposed to a qualitative vs. quantitative debate) that assessment designers intend to use that is of importance here. Our thesis is that different positions mandate different obligations and interpretations for each component of assessment. As such, left unattended, incongruencies can emerge between these components that can undermine assessment work.

Example 1: the use of observers in assessment

Consider, as a hypothetical example, assessments of clinical competence based on assessors' judgments and the importance of sampling their observations. A number of decisions are made that are unavoidably linked to underlying philosophical positions. For example, adopting a post-positivist position (Phillips and Burbules 2000) might include positioning competence as a collection of latent attributes existing within individuals (or teams/groups) that are measurable but with some degree of imperfection. Post-positivists may position assessors and their contributions as opportunities to objectively and accurately transfer information based on candidate behaviors exhibited in response to clinical stimuli, but with a recognized degree of bias or human influence that is considered noise or error to be eliminated. Rater training is applied to further mitigate the degree of error attributable to assessors. Rating tools may be highly structured (to assist raters and minimize error) with or without text. Where text is used, it is to further elaborate on latent attributes. Sampling in this context is a method for reducing error (mathematically) attributable to raters. Finally, justification processes may use validity frameworks that include metrics such as reliability, inter-rater reliability, and item analyses to explore assumptions related to (and eventually to correct) rater behaviors.

However, if viewed from an interpretivist/constructivist perspective, competence may be more of a negotiated sociocultural construction. Raters are recognized and valued as subjective, meaningfully idiosyncratic (the more the better), values-driven, and having views and contributions that are acceptably shaped by social rules and contexts (e.g., environment, society, professional culture). Rater orientation—rather than rater training intended to correct behavior—is intended to have assessors understand their role and how their contributions may be used. Rating scales may be less structured and textual data is prioritized. Here text is most valuable when it is sufficiently varied and is recognized as one of multiple meaningful (i.e., rich and diverse) source of data. Sampling may be intended to ensure that the collection of raters, in their unique contributions, sufficiently represent the construct. Justification may also involve validity frameworks (Cook et al. 2016), but the arguments for a justified belief may draw on considerations of rater reflexivity, the diversity of rater contributions, and the credibility and trustworthiness of the data being generated.

Example 2: the use and conversion of textual data

Alternatively, consider a recent study exploring the use of numeric versus narrative evaluations (Bartels et al. 2017). The authors' ontological and epistemological positions are unclear, but we are left to assume a constructivist stance—though we cannot be certain—based on their thorough review and fair criticism of alternative positions. The authors retrospectively collected numerical and narrative data from in-training evaluation reports, had experts assign numbers to the textual data, then calculated correlation and reliability analyses. At this point, compatibility between ontology and epistemology becomes a concern *if* our assumption about the authors' philosophical position is accurate. Using the statistical data as a source of justification, the authors conclude that converted narrative evaluations are as reliable as numeric scoring. The authors, we assume, found the study acceptable because of an implicit set of views. However, multiple interpretations are now possible including: (a) that the conversion from text to numerical data to reliability was appropriate; (b) that the use of text for the purposes of capturing meaningful variation is inconsistent with this approach; or (c) that we (as knowledge users) are having difficulty resolving a potential compatibility issue and have misinterpreted the study. Many alternative interpretations are thus possible, leaving the interpretation to be guided by the philosophical positions of the user, including what counts as evidence and what is problematic. Further, the ways in which this work can be positioned within broader conversations about assessment that would advance that field become less certain, not necessarily to the authors, for whom these may seem clear, but to users of their contribution who are left to impose their own positions. In fairness to the authors whose paper we have discussed in this paragraph, we note that we too have been guilty of not being explicit about our own philosophical positions in assessment-related papers; see for example Tavares et al. 2016. Also, the lack of explicit philosophical positioning in their paper is in no way unusual or below the current standard of practice in the field; indeed, it is illustrative of the extremely common phenomenon that provided part of the impetus to write this article.

These examples illustrate how assessment strategies that appear the same on the surface (e.g., sampling, rater contributions) can hold different meanings and roles that may align more or less with certain conceptualizations of competence: this is parallel by differences in how validity and validation are applied. They also make clear that if philosophical positions, as opposed to methods, are mixed within an assessment program, conflicting interpretations of its results are possible. This places both educators and assessment researchers in a position of uncertainty about how to proceed with future assessment plans or standards. See Table 1 for a side by side summary of the ways in which assessments programs using different philosophical positions could approach similar features of assessment in different ways.

Further compatibility issues arrive when attempting to combine very different assessment strategies in a single program without considering their philosophical assumptions. Consider, for example, the effort to combine both simulation and workplace based assessments for a more robust summative decision making process regarding competence. In a simulation context, competence is often clearly defined and assessments promote standardization, consistency, reliability and objectivity. These are common attributes closely aligned with a post-positivist philosophical position. However, in workplace contexts, for some, competence may be conceptualized as dynamic, contextually bound, constructed and socially situated with formal and informal dimensions. Structuring assessment programs may mean promoting diversity, richness,

Table 1 A side by side summary of the different ways in which assessments programs using different philosophical positions could approach similar features of assessment in different ways

Legend	Post-positivism/objectivism	Constructivism/interpretivism
Definition	Phenomena exist within individuals, there is some degree of a True level of ability (i.e., a true score) but only imperfectly and probabilistically apprehensible. Assessment proceeds through standardization, sampling, objectification but with a recognized degree of subjectivity/bias or human influence	While there is a physical reality, the meaning made of it is necessarily a negotiated socio-cultural construction, as are all understandings of the social world more broadly. Meaning comes into existence due to our engagement with learners/candidates in social contexts/world. Assessment proceeds through rich diverse usually qualitative data collection (although not exclusively) that is intentionally subjective
Competence/construct	Discoverable. Latent attributes exist within individuals and can be measured, but their enactment and measurement are influenced by social factors. Discovered through measurement models while attempting to mitigate error. A property of the individual	Competence is not a characteristic of individuals but is embedded in collectivities. It is contextual, constructed and changeable. Socially constructed. A property of the professional body of knowledge
On the assessment process	Involves a set of rules for combining data to reach consistent and reproducible inferences on individuals. Decisions are fully traceable, debatable, revisable	A social, contextual, specific process of inquiry and of interpretation. Multiple interpretations are encouraged, and is an active assessor-learner/candidate activity
On justification	A statistical principle applied in a number of ways (e.g., generalizability theory)	Based on principles of rigour in qualitative research (e.g., credibility, transferability, dependability, confirmability)
On reliability	Refers to the reproducibility of assessment data or the ability to consistently differentiate. Serves as an indicator of the degree of error in the assessment process	Not a strongly held concept, certainly not in statistical terms
On how assessors are used/viewed	As objective/subjective measurement devices with some controllable and uncontrollable biases to account for. Intent is to have consistency between raters. Error prone. Unconsciously biased. Contributes scores or narratives. Rater training to calibrate	Subjective participant. To engage in dialogue/relationship with the learner/candidate. Contributes rich and unique information. Intentionality (i.e., close and contributory interaction between candidate and observer). Dependent on values, culture, background, expectations. Can contribute scores but largely narrative focus. Rater training is counter intuitive. Orientation to assessment goals may be helpful
On how sampling is used/viewed	To eliminate sampling error/mitigate other sources of error (e.g., context, rater error)	Communal interpretation influenced by values. Promotes diversity of input
On the role of context	Source of variance to be mitigated (i.e., context specificity) through sampling. Unattractive nuisance threatening validity. Low degree of within-person consistency of performance across assessments	Source of information; contributes meaning. Context is central/inextricably linked. Creates meaning and is acknowledged as part of the process. Absence of context distorts the assessment

Table 1 (continued)

Legend	Post-positivism/objectivism	Constructivism/interpretivism
On subjectivity	To be avoided/minimized. Considered a source of error threatening validity. Can be minimized	Embraces intentionality. Different interpretations of the same subject are expected/valued. Influenced by context. Cannot be avoided
On how text is used	To provide support for claims about truth	To formulate meaning through value laden interpretations
On standardization	Standardization of assessment processes is promoted to support objectivity	Standardization is not an inherent feature
On candidate involvement	Exhibiting behaviours in response to clinical stimuli and "distanced" from the assessor	Exhibiting behaviours in response to clinical stimuli and may engage in interactions with the assessor. Contributes to the construction of "competence"
On the role of values	Reject the roles of values in practice as much as possible but recognizes they exist	Values are central

comprehensiveness and subjectivity. This is more closely aligned with a constructivist philosophical position. When the two are combined, we might find ourselves in a conceptual struggle. For instance, under which philosophical position are we to inform justification strategies and in what way does data collected under a different "competing" positions matter? These issues are not inconsequential, particularly when assessment decisions have to be rigorously justified.

Logical extensions: pragmatism and other considerations

In order to move their practical work forward—to get on with the business of doing and justifying assessment activities—some have taken up various counter-positions. One of the claims to which our argument for philosophical compatibility may be vulnerable is the role and suitability of pragmatism as *the* ideal philosophical position to guide assessment science (recognizing that other positions, such as realism or philosophical pluralism, may be of equal interest). We have described assessment as a social process of inquiry, shaped by fundamental philosophical positions about what we are interested in understanding, predicting, and/or making claims. Some may argue that pragmatism, as one such position, is the most appropriate philosophical position in assessment today in part because assessment in the health professions involves socially enacted subjective and abstract constructs that are difficult to define (Guyon et al. 2018). Further, a number of pressures act on assessment plans, including the need to balance feasibility/resources, opportunities, cost, acceptability, rater behaviors, educational impact, etc. These are important practical questions to consider of which pragmatism, for some, may be a solution. Pragmatism does represent a degree of fluidity and openness and may provide a way of clarifying philosophical disputes (Bernstein 1989). These may include resolving issues related to how to choose one philosophical position over another, or tensions with "truths" that are in conflict, are unknowable, or have no impact on assessment (Bernstein 1989; Morgan 2014). However, we draw a clear distinction between the lay definition of being practical or pragmatic as an operational approach ("what works"), which may be what is most commonly meant by pragmatism within assessment science, versus pragmatism as an explicit and deliberate philosophical position with a long intellectual history (Dewey 2018; James 1907) that has much to offer as a workable approach to assessment problems.

Other ideas challenging a compatibility principle are also worth considering. For instance, assessment science may have a number of exemplars that represent a shared belief about the nature and conduct of assessments (e.g., programmatic assessments). Assessments grounded in pragmatism or in the mixing of philosophical positions may be presented as exemplars that may challenge the need for a compatibility principle. A similar concept has been described in mixed methods research as paradigms-as-exemplars and may be argued to provide a way of accommodating nuances in assessment work and the numerous philosophical positions that are likely to populate the assessment field. In addition, there are likely to be instances in which assessment programs can be conducted through different ontological and epistemological pathways or pairings thereby facilitating different points of view and promoting helpful examination of taken for granted understandings. (Greene 2006; Uprichard and Dawney 2016; Ghiara 2019). This complex undertaking requires careful philosophical forethought.

Ways forward

Assessment design and the determination of assessment quality are subject to the beliefs and standards of those who create indicators and evaluate practices. These beliefs and standards are, in turn, applied social processes shaped and informed by numerous philosophical positions. We are encouraged by the growing diversity of philosophical views on assessment in health professions education but also recognize that assumptions and commitments associated with the application of different philosophical positions require new areas of focus. For example, we now must attend to how various positions translate into conceptualizations of constructs and how those are aligned (or not) with specific assessment activities in order to avoid the multiple meanings or conflicting interpretations of results and justifications. As the range of philosophical positions informing health professions assessment broadens, attention to the *logic of assessment practice and justification* is needed. In this way, debates or boundary work over methods can be avoided, since methods are not associated with a specific philosophical position, a philosophical position can be congruent with several similar methods, and researchers can mix methods while adopting one philosophical position. Also, debates over the superiority of philosophical positions can be deemphasized. What is of interest is where and how philosophical positions exhibit clarity, congruence, and contradictions in assessment scholarship and practice.

We do not suggest that the boundaries between philosophical positions are always highly distinct or incompatible. The permeable boundaries across them loosen compatibility constraints in many cases. Much like the value applied to conceptual and theoretical frameworks in education research (Bordage 2009), or advances in mixed methods research (Teddlie and Tashakkori 2012), there is room for philosophical pluralism in the field of assessment. Kuhn described the value of a commonly agreed upon vocabulary and set of principles (not a specific position) to avoid breakdowns in communication. (Kuhn 2012; Morgan 2007) Encouraging any more organization at this point may otherwise be premature and a misguided attempt to establish unclear boundaries or uniformity or, worse, to silence diverse perspectives and opportunities to further understanding.

As a way forward, we recommend the following:

a. Clearly articulate and rationalize philosophical positions—including when they are mixed—so that knowledge users may understand assumptions and make sense of and examine justifications in relation to the stated position(s). For example, an educator may structure guidance for observers differently depending on their choice of philosophical position (e.g., minimize or promote subjectivity). Within the same assessment plan and to the knowledge user, either could be an appropriate or flawed strategy depending on the applied philosophical position(s).
b. Avoid premature claims that one philosophical position is better than another and instead focus on developing the merits, boundaries, relationships, mixing and compatibility within and/or between each. This includes guarding against discussions that degrade to data collection (e.g., quantitative vs. qualitative or other similar distractors), as many philosophical positions can employ several similar methods.
c. Question, refute, refine, extend and/or determine further a core set of values/principles related to the role and relevance of philosophical positions and their compatibility in an assessment context. For example, should compatibility be rejected for philosophical pluralism?

d. Establish strategies and criteria to delineate along a continuum (as opposed to a dichotomy) from 'compatible' to 'incompatible'. Delineating positivism from constructivism may be less complex than what might exist when considering realism, relativism, post-structuralism, pragmatism, or philosophical pluralism.
e. Given the broadening philosophical positions informing assessment, designers (and researchers) may require a broadened knowledge base and set of competencies related to these issues. This may include for example, complementing traditional texts about assessment in the health professions with texts elaborating on the inherent assumptions and commitments associated with various applicable philosophical positions.

Box 1 Examples of implications and processes informed by our recommendations

Recommendation	Examples of implications or process
Articulating and rationalizing philosophical positions	This provides support for interpretations and claims of quality. As an example, depending on philosophical position(s), the role and value of reliability in promoting defensibility might be necessary or it might be inappropriate (and replaced with evidence supporting the dependability of assessments); wide variation in rater contributions may need to be treated with rater training/calibration, or left alone. This may refine standards for reporting on assessment programs
Avoid premature claims that one philosophical position is better than another	Introducing news ways of approaching the complexity associated with the assessment of intrinsic roles, as an example, can be valued and supported when the assumptions and commitments associated with philosophical positions are considered and used. Some philosophical positions may be more suited for assessment of certain constructs over others (e.g., procedural skills vs. intrinsic roles)
Further a core set of values/principles related to the role and relevance of philosophical positions and their compatibility	Exploring the conditions under which philosophical positions may be necessary/unnecessary or threatening to good assessment work (and why) can further advance assessment science. This may include, as an example, identifying if there are circumstances where philosophical positions can effectively be mixed in ways that add rigor and defensibility to assessment programs
Establish strategies and criteria to delineate along a continuum from 'compatible' to 'incompatible'	Analogous to the critical appraisal of research, assessment scholars and educators may develop the same form of appraisal involving philosophical positions in assessment practice and research
A broadened knowledge base and set of competencies related to the role of philosophical positions	Assessment training and texts should broaden foundational knowledge to include broad ontological, epistemological and methodological assumptions, including when these can be mixed and why. This may provide opportunities for stronger assessment work but also opportunities for divergence and dissonance when/if merging positions

See Box 1 for additional examples of implications and processes that are intended to support the advancement of assessment science.

Conclusions

Health professions educators are tasked with assessing multiple nuanced aspects of clinical competence. This has led the assessment community to explore competence, assessment and justifications in ways that involve broadening the range of (often implicit) underlying philosophical positions. This variety of philosophical positions helps to expand our understanding and uses of these component parts in new ways but also introduces a risk of incompatibility between them, that can undermine assessment programs. This includes leaving interpretations of inferences, quality, and/or defensibility difficult, uncertain, or unintended. In this reflection we demonstrate the value of applying a philosophically informed logic of inquiry in the assessment of clinical competence.

Philosophical positions ultimately influence how we act in assessment and the claims we make. As Burbules (1993) wrote about rules according to Wittgenstein's philosophy of investigations: "Like rules [ontologies and] epistemologies guide practice but do not rule it; they are embedded and unconsciously affect the "state of play" of the actors and their values. Not so much a choice as it is a disposition"(Burbules 1993). In order to advance assessment science, we need to make those dispositions explicit in the work that we do.

References

Bartels, J., Mooney, C. J., & Stone, R. (2017). Numerical versus narrative: A comparison between methods to measure medical student performance during clinical clerkships. *Medical Teacher, 39*(11), 1154–1158.

Bernstein, R. J. (1989). Pragmatism, pluralism and the healing of wounds. *Proceedings and Addresses of the American Philosophical Association, 63*(3), 5–18.

Bordage, G. (2009). Conceptual frameworks to illuminate and magnify. *Medical Education, 43*(4), 312–319.

Borsboom, D., & Markus, K. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50*(1), 110–114.

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review, 20*(2), 144–157.

Burbules, N. C. (1993). *Dialogue in teaching: Theory and practice*. New York: Teachers College Press.

Cizek, G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31.

Cizek, G. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy and Practice, 23*(2), 212–225.

Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: Validity evidence for qualitative educational assessments. *Academic Medicine, 91*(10), 1359–1369.

Denzin, N. K., & Lincoln, Y. S. (2011). Preface. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 9–16). Thousand Oaks, CA: Sage.

Dewey, J. (2018). *Logic-the theory of inquiry*. Redditch: Read Books Ltd.

Epstein, R., & Hundert, E. (2002). Defining and assessing professional competence. *JAMA, 287*(2), 226–235.

Eva, K. W., Bordage, G., Campbell, C., Galbraith, R., Ginsburg, S., Holmboe, E., et al. (2016). Towards a program of assessment for health professionals: From training into practice. *Advances in Health Sciences Education, 21*(4), 897–913.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' non-technical skills (ANTS): evaluation of a behavioural marker system. *British Journal of Anaesthesia, 90*(5), 580–588.

Ghiara, V. (2019). Disambiguating the role of paradigms in mixed methods research. *Journal of Mixed Methods Research*. https://doi.org/10.1177/1558689818819928.
Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box'differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055–1068.
Ginsburg, S., Regehr, G., Lingard, L., & Eva, K. W. (2015). Reading between the lines: Faculty interpretations of narrative evaluation comments. *Medical Education, 49*(3), 296–306.
Govaerts, M., Van de Wiel, M., Schuwirth, L., Van der Vleuten, C., & Muijtjens, A. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education, 18*(3), 375–396.
Govaerts, M., van der Vleuten, C., Schuwirth, L., & Muijtjens, A. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*(2), 239–260.
Govaerts, M., & Vleuten, C. P. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education, 47*(12), 1164–1174.
Greene, J. C. (2006). Toward a methodology of mixed methods social inquiry. *Research in the Schools, 13*(1), 93–99.
Guyon, H., Kop, J., Juhel, J., & Falissard, B. (2018). Measurement, ontology, and epistemology: Psychology needs pragmatism-realism. *Theory and Psychology, 28*(2), 149–171.
Hanson, J. L., Rosenberg, A., & Lane, J. (2013). Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Frontiers in Psychology, 4,* 668.
Hathcoat, J. (2013). Validity Semantics in educational and psychological assessment. *Practical Assessment in Research, and Evaluation*, *18*(9), 1–14.
Hathcoat, J. D., & Meixner, C. (2017). Pragmatism, factor analysis, and the conditional incompatibility thesis in mixed methods research. *Journal of Mixed Methods Research, 11*(4), 433–449.
Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568.
Hodges, B., & Lingard, L. (2013). *The question of competence: reconsidering medical education in the twenty-first century*. Ithaca: Cornell Univeristy Press.
Hood, S. (2009). Validity in psychological testing and scientific realism. *Theory and Psychology, 19*(4), 451–473.
Howe, K. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10–16.
James, W. (1907). *Pragmatism: A new name for some old ways of thinking*. New York, NY
Johnson, R. B., & Onwuegbuzie, A. J. J. E. R. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.
Kane, M. (1992). The assessment of professional competence. *Evaluation and the Health Professions, 15*(2), 163–182.
Kane, M. (2013). Validation as a pragmatic, scientific activity. *Educational Measurement, 50*(1), 115–122.
Kim, J., Neilipovitz, D., Cardinal, P., Chiu, M., & Clinch, J. (2006). A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Critical Care Medicine, 34*(8), 2167–2174.
Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.
Kuhn, T. S. (2012). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Kuper, A., Reeves, S., Albert, M., & Hodges, B. (2007). Assessment: do we need to broaden our methodological horizons? *Medical Education, 41*(12), 1121–1123.
Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry* (Vol. 75). Thousand Oaks: Sage.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: Amercian Council on Education and National Council on Measurement in Education.
Morgan, D. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1*(1), 48–76.
Morgan, D. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry, 20*(8), 1045–1053.
Neuendorf, K. (2016). *The content analysis guidebook*. Thousand Oaks: Sage.
Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., et al. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33*(3), 206–214.
Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., et al. (2018). 2018 Consensus framework for good assessment. *Medical Teacher, 40*(11), 1102–1109.

Norcini, J., Blank, L., Duffy, F. D., & Fortna, G. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138,* 476–481.
Phillips, D., & Burbules, N. (2000). *Postpositivism and educational research*. Lanham: Rowman and Littlefield.
St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017). Validity: One word with a plurality of meanings. *Advances in Health Sciences Education, 22*(4), 853–867.
Tavares, W., Boet, S., Theriault, R., Mallette, T., & Eva, K. W. (2012). Global rating scale for the assessment of paramedic clinical competence. *Prehospital Emergency Care, 17*(1), 57–67.
Tavares, W., Ginsburg, S., & Eva, K. (2016). Selecting and simplifying: Rater behavior when considering multiple competencies. *Teaching and Learning in Medicine, 28*(1), 41–51.
Teddlie, C., & Tashakkori, A. (2012). Common "core" characteristics of mixed methods research: A review of critical issues and call for greater convergence. *American Behavioral Scientist, 56*(6), 774–788.
Uprichard, E., & Dawney, L. (2016). Data diffraction: Challenging data integration in mixed methods research. *Journal of Mixed Methods Research, 13,* 19–32. https://doi.org/10.1177/1558689816674650.
van Heerden Gideon, D., & Mellenbergh, J. (2013). Validity and truth. In *New developments in psychometrics: Proceedings of the international meeting of the psychometric society IMPS2001. Osaka, Japan, July 15–19, 2001* (p. 321). Springer.
Whitehead, C. R., Kuper, A., Hodges, B., & Ellaway, R. (2015). Conceptual and practical challenges in the assessment of physician competencies. *Medical Teacher, 37*(3), 245–251.
Wiliam, D. (2017). Assessment and learning: Some reflections. *Assessment in Education: Principles, Policy, and Practice, 24*(3), 394–403.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Walter Tavares[1,2] · Ayelet Kuper[1,3,4] · Kulamakan Kulasegaram[1,5,6] · Cynthia Whitehead[1,5]

1 The Wilson Centre, Department of Medicine, University of Toronto/University Health Network, 200 Elizabeth Street, 1ES-565, Toronto, ON M5G 2C4, Canada

2 Post-MD Education (Post-Graduate Medical Education/Continued Professional Development), University of Toronto, Toronto, ON, Canada

3 Division of General Internal Medicine, Sunnybrook Health Sciences Centre, Toronto, Canada

4 Department of Medicine, Faculty of Medicine, University of Toronto, Toronto, Canada

5 Department of Family and Community Medicine, Women's College Hospital, Faculty of Medicine, University of Toronto, Toronto, Canada

6 MD Program, University of Toronto, Toronto, Canada