Title: Assembled Validity: Rethinking Kane's Argument-Based Approach in the context of International Large-Scale Assessments (ILSAs)

Authors: Camilla Addey (Autonomous University of Barcelona), Bryan Maddox (University of East Anglia and University of Oslo, and Bruno D. Zumbo (University of British Columbia).

Abstract

Drawing on Kane's argument-approach to validity and Toulmin's later work on cosmopolitanism and diversity, this paper asks *whose* validity arguments and evidence are being presented in International Large-Scale Assessments (ILSAs), *where* and *when.* With a case study of the OECD's PISA for Development, we demonstrate that validity arguments are assembled, negotiated and transformed by the network of actors who come together in ILSAs. We claim that the challenge of ILSAs is not to establish a single authoritative argument through the displacement of plural interpretations and uses. We argue that one of the tasks of an argument-based approach to validity is to create a democratic space in which legitimately diverse arguments and intentions can be recognised, considered, assembled and displayed. We therefore suggest that 1) this socio-material practice of assembling validity should be integrated into validity theory and practice, and 2) the task of assembling validity should be informed by democratic principles of diversity and inclusion.

Keywords

Assembled validity, validity, validation, International Large-Scale Assessments (ILSAs), PISA for Development (PISA-D), OECD, Kane

Introduction

How should different validity arguments and evidence be reconciled in situations where there are diverse stakeholders and multiple contexts of use? In this paper, we address that question in relation to International Large-Scale Assessments in Education (ILSAs), with a case study of the OECD's PISA for Development (PISA-D). To do so, we re-think Kane's argument-based approach to validity (Kane, 2013, 2016), considering how validity is assembled by diverse stakeholders as a socio-material practice.

Validity evidence and judgements are always 'assembled' to some extent with evidence from multiple sources. That is, they involve pragmatic attempts to bring together and consider arguments, theory, evidence about interpretations, uses and consequences of assessment (Messick, 1989; Hubley and Zumbo, 2011; Stone and Zumbo, 2016; Kane, 2016). However, in this paper, we suggest that the practice of assembling validity – of reconciling different evidence, interpretations and arguments, is more challenging than is often acknowledged in contemporary theory, or in the practice of assessment programmes. We suggest that the socio-material practice of assembling validity should be integrated into validity theory and practice.

Most contemporary writing on validity promotes clarity of argument as a virtue (Newton, 2012). Hence, we see Messick's (1989) unitary framework, and Kane's (2016) call to 'explicate'

validity. That commitment to clarity, as Kane notes, has a role in 'defining the scope and structure of the discipline' (2016, p. 198). It provides a common language, moves the argument forward, and enables setting of professional standards such as 'The Standards in Educational and Psychological Testing' (2014). However, in the quest for clarity and consensus, validity theory can become rarefied and idealised, and recognition of diversity diminished. In the process, the multiple socio-material practices that constitute and contest validity may become peripheral (Zumbo, 2014).

In this paper, we shine a light on the socio-material validation practices of assessment actors as they assemble validity. Thus, instead of treating validation practices (including those of argumentation) as somehow second order activities, we consider them as part of the essence of validity as it is constituted in social practice. The paper draws on two sets of theory. We reconsider Kane's Argument Based Approach (Kane, 2013, 2016) from the perspective of diversity, drawing on Toulmin's philosophy to tease out contrasting theoretical perspectives. We also draw on Actor Network Theory to consider how networks of human actors and material artefacts come together in scientific projects (Callon 1986; Latour 1987).

In the first part we re-evaluate Kane's Argument Based Approach to validity from the perspective of diversity. To do that we draw on the philosophy of Toulmin, whose influential work on 'The uses of argument' (1958) is central to Kane's work. However, we also use Toulmin's later philosophy (1972, 1990) to critique Kane's work. Toulmin's later philosophy was influenced by the democratic and intellectual revolutions of the 1960's, with a strong emphasis on a pluralistic approach to argument, context and diversity. [1]

Informed by that work we ask: *whose* validity arguments and evidence are being presented, *where* and *when*? We suggest that the challenge of large-scale assessment programmes is not to establish a single authoritative argument (as Kane's work suggests). Instead we argue that the goal of validation practice should be to actively recognise and reconcile the arguments and evidence of diverse actors, who may have legitimate, but different ideas about the purposes of assessment programmes, the appropriate interpretation and use of data and the desirability of their consequences. We therefore suggest that the task of assembling validity should be informed by democratic principles of diversity and inclusion.

The paper then explores that process in the context of an International Large-Scale Assessment (ILSA) and a case-study of the OECDs '*PISA for Development*' (PISA-D). The success of all assessment programmes depends on their ability to recruit and retain necessary actors and resources. For ILSAs, that process involves special demands. Over 80 countries have participated in PISA. PISA-D extended the scope of PISA further by including low- and middle-income countries from South Asia, Central Asia and Africa. The empirical data that the paper analyses was gathered by Addey in her research project 'PISA for Development for

[1] Within the field of educational programme evaluation there were similar concerns at the time about the democratic representation of diverse stakeholders (e.g. Cronbach, 1988; House, 1980, 1990).

Policy' (Addey 2019), and draws on observations of PISA-D meetings and social events, and approximately 30 interviews carried out with OECD staff, OECD contractors, and high level policy actors in Ecuador and Paraguay between 2016 and 2019.

As a result of its global expansion, PISA has the challenge of assembling and holding together a diverse global network of actors, and somehow reconciling their differences within a coherent approach to assessment use and validity. In the second part of our paper, we consider what that process might mean, with the case study of PISA-D in which we present interview testimony to explore the perspectives OECD staff and national level representatives.

The concept of *assemblage* (as we use it in this paper) is drawn from Actor Network Theory (ANT) and the influential work of Callon (1986), Latour (1987, 2005) and Law (2009), though the assemblage concept has wider intellectual origins, for example, in the work of Deleuze and Guattari (1983). Informed by ANT, we consider assemblages as relating to the way that unstable networks of human actors and material artefacts align temporarily to achieve shared goals[2].

Our engagement with ANT has led us to question the implications of the stated conceptual dichotomies in validity theory, between the social and technical, and the distinction between validity theory and practice. As Latour (1987, 1992) has argued, the symmetry (and co-presence) of the social and material is integral to scientific method. Our reading of Kane's work suggests that such symmetry is welcome in validation practice, where the presence of the social within the technical is not viewed as a threat to the rigour of validation. Indeed, Kane's approach considers argument and persuasion, the weighing up of evidence and its interpretation as central to the practice of validation.

In this paper we therefore depart from the established dichotomies in validity theory between the *scientific and ethical*, and the *technical and social*, as proposed by Messick (1980) and Newton and Shaw (2012) to consider the co-existence of socio-material and technical characteristics of validation practice.

Validity theory also tends to make an implicit dichotomous distinction between the initial formulation of validity arguments (i.e. as the source of conceptual purity), and down-stream validation practice. Kane makes references to validation 'in practice', which he illustrates with short illustrative examples in the style adopted in earlier work of Messick, rather than as actual research-based empirical evidence (Kane, 2016, pp. 199-201). That approach clearly has value, as it is worthwhile to consider the validity argument from the outset, just as it is necessary to consider the construct that is under investigation, and the intended use of assessment data. However, we question whether such a dichotomous way of thinking about validity really holds in practice. Instead, we consider the way that validity arguments are assembled, negotiated and transformed by the network of actors who come together in assessment programmes (i.e. with temporal and spatial dimensions).

[2] ANT has previously been applied to study the socio-material and technical aspects of assessment programmes (Gorur, 2011; Addey, 2019; Gorur, Sorensen and Maddox, 2019).

In this paper we therefore apply ANT themes from Callon's (1986) work to consider the initial '*problematization*' as actors identify shared interests and rationales for participation; the operation of technical standards and procedures as '*obligatory passage points*'; and the choreographed processes of '*mobilization*' in which all actors come to speak with a single voice.

Validity in Theory

> '*...a man who asserts something intends his statement to be taken seriously: and, if his statement is understood as an assertion, it will be so taken. Just how seriously it will be taken depends, of course, on many circumstances—on the sort of man he is, for instance, and his general credit.' (Toulmin, 1958, p. 11)*

Kane's argument-based approach to validity is globally dominant in international educational assessments. The authority of Kane's work is enhanced by its scholarly discussion of earlier approaches to validity including the foundations of an argument-based approach in the work of Cronbach (1988) and others. Kane's work also has the backing of powerful institutions such as Educational Testing Service (ETS) and the OECD. It is hard to argue with its credentials (see Chapelle, Enright and Jamieson, 2010).

Kane argues that validity judgements are informed by arguments about the proposed interpretation and uses (IUAs) of test scores rather than an intrinsic property of the test, or its technical performance (Kane, 2013, 2016).

> '*I think of validity as the extent to which the proposed interpretations and uses of test scores are justified. The justification requires conceptual analysis of the coherence and completeness of the claims and empirical analyses* of the inferences and assumptions inherent in the claims.' (Kane, 2016, p. 198)

His approach is informed by Toulmin's philosophical work on 'The Uses of Argument' (1958) with its framework of interpretive, evidence and qualifiers, warrants, backing and rebuttals[3]. Kane applies Toulmin's early work to develop the core components of his approach (see Kane, 2013, p12). Furthermore, Kane adopts Toulmin's philosophical stance in rejecting absolutism and universalism. Validity judgments are therefore viewed as *contingent*, open to re-evaluation or indeed rebuttals based on new evidence, or new information about test uses or consequences (see Kane, 2013). Validity is not established in a-priori arguments about the intended interpretations and use of test results (Kane's Interpretation and Use Arguments, IUA), as it involves on-going judgements about uses and consequences as they play out in practice:

> '*Validity is a matter of degree, and it may change over time as the interpretations/uses develop and as new evidence accumulates. The plausibility of a proposed IUA will increase if ongoing research supports its inferences and assumptions (especially those that are most questionable a*

[3] Toulmin's (1958) work on argument is influential beyond assessment in studies of literature and rhetoric.

priori). Validity may decrease if new evidence casts doubt on the proposed IUA.' (Kane, 2013, p. 3)

Secondly, Kane's recognises the *implications of diversity* (i.e. of context, actors and time). The challenge of diversity is discussed in Toulmin's (1958) early work on 'The Uses of Argument' in the distinction between 'field invariant' and 'field dependent' sources of argument and evidence. The distinction is acknowledged by Kane:

'...we often assume other kinds of invariance, without investigating the impact of violations of these assumptions. […] These assumptions may be plausible if the attribute being assessed is expected to be stable over occasions and contexts, but they are assumptions, and should be acknowledged and considered (even if they are not evaluated empirically).' (Kane, 2016, p. 201)

Applying Toulmin's early philosophy, Kane leaves a door open to *acknowledge* and *consider* the place of diversity. Kane's work contrasts with the more overt commitment to democratic inclusion and recognition of diversity advocated by Cronbach (1988). Cronbach's later position (in contrast to Cronbach and Meehl, 1955) reflects his rhetorical turn towards social science methodology as a form of democratic action embodied in his view of programme evaluation. In comparison, those concerns remain peripheral to Kane's work as exceptions that prove the rule. As a result, validation practices may consider it adequate to recognise diversity, but treat such differences as special cases that somehow lie outside the scope of the validity argument.[4]

Kane cites Toulmin's work on 'The Uses of Argument' (1958) extensively, and it is clearly central to Kane's argument-based approach. However, he does not discuss the significant implications of Toulmin's later work on his approach. We can therefore see a challenge to Kane's work coming from Toulmin's later work, which was already published as Kane wrote his most influential work on validity. In this paper, we consider Toulmin's work as particularly useful for an analysis of assessment validity in contexts of ILSA diversity.

Toulmin's later work radically revised some of the fundamental tenets of his philosophy in response to the socio-cultural and intellectual movements of the 1960's.

'*For myself, in the late 1960's I began to be uneasy about the received account of 17th Century ideas. The cultural changes that began around 1965 were (it seemed to me) cutting into our traditions more deeply than was widely appreciated.' (Toulmin, 1990, ix)*

In his later work, Toulmin's emphasis on establishing authoritative arguments and their credentials was replaced by a concern with the legitimacy of diverse forms of understanding. In 'Human Understanding' (1972), Toulmin's earlier concern with argument was replaced by an extensive discussion of anthropology, intellectual ecology and cultural diversity.

[4] In a recent paper, Schaffner (2020) criticises Kane's 'rigid form of the Toulmin framework' (p4). He considers the 'quasidebate' suggested by Kane's application of Toulmin to be 'at variance with the ways that scientific writers actually present arguments' (p4).

By the time of 'Cosmopolis' (1990), Toulmin had rejected universal and absolutist truths and advocated a pluralist approach that recognised diverse cosmologies, and recognised the way that reason is shaped by the local, temporal and particular. Toulmin's earlier concern with argumentation moves from a concern with 'written propositions' to that of 'oral utterances' (1990, p. 187) and their contexts.

> '*We may temporarily ("for the purposes of calculation") shelve the contexts of our problems, but, eventually, their complete resolution obliges us to put these calculations back into their larger human frame, with all its concrete features and complexities.' (Toulmin, 1990, p. 201)*

What then, are the implications of Toulmin's later work for an argument-based approach to validity? While Kane does not cite his later work, the seeds of his philosophy are already present in his 1958 work on argument. What Toulmin's later work adds is a shift from a winner takes all approach to reason and argument, to one that recognises the possibility that multiple arguments might legitimately co-exist. That stance is in keeping with democratic and social justice movements that have become especially strong since the 1960's, such as those of feminism, anti-colonialism, and post-modernity. In the context of ILSAs, as globalised assessment programmes, those concerns to situate and recognise diverse arguments and to promote inclusion seem especially appealing.

The later Toulmin therefore suggests a subtle, but important shift of argument-based approach to validity. Instead of seeking to establish the authority of a single IUA, and to rebuff the arguments of others, the task is to identify and reconcile the plural arguments that might legitimately be made about the validity of assessment programmes. The process of assembling validity would be to listen to the arguments of diverse actors, in contexts in which power differentials might otherwise render them unable to make their arguments heard. We move from singular, powerful texts, to oral and written texts and polyphony, and localised evidence.

In the case study that follows we describe the practice of assessment validity as a process of assemblage in the context of the OECD programme PISA for Development. Following the work of Kane, we consider the Interpretation and Use (IUA) arguments put forward, the different types of evidence involved, and the authority of institutional actors to promote or refute those arguments.

Validity in Practice

PISA for Development (PISA-D) explicitly sought to adapt PISA to the diverse needs of low- and middle-income countries, and involved changes to PISA to adapt its *relevance* for those participating countries, their *capacity* to administer the survey and make use of its data, and as a contribution to monitoring the quality of learning outcomes for the UN sustainable development goals (OECD, 2013). The programme has involved a number of key adaptations to PISA, most notably, the extension of the measurement scale to more accurately capture performance at lower levels of ability; a modified contextual questionnaire; and the inclusion of an out-of-school youth survey. Initiated in 2013, PISA-D involved nine countries (Bhutan, Cambodia, Ecuador, Guatemala, Honduras, Panama, Paraguay, Senegal and Zambia). The

programme has been described by the OECD as a pilot, in the sense that it attempts to resolve validity concerns (e.g. relevance, data quality, data use), not only in participating countries, but also in a wider set of participating countries in PISA. For that reason, it provides a useful case-study of the way that validity is assembled in assessment practice. Our case-study presents testimony evidence from key stakeholders, and draws on wider studies of the way that PISA validity arguments, technical standards and procedures are translated and recontextualised in PISA-D (Gorur, Sorensen and Maddox, 2019).

> *'The modern approach to construct validity, defined here as the degree to which evidence and theory support the interpretation of test scores for the proposed uses of tests, is based on the Interpretive/Use Argument pioneered by Michael Kane. The PISA-D test of construct validity begins by defining the purpose of the project's data-collection instruments. PISA-D enhances PISA's cognitive instruments to better measure the lowest student performance in reading, mathematics and science.' (Policy Brief 'PISA for Development Construct Validity', OECD, 2018.)*

> *'In terms of implementing the survey, everything works according to the technical standards and that we have got quality data, and we are very pleased that is the reality, we have that. The Advisory Group meeting in August that reviewed what is called the Data Adjudication Database confirmed data collected […] meets all the technical standards and we have data that is comparable in terms of quality and quantity to a normal PISA cycle.' (Interview in 2018, OECD #50)*

As we can see from the quotation above, the OECD made strong claims early in the programme about its validity – drawing explicitly on the Kane's argument-based approach to validity, and on the PISA technical standards (i.e. validity as a characteristic of the test). Here, we see an example of how human (Kane) and non-human actors (technical standards and procedures) are assembled to establish validity arguments.

As we observed above in relation to Kane's theory of validity, the initial validity claims are never fully settled, but always in the making as new actors are enrolled, and as new evidence and experience becomes available. Hence, the validity arguments of PISA-D remained open-ended and plural, at least up until the completion of the programme in 2018, and the public launch of its data and findings. In the discussion below we consider the 'front stage' and the 'backstage' (Goffman, 1959) events in which PISA-D validity was assembled and constructed. At times, that involved overt public displays of criticism (Auld, Rappleye and Morris, 2019). However, as we argue, much of the PISA-D process took place in more secluded environments ('laboratories' in ANT jargon) of technical meetings as different actors grappled with how to administer the programme, and to ensure its relevance in their own policy contexts (Gorur, Sorensen and Maddox, 2019).

By studying the way that PISA-D validity is assembled as a social practice we observe the numerous local moments, events and places through which validity arguments are developed, presented, and contested (Gorur, Sorensen and Maddox, 2019).

As we discuss below that process involves significant power imbalances which play out in terms of authority of different actors to promote and refute validity arguments. As the work

of Toulmin (1958) suggests, the participants of PISA-D are positioned differently in relation to their respective authority, warrants and institutional backing and geopolitical position.

The process of 'enrolment' (Callon, 1986) into PISA-D involves the allocation of different institutional roles, which impact on the authority attributed to different actors and material artefacts. In PISA-D, those power differentials play out in terms of who is allocated front-stage talking roles, and those who have to make their case in back-stage interactions, for example, during coffee breaks or over a glass of wine at social events.

We adopt a non-hierarchical 'flat ontology' (Latour, 1993, Moll 2002) to describe the moments in which PISA-D validity is assembled. That is, we recognise the multiple locations and actors involved in the assembly, the symmetry between humans and technology as well as the technical and the material. These validity moments include events such as: technical group meetings, decision-making advisory board meetings, associated social events, report writing retreats, policy 'surgeries', choreographed data launches, and the minutiae of WhatsApp groups, weekly conference calls, and frequent email correspondence. Those processes have a material dimension as they enrol durable material artefacts such as guidelines and technical standards, protocols for data collection, a 'Data Adjudication Database', psychometric products such as Item Characteristic Profiles and Differential Item Functioning tables, frameworks, batteries of test items, questionnaires, sampling procedures, and the minutiae of regular quality assurance reports.

Global Rationales and Local Actors: (Re) Assembling Validity in PISA-D

> *'I think of validity as the extent to which the proposed interpretations and uses of test scores are justified. The justification requires conceptual analysis of the coherence and completeness of the claims and empirical analyses of the inferences and assumptions inherent in the claims.' (Kane, 2016, p. 198)*

If validity is assembled by a network of actors over time, and across different geographical locations, then our genealogical task is to piece together a sense of that process from fragments of testimony, documents and technical artefacts, while recognising the potential for diversity of that experience, including the stated purposes, interpretation and uses of test scores presented by different actors.

Auld, Rappleye and Morris (2019) capture that diversity of experience when they discuss the way that arguments about the purpose of PISA-D were developed, negotiated and resisted in Cambodia. They describe extensive lobbying and argument by the OECD and the World Bank, in a combination of (in Goffman's 1959 terms) 'front stage' public policy meetings and conferences, and 'backstage' closed meetings, private phone calls and emails between the OECD, World Bank and the Cambodian Minister of Education.

Addey and Sellar (2019) have shown that there are multiple rationales for national actors to support (but also resist or drop out of) participation in ILSAs, including political, economic, technical and socio-cultural explanations, mostly unrelated to data and education. Hence, the policy rationales for ILSA participation are unlikely to be uniform, as they are informed by both international policy discourses and domestic politics (Steiner-Khamsi 2014; Steiner-

Khamsi & Waldow, 2018; Addey et al, 2017; Addey and Sellar, 2019). Research on rationales for participation has shown how governments participate in ILSAs, particularly in PISA, for a multitude of political, economic, technical and socio-economic reasons. A study of PISA-D in Ecuador and Paraguay showed how participation was driven by rationales as varied as demonstrating global accountability, mobilizing domestic resources, acquiring technical capacity, attracting foreign and domestic investors, accessing the transnational accreditation PISA provides educational systems, and for the legitimacy that the relationship with OECD transfers to governments (Addey 2019). We might therefore ask how those different agendas come together in the PISA-D programme and how they impact on a shared validity argument.

According to Callon (1986) the process of assembling a network of actors around a common goal begins with a process of '*problematisation*' in which the actors identify a set of interests that hold the network together. For the purposes of this paper, we make a distinction between the authorised validity arguments in formal publications and conference presentations, and those informal, unauthorised arguments that take place, for example, in informal conversations over coffee or beer, email and WhatsApp correspondence. For reasons that we highlight below, we note that more and less powerful actors take part in *both* sets of interaction. Challenges to validity arguments may take place more frequently as informal, 'hidden transcripts' (Scott 1992) rather than overt public rebuttals. However, we can also observe the presence of unauthorised narratives practised by influential actors as Auld, Rappleye & Morris (2019) observed, as well as overt, choreographed public displays of support by less powerful actors as we describe below.

The authorised validity arguments of PISA-D reveal a consistency of themes in published documents (see, Bloem, 2013, 2015; OECD 2016, 2018), stating that PISA-D will:

- (VA1) Enable low- and middle-income countries to produce and use high quality assessment data that is compliant with the PISA technical standards and scales;
- (VA2) Support participating countries to increase their capacity to administer large-scale assessments, including sampling;
- (VA3) Adapt PISA test instruments (item difficulty, relevance, contextual data) for use in low- and middle-income countries and to capture more data on lower performance levels, and;
- (VA4) create a more representative sample with the inclusion of ('Strand C') assessment of out of school youths.

For the OECD, the primary purpose and rationale for PISA-D was to operate as a vehicle for the global expansion of PISA (Bloem, 2013):

> '*As more countries joined PISA, it became apparent that the design and implementation models for the assessment needed to evolve to successfully cater to a larger and more diverse set of countries, including a growing number of middle- and low-income countries.*' (OECD, PISA-D Website)

Bloem (2013, 2015) argued that the expansion of PISA was limited by the costs of participation; low levels of institutional capacity; difficulties in sampling that undermine the validity of international comparison; and a lack of granular data on performance at lower

levels of the achievement scale. Therefore, if PISA-D participating countries were able to demonstrate capacity to comply with PISA technical standards, and if the OECD were able to make the necessary accommodations, it would 'validate' PISA for use in an expanded set of low- and middle-income countries.

As Bloem (2013) noted, several countries participated in one round of PISA, and dropped out before subsequent rounds, limiting their ability to use PISA data to examine performance trends. A key moment in the '*problematization*' of PISA-D was the participation and subsequent rejection of the data from India for its 2012 results on its participating states of Tamil Nadu and Himal Pradesh. While the OECD had argued that those Indian States 'did not meet the PISA standards for student sampling' (Bloem, 2013), the Indian Government rejected the results on the grounds that the test was not suitable for India's socio-cultural diversity (Chakrabarty, Molstad, Feng and Pettersson, 2019), and dropped out of two subsequent cycles. The Indian press was highly critical of India's performance, with headlines such as '*Indian students fare poorly in international evaluation test*' (the Hindu), and '*Indian Students rank 2nd last in global test*' (Times of India). The Indian PISA-shock, was highlighted by an OECD member of staff in an interview with ATUHOR 1:

> *"Tamil Nadu and Himachal Pradesh as the two states because they are the highest performing states in India, but yet even the highest performing states in India were performing the lowest in PISA. And so MHRD [Ministry of Human Resource Development] built a story that 'This was not an appropriate test, look there is a question here about taking money out of an ATM, our children have never seen an ATM, they don't know what an ATM is'. And so [at the OECD we said] 'Look, we should have been there, we should have been helping them to unpack these results, and to present these results, instead we have got this situation.' And that all informed what [the OECD is] doing here in terms of designing PISA-D*. […] *I think the Indian experience is central, for how you prepare a country to come into PISA, how you then do more than just enable the country to take part, you have to be with them all the way through, in terms of presenting to their population why they are in PISA, helping them to understand the results, and to present those results, and managing the fall out.'* (Interview in 2015, OECD #30).

That problematization, as the desire to enable the full globalisation of PISA and to involve low- and middle-income countries was articulated by the OECD in documents and interviews, and by national level actors. Responding to that validity challenge the OECD repositioned its PISA-D arguments and its partnerships with UNESCO, and UNICEF to respond to the globalising discourse of the UN Sustainable Development Goal of Education (SDG4), with its focus on the measurement of learning outcomes and performance levels (OECD 2014[5], Addey 2017; Auld, Rappleye and Morris, 2019).

Validity as Technical Standards

The nine countries that completed the first PISA-D cycle in 2019 had to comply with the technical standards and frameworks established by the OECD in PISA. Five other countries - Mongolia, Pakistan (Punjab Province), Rwanda, Sri Lanka and Tanzania - were initially interested to participate in PISA-D, and for various (largely non-financial) reasons decided not. Mongolia is currently due to participate in PISA 2020 (Per com. 2020).

One explanation for countries dropping out in the development stage was the formation of what Callon (1986) has called '*obligatory passage points*'. In the case of PISA-D, those passage points involved compliance with the technical standards and frameworks established by PISA - in other words, agreeing to the single voice that would validly represent their educational countries in PISA data. Hence, we can see a shift in the PISA-D discourse on validity from its early emphasis on dialogue and validity arguments in Kane's sense (OECD, 2018; 2020) towards one of validity as a characteristic of the test, which is contrary to what Kane proposes.

> '*The PISA-D test of construct validity begins by defining the purpose of the project's data-collection instruments. PISA-D enhances PISA's cognitive instruments to better measure the lowest student performance in reading, mathematics and science; it enhances contextual data instruments to better capture the diverse contexts in middle- and low-income countries.*' (OECD, 2018)

> '*PISA's technical standards were applied at every stage of the project. The main survey data collection is subject to a strict adjudication process, particularly for the sampling and translation/adaptation parts of the implementation.*' (OECD, 2018)

We can see not only that those technical standards are significant as claims to validity, but that it was necessary to be compliant with those standards to participate in PISA-D. The initial discussions with governments about PISA-D participation included more than the nine participating countries: many were initially interested in a PISA that would value diverse educational contexts and dropped out upon understanding the instruments would be sensitive only in so far as comparability with the main PISA would not be compromised. On the other hand, the nine participating countries valued the comparability with PISA over the instruments' capacity to validly capture their contexts.

How did the successful participating countries negotiate with those technical passage points, and how did they impact on the assemblage of validity arguments? In Addey's interviews with national level participants in PISA-D we get a sense of the way that actors reconciled those demands with their wider rationales in participation:

> *'It's as if you have all the freedom in world but with limits, you cannot play that much and these limits are not only OECD-imposed limits, the countries impose them too. You take part in PISA because you want PISA. […] It cannot reflect my reality much, but I can sacrifice that. [...] And yes, many things countries wanted were omitted. At the end of the day you need to decide what you*

prefer. And what was preferred was to have a stronger tie with PISA.' (Interview in 2016, PISA-D Country *#NoNumber*)

The interview extract gives a sense of the way in which actors from the participating countries traded off their rationales for participation and their own validity concerns with the technical demands of the programme. Some field-based technical and administrative practices must be translated to match with the context of participating countries (Gorur, Sorensen and Maddox, 2019), but in daily practice, the validity of PISA-D was established through technical standardisation.

Assembling Validity as Technical Practice

In the testimony above we get a sense of the trade-off that some actors had to make between their validity concerns and obligations of technical standardisation. When we look at the technical practice of PISA-D, we can see how that impacted on what Callon (1986) describes as '*Enrolment*', i.e. the institutional roles that are assigned to actors in the PISA-D network and how that impacted on their authority and practice. The requirements for compliance with technical standards and the routines of technical meetings can be viewed as a way of the OECD enrolling non-human actors (technical processes, statistical artefacts, sampling procedures, reporting forms) to regulate the roles and behaviours of human actors:

'The ideal scenario which we are aiming for, is that the countries successfully collect the data in accordance with technical standards. In terms of implementing the survey, everything works according to the technical standards and that we have got quality data, and we are very pleased that is the reality, we have that.' (Interview in 2018, OECD, #50).

However, that sense of an 'ideal scenario' was not necessarily shared by national level actors:

'The specialists from these Organizations and companies present their instruments and there is not much time. And one is just concerned with understanding what they are saying, and sometimes with trying to adapt, but in most cases you do not question the content.' (Interview in 2016, PISA-D country #NoNumber)

In practice, this latter extract shows that the dynamics of PISA-D meant national programme managers (and all local experts invited to share their expertise) had limited space to raise validity concerns in formal technical meetings, while their activities outside of those meetings were regulated though a SharePoint calendar sending automated email reminders about achievement of necessary tasks and fast-approaching deadlines. In formal technical meetings, there is little opportunity for participants to raise validity concerns. Each meeting is minuted and is completed with a record that the participants agree on the activities and decisions (often the discussed and agreed decisions are pre-written).

'There is really a lot of asymmetry, it is not a shared building process, it is not that we are building the study together. The background questionnaires are where I noticed more openness, but not in the rest. There was a lot of distance in the technical management, between the specialists, who are the very top,

and the countries, and so the reality is it is not a dialogue.' (Interview in 2016, PISA-D country #NoNumber)

'Well, I want PISA and if I am in PISA-D it's because I want to compare myself with other countries that are part of PISA and not just to compare with you lot. And so yes, I do need those questions there, yes, I do need the socio-economic index fits with the PISA one, even if it does not reflect my reality. It cannot reflect my reality much, but I can sacrifice that.' (Interview in 2016, PISA-D country #NoNumber)

We get a sense from these testimony extracts of the way that socio-material practice took place in the secluded context of the PISA-D 'laboratory'. We see how the assemblage functions in a way that allows certain allies to exercise power. As Addey observed at PISA-D meetings, the sophisticated nature of assessment methodologies, which are kept in the hands of few actors (i.e. the private contractors), is a strategy used to settle disputes. Other ways in which relations of power are settled include the use of English as the language of communication which not all participants are sufficiently comfortable in to be able to defend non-conforming views. Interviewees also describe the lack of a voting system during meetings as a way to 'decree' consensus, thus silencing the plural interpretations and uses of multiple actors in PISA-D.

The emphasis on validity as technical standards is also evident in the PISA 2021 'National Project Manager Manual' (OECD, 2019) with its emphasis on necessary validity checks and reporting. For PISA-D at least, there appears to have been little front stage space for NPMs to raise and discuss wider validity questions, for example, the relevance and universal application of test items, or the appropriateness of contextual questionnaire content. The potential for the kinds of heated debates and discussions that one might expect in scientific projects (Callon et al, 2011; Latour 1987) were managed and resolved – pushed into the informal context of discussions in coffee breaks, WhatsApp chats and social activities, leaving little space for plural interpretations and uses to be included.

Choreography and Display

Much of the activities of PISA-D took place in the seclusion and technical routines of the laboratory or were managed remotely by reporting processes. However, for the assembled allies to return their work as data and reports to the world, they had to adopt strategies to counteract the 'fierce adversaries' who may not be supportive of their work (Addey and Gorur, forthcoming). Hence, significant amounts of work took place in the run up to the publication of the PISA-D results in 2019, that involved the enrolment of a wider network of allies and the choreography and display of findings and their validity in public events and reports:

'We certainly feel that if this report is going to have any impact, those key people who are responsible for taking it forward, need to be brought on board, what we don't want is a situation where the report is presented and then people start publicly disagreeing with findings, with implications, that would not be helpful.' (Interview in 2015, OECD #30)

As this extract shows, the process of defeating adversaries remained a concern throughout the entire development of PISA-D. This led the OECD and partners to carry out communication campaigns to inform all educational stakeholders about the value of PISA-D and by organising high level events with the most senior OECD officials and education ministers.

> '*We ran 'surgeries' for each country. So we spent about three hours with each country reviewing particularly chapter 6 the policy implications and suggested ways forward and those reports will have three or four very clear policy messages arising from the analysis of the data. […] And then having gone through each of the three or four key policy messages and identified who it is that is going to own it, we then worked with them about how they can develop a plan to go to that person and have some structured discussion about how to take this forward, firstly to inform them of the results, all of it in confidence as it in under embargo before the launch, and get the input of those individuals into chapter 6 of the report, so what is elaborated there, as a suggested way forward is something the owners of this, whether it is curriculum, whether it is teacher education, administration, resources, whatever it is, that those who are responsible own it, they own it, they say 'Yes, okay we agree with these findings, we agree with the implications, we agree with what you are saying is the way forward, we support this proposal.' (Interview in 2018, OECD #50)*

The choreography and display of the reporting processes involves attention to the fine detail of launch ceremonies, the reporting process, and the presentation of material artefacts - glossy reports, press releases and power-point presentations.

> *'That then becomes our dream ideal, scenario if you like. When the national report is launched in the week beginning 10th December [2018], it is the minister who is presenting the report to the population, to the media, to the stakeholders, everyone who is waiting to have this report. The minister of education is presenting it, and then when it comes to the policy implications, those bits of the Ministry or of the government which are responsible for each of those areas, are standing up and saying yes, absolutely, this is the way forward, this is what we must do, and that leads to a broad consensus in terms of how to move forward in terms of an evidence-based policy dialogue in education'. (Interview in 2018, OECD #50)*

The staff at the relevant 'bits of the Ministry' are asked to demonstrate their acceptance of the PISA-D validity at high level, public events which are a choreography on a global stage meant to not leave space for philosophical doubts on the validity of such an international comparative endeavour. In other words, this assembling of allies who stand up to show their agreement can be viewed as a public demonstration of PISA-D's validity. This, we argue, is equally a practice of validation which current theories of validity would not recognize.

The UK launch of PISA-D at Central Hall, Westminster, London (September 2019) illustrated a careful choreography and display, with formal speaking roles assigned to influential actors including the OECD, DFID and the World Bank. In the opening session, Andreas Schleicher

(Director of the Education and Skills Directorate at the OECD), whose presence creates a 'global stage show' that PISA-D countries sought during their participation (Addey 2019), began by saying that '*Everyone supports PISA for development because it has been so successful*'. He noted that the initial aim of PISA-D was to examine '*to what extent these instruments that we have are valid in a wider range of contexts'.* He noted that *'We ended up with nine countries that all met the PISA technical standards in full, and it was almost you know, a miracle' [he presented data on 7 countries, without data on Panama and Bhutan]*, and argued that it has shown, that '*we have now instruments that can extend throughout the world, we have proven that'*. This public choreography and display of validity, which was presented to actors who had been involved in the PISA-D validation practices for many years, sought to demonstrate that the validity of PISA-D had been assembled in a durable manner, settled and agreed upon by all allies although this paper's main argument shows this was achieved through mechanisms that displaced plural interpretations and uses.

Discussion

We have contrasted two opposing perspectives on validity. Each draws out different strengths and weaknesses of Kane's argument-based approach. The first perspective draws on the early work of Toulmin, and considers the purpose of validation arguments as to establish and promote what we might call an 'original position' (Rawls, 1971). That is, in the sense used by John Rawls, with its emphasis on explicating a coherent and logical argument that is indifferent to the practical contexts and concerns in which assessment programmes are applied. The contrasting position that we have developed in this paper is informed by Toulmin's later work. Specifically, the function of validity is one of assemblage – to recognise and bring together the legitimate arguments, evidence and perspectives of diverse actors and stakeholders who participate in assessment programmes.

Messick's (1989, 1995) approaches to validity were grounded in construct judgments. He identified the end user as being in the best position to evaluate the meaning of scores obtained in a given context. In addition, the end user is in the best position to determine the extent to which the intended meaning of those scores may have been eroded by contaminating influences within that context. In contrast, as we have argued and demonstrated, it is right that the various actors should seek to understand and produce the meaning of test scores, and that the contextual factors inform rather than erode test validity. Kane on the other hand, emphasizes the need to produce a validation argument for whatever test score interpretation and use one intends to defend, with the centerpiece being the interpretive/use argument justification. The need to seek and produce construct meaning does not feature prominently in Kane's work. In contrast, this pursuit is central to our argument. As such, one may use the tensions among the animate or inanimate actors in ILSAs to produce and illustrate the valid interpretations and uses of test scores.

Although validity remains grounded in score interpretation and use, its theoretical foundations as well as scope remain fundamental and contested (Zumbo and Hubley, 2016). Since the publication of Messick's groundbreaking review of validity (Messick, 1989), the field of testing has been calling out for an expanded evidential basis for test validation (Zumbo, 2017). We respond by drawing the practical into the abstract. As such, the apparent

vicissitudes of test context and actors enrich rather than pollute the integrity of the assembled validity argument.

To return to our case-study, if the purpose of PISA-D was simply to demonstrate the OECD's ability to apply PISA technical instruments on a global scale whilst claiming its validity and comparability with the main PISA, then one might say that it has been a success. As Andreas Schleicher said, '*we have proven that*'. Furthermore, with India re-joining PISA in the 2021 cycle, and the ever-growing participation from low- and middle-income countries (including 8 PISA-D countries in either the 2021 or 2024 cycle), the OECD seems to have resolved the problem that had initially motivated the PISA-D programme. As Andreas Scheichler noted at the London data ceremony, the OECD and partners learned a lot from PISA-D, for example about the involvement of a diverse set of countries and actors, and on testing at lower performance levels. The OECD has also successfully presented the rationale for global expansion of PISA in terms of developing a valid global measurement of SDG learning outcomes.

However, we would rebuff the idea that country compliance with a set of technical instructions and manuals is sufficient to establish the warrants for the claims made by the testing programme. Technical standards are certainly an important feature of a testing programme, but we would argue as Kane does, that '*validity is not a property of the test*' (Kane, 2016, p198).

By conducting PISA-D, the OECD was successfully able to mobilize the resources, and assemble, at least temporarily as ANT reminds us, a network of actors, institutions and material artefacts around a shared interest and '*problematization*' (Callon, 1986). Those allies included national actors and test participants in the participating countries, and a wide set of global actors from development organisations (e.g. the World Bank, UNICEF, UNESCO), technical partners, and businesses. Nevertheless, as we noted, five countries who had initially planned to join the programme decided not to continue. For them, the problematisation, resources and intended benefits of participation were not sufficient to maintain their involvement, and to navigate the '*obligatory passage points*' (Callon, 1986). They opted out from assembling the validity of PISA-D - always an option in the temporary nature of social projects that require numerous allies and resources.

The practices of PISA-D reveal asymmetries of power and representation that speak to our discussion of Kane's argument-based approach to validity, and the different perspectives offered by the early and late philosophy of Toulmin. We do not contend that the acts of resistance, debate and persuasion observed by Auld, Rappleye and Morris (2019) are somehow inimical to programme validity, to the extent that they are anticipated by Kane's (2013) argument-based approach (i.e. establishing backing for warrants, and the space of challenge and rebuttal). However, since the key feature that makes PISA-D different to PISA is the extent of cultural, economic and educational diversity present in low- and middle-income countries (Bloem, 2013), it would seem appropriate to recognise and incorporate that plurality in the validity argument and validation practices. That would mean opening a space for discussion and representation of validity arguments and evidence that extends beyond

compliance with technical standards and procedures, and in which all participants are not required to agree with and support all the findings.

Conclusion

This paper has discussed the implications of Kane's argument-approach to the validity of assessment contexts such as ILSAs, with their diverse networks of actors and national contexts. As we have argued, in those situations, while it may be advantageous to promote an initial clarity of validity arguments as Kane and others have suggested, in practice that means that, there is a need to recognise and reconcile the plural interpretations and uses of those diverse actors and their assessment contexts. We have suggested that Kane's argument-based approach can accommodate those demands if it extends its philosophical foundations from Toulmin's early work on 'The Uses of Argument', to Toulmin's later work on cosmopolitanism and diversity.

V*alidity in theory* may seek to 'explicate validity' (as Kane 2016 argues) with precision of argument about the intended interpretation, uses and consequences of assessment. However, *validity in practice* is 'assembled' over time and place (in the sense implied by ANT) by a network of actors with shared rationales and what may be diverse interests and intentions about the interpretation, uses and consequences of assessment. As a result, one of the tasks of an argument-based approach to validity is to create a democratic space in which legitimately diverse arguments and intentions (in the sense implied by Toulmin's later work) can be recognised, considered, assembled and displayed.

Those obligations to recognise and validate diversity are especially significant in the context of ILSAs, which involve heterogeneous socio-cultural, economic and educational contexts, multiple rationales for participation and plural policy contexts for the interpretation and use of assessment data. This obligation is all the more significant when the global education agenda is pledged as a goal that was collectively put forward and influential international organizations and working groups claim that its benchmarking requires the use of a single metric for all educational systems.

In the case study of PISA-D, we have seen that compliance with the PISA technical standards was emphasised as if validity was a property of the test (in contradiction of Kane's 2013 argument), and how the programme has attempted to present 'front-stage' (Goffman) validity arguments in public occasions and technical meetings as if the diverse set of actors speak with a single voice, and to push debate and dissent into informal, 'back-stage' arenas. An alternative approach, inspired by Toulmin's later philosophy, would be to see the potential for a legitimate diversity of arguments as an opportunity to actively assemble validity in a way that recognises and reflects the contexts of diverse participants. In that way, the validation practices that are developed might also reflect that diversity in the different warrants and evidence involved and expand the use of data on contextual differences and response processes across those contexts.

References

Addey, C. (2017). Golden relics & historical standards: how the OECD is expanding global education governance through PISA for Development, Critical Studies in Education, 58 (3): 600, 311–325, DOI: 10.1080/17508487.2017.1352006

Addey, C. (2019a). Researching inside the international testing machine: PISA parties, midnight emails & red shoes. In, Maddox, B, *International Large-Scale Assessments in Education. Insider Research Perspectives*. London: Bloomsbury. Pages 13 – 29.

Addey, C. (2019b). The appeal of PISA for Development in Ecuador and Paraguay: Theorising and applying the global ritual of belonging. *Compare: A Journal of Comparative and International Education*. DOI: 10.1080/03057925.2019.1623653

Addey, C. & Sellar, S. (2017). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. In A. Verger, M. Novelli, & H.K Altinyelken (eds) Global education policy and international development. Second edition.

Addey, C., & Sellar S. (2019). 'Is it worth it? Rationales for (Non)participation in international large-scale learning assessments'. *Education Research and Foresight Working Papers Series*, No. 24. Paris: UNESCO Publishing.

Addey, C., S. Sellar, G. Steiner-Khamsi, B. Lingard & A. Verger. (2017). The rise of international large-scale assessments and rationales for participation, *Compare: A Journal of Comparative and International Education*, 47 (3): 434–452, DOI: 10.1080/03057925.2017.1301399

Addey, C. and Gorur, R. (Forthcoming). *Translating PISA, Translating the World*. Comparative Education.

Auld, E., Rappleye, J., & Morris, P. (2019). PISA for Development: How the OECD and World Bank shaped education governance post-2015. *Comparative Education*, 55(2), 197-219.

Bachman, L. F. (2005). Building a test use argument. *Language Assessment Quarterly*, 2, 1–34.

Cronbach, L. J. (1988). Five perspectives on validity argument. In, H. Wainer & H. Braun (Eds.), *Test validity.* Hillsdale, NJ: Lawrence Erlbaum. (pp. 3-17).

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? Educational Measurement: Issues and Practice, 29(1), 3-13.

Deleuze, G, & Guattari, F. (1983). *Anti-Oedipus*. Translated by Brian Massumi. Minneapolis: University of Minnesota Press.

House, E. (1980). *Evaluating with Validity*. Beverly Hills, Sage.

House, E. R. (1990). Trends in Evaluation. Educational Researcher. Vol. 19, (3), pp. 24-28

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319–342. Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21, 31–41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2, 135–170.

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement*, 4th edn. (pp.18–64). Washington, DC: American Council on Education/Praeger.

Kane, M. T. (2013). 'Validating the Interpretations and Uses of Test Scores'. *Journal of Educational Measurement*, Vol. 50, No. 1, Special Issue on Validity (Spring 2013), pp. 1-73

Kunnan (2010) Test fairness and Toulmin's argument structure. *Language Testing*. 27 (2), 183-189.

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge MA and London: Harvard University Press.

Latour, B. (1993). *The Pasteurization of France*. Harvard University Press. Cambridge M.A.

Latour, B. (1987). *Science in Action*. Milton Keynes, Open University Press.

Latour, B. (2005) *Reassembling the Social: an introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory Life*. Princeton: Princeton University Press.

Latour. L. (1992). 'Where are the missing masses: The sociology of a few mundane artefacts'. In Wiebe E. Bijker and John Law, eds., *Shaping Technology/Building Society: Studies in Sociotechnical Change* (Cambridge, Mass.: MIT Press, 1992), pp. 225–258.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement*, 33, 379–416.

Moll, A. (2002) *The Body Multiple.* USA: Duke University Press.

OECD (2016). *PISA for Development* (Brochure). OECD, Paris.

OECD (2018). 'PISA for Development construct validity', *PISA-D Policy Brief*, number 24, August 2018. OECD, Paris.

OECD (2014). *The OECD's contribution on education to the post-2015 framework: PISA for development*. Paris: OECD Publishing.

OECD (2020). *PISA Technical Report*. Paris: OECD Publishing.

Ozga, J. (2012). Assessing PISA. *European Educational Research Journal*. 11 (2). 166-171.

Rawls, J. (1971). *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Schaffner, K. F. (2020). 'A Comparison of Two Neurobiological Models of Fear and Anxiety: A "Construct Validity" Application?' *Perspectives on Psychological Science*. pp1-14.

Scot, J. (1990) *Domination and the Arts of Resistance*. New Haven: Yale University Press.

Sellar, S., Lingard, B., Rutkowski, D., and Takayama, K. (2018). Student preparation for Large-Scale Assessments: A comparative analysis. In, B. Maddox (Ed). *International Large-Scale Assessments in Education: Insider Research Perspectives.*

Steiner-Khamsi, G. (2014). 'Cross-national policy borrowing: Understanding reception and translation'. *Asia Pacific Journal of Education*, 34 (2), 153-167.

Steiner-Khamsi, G. & Waldow, F. (2018). 'PISA for scandalisation, PISA for projection: the use of international large-scale assessments in education policy making – an introduction'. *Globalisation, Societies and Education*. Vol. 16 (5), pp 557–565.

Toulmin, S. (1958). *The uses of argument.* Cambridge, UK: Cambridge University Press.

Toulmin, S. (1972). *Human understanding*. Princeton, NJ: Princeton University Press.

Toulmin, S. (1990). *Cosmopolis: The Hidden Agenda of Modernity*. Chicago: The University of Chicago Press.

Zumbo, B.D., & Hubley, A.M. (2016). Bringing Consequences and Side Effects of Testing and Assessment to the Foreground. *Assessment in Education: Principles, Policy & Practice*, 23, 299–303.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics*, Vol. 26: Psychometrics (pp. 45–79). Amsterdam, Netherlands: Elsevier Science B.V.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte, NC: Information Age.

Zumbo, B.D. (2017). Trending Away From Routine Procedures, Towards an Ecologically Informed 'In Vivo' View of Validation Practices. *Measurement: Interdisciplinary Research and Perspectives*, 15:3-4, 137-139.