TALK AND GESTURE AS PROCESS DATA

Forthcoming in *Measurement: Interdisciplinary Research and Perspectives*

Bryan Maddox, University of East Anglia, Norwich, UK. B.maddox@uea.ac.uk

This paper discusses talk and gesture as a neglected source of data on interaction and response processes in testing situations (Maddox, 2015, Maddox and Zumbo, 2017). The significance of the paper is the growing use of various sources of 'process data' in computer-based testing (Ercikan and Pellegrino (Eds.) 2017; Zumbo and Hubley (Eds.) 2017). That involves the use of process data on assessment response processes as an additional source of information about test performance and validity (Kane & Mislevy, 2017).

Process data in assessment are studied using a range of techniques, including: think aloud protocols; eye tracking studies, video-ethnography and computer-generated log files (e.g. Goldhammer, Naumann and Greiff, 2015; Lee and Haberman, 2016; Greiff et al. 2016; Orange, Gorin, Jia and Kerr, 2017; Maddox and Zumbo, 2017). They can be thought of conceptually as '.. *the mechanisms that underlie what people do, think, of feel, when interacting with, and responding to, the item or task and are responsible for generating observed test score variation*' (Hubley and Zumbo, 2017, p2). Process data are increasingly used in assessment as an extension of the test (e.g. Greiff et al, 2016), as an extension of validity (e.g. Newton, 2016; Kane and Mislevy, 2017), and as a source of performance data for

'next generation' assessment design (e.g. Orange, Gorin, Jia and Kerr, 2017). As they are integrated into digital processes (e.g. in predictive analytics and machine learning), process data of various kinds also impact on the evaluation and management of education systems, schools and teachers (Cope and Kalantzis, 2016; Williamson, 2016).

This paper focuses on a particular source of process data, namely, the information that can be gleaned from talk and gesture in real-life testing situations (also Maddox and Zumbo, 2017). As the paper argues, such information can provide insights into areas such as respondent engagement, interviewer effects, and ecological impacts on test performance. Yet, talk and gesture is not captured by computer-generated log files on response times and key strokes. As a result, there is a risk that the integration of process data into 'next generation' assessment design may be based on faulty or partial information. Information on talk and gesture is especially important in international large-scale assessments (ILSAs), as it has potential to identify and explain variation in the ways that tests are received and performed across diverse socio-economic and cultural settings (Maddox, 2014, 2015).

In previous eras, data on talk in tests was likely to be regarded as anecdotal accounts, an 'intrusion' of construct irrelevant contextual noise or as a threat to validity (see Messick, 1989, McNamara and Roever, 2006). Even researchers such as McNamara and Roever (2006) with a commitment to the study of language in assessment regarded the integration of data on face-to-face interaction as a challenging and perhaps insurmountable task:

'*The social nature of interaction is difficult to capture in a testing setting, and what can most reliably and practically be captured are isolated pieces of a much larger mosaic*' (McNamara and Roever, 2006, p78).

What's new in assessment is the potential for such 'isolated' pieces of data on testing situations to be systematically integrated into the larger-mosaic that McNamara and Roever describe. The 'micro-analytic' approach set out in this paper takes advantage of the affordances of digital data in assessment to locate small-scale observations on talk and gesture with precision in the analysis of larger-scale assessment data. The use of digital video recording enables observations of assessment response processes and interaction to be synchronised with log file data, and to be slowed and examined down to the millisecond. That precision enables respondent and interviewer talk and gesture to be located within a sequence of test items (i.e. the unfolding temporal character of the test as an event), and within response processes on specific test items. The approach collapses conventional distinctions between micro and macro, and between quantitative and qualitative, creating new spaces for multi-disciplinary (or even post-disciplinary) enquiry (Maddox, Zumbo, Tay Lim and Qu, 2015). The paper presents examples of talk in tests, and then discusses the significance and place of such data in validity practice, and in wider considerations of how we integrate process data into test design.

The 'micro-analytic' approach discussed in this paper has its disciplinary roots in the study of face-to-face communication in linguistic anthropology and conversation analysis, and is most notably articulated in the work of Erving Goffman:

> '*My concern over the years has been to promote acceptance of this face-to-face domain as an analytically viable one – a domain which might be titled, for want of any happy name, the interaction order – a domain whose preferred method of study is microanalysis*' (Goffman,1983a, p. 2).

As Goffman (1963, 1981, 1983a) argued, social situations are subject to distinctive cultural rules that inform the role of actors, their conduct and obligations. These are observable though the study of every-day talk. This paper follow's Goffman's lead to study rea-life testing situations as distinctive social occasions (see McNamara, 1997; Maddox, 2015). The application of these perspectives is evidence of a trend away from routine procedures of validation practice (Zumbo and Chan, 2014), towards more eclectic and improvised study of assessment performance that integrates sources of environmental feedback from the testing situation (author names removed, 2017). This supports a more sensitive distinction between ecological noise and signal to inform decisions about what counts as relevant sources of variance (Maddox, 2015).

The purpose of the paper is not to introduce 'interaction' as a spoiler, as if its presence should diminish the validity of standardised tests. It is not an attempt to introduce epistemological anarchy or to bolster claims about test fairness or item bias – although though small-scale studies of the testing situation are a useful source of information on how well tests travel, and how diverse populations receive standardised tests (Maddox, 2014, Maddox, Zumbo, Tay-Lim and Qu, 2015). Rather, following Zumbo's (2007) argument in his paper on 'Three generations of DIF analysis', the paper aims to use data on talk and gesture in tests to identify and

discuss sources of variation in assessment performance that can be observed in the testing situation but are not necessarily relevant to the trait or construct under investigation (Zumbo, 2007, p229).

The presence of a computer as a material artefact in the testing situation has not generated as much attention as it deserves. The extensive literature on 'mode effects' has sought to examine the equivalence of test scores in paper-based (PBT) and computer based testing (CBT). That includes some discussion of how performance may be related to people's familiarity with computers and their cultural orientation toward their use in testing (Russell, Goldberg and O'Connor, 2003; Jeong, 2012; Jerrim, 2016). Those issues are evident in international testing such as the OECD Programme for International Assessment of Adult Competencies (PIAAC), which highlights the differences in the prior experience of populations regarding their use of computers (OECD 2016b). 'Mode effects' introduce interactive, subjective and affective sources of variation that extend beyond score equivalence and experience, for example, with implications for measurement quality and administration. This is especially relevant for test items that make use of the affordances of computer-based design (e.g. multi-stage and adaptive items, the use of different media). In those situations (such as PIAAC items on 'Problem Solving in Technology Rich Environments'), tests of equivalence are redundant as there is no paper-based alternative.

The study of talk and gesture in tests provides insights into the intimate character of computer-based testing as it captures 'face-to-face' interaction between the interviewer and respondent, and 'face-to-screen' interaction between humans and computers. The structure of this paper is as follows. The next part introduces

the subject of talk in tests, and the theoretical and methodological orientation of the paper. Transcripts based on video-ethnographic observations of computer-based assessment are presented from the 'PIAAC' assessment in Slovenia. These focus on three themes – the first being interactive 'repair' of respondent face or dignity at the start of an assessment process. Secondly, on how interviewers deal with respondent fatigue mid-way through the assessment. Thirdly, the paper discusses respondent engagement, and considers how 'micro-analytic' data can be integrated into validation practice.

Talk in Tests

In studying talk and gesture in real-life testing situations, the paper applies established theory and methods in Linguistic Anthropology (Duranti, 1997), Conversation Analysis (Sacks, 1994) and Gesture Studies (McNeil 1985; Kendon, 2007). These involve the study of individual sequences of talk and gesture obtained from real-life social situations (Schegloff and Sacks 1973; Goffman, 1981; Schegloff, 1988). Gesture Studies add 'environmentally coupled' gesture – such as pointing and gaze in face-to-face communication (Goodwin 2007). As Goodwin argues, participants orientate themselves through talk and gesture to a shared focal activity. The study of talk and gesture therefore reveals the affective 'stance' of participants – their orientation to a task or topic as it is communicated in sequences of interaction (Du Bois and Karkkainen, 2012; Goodwin, Cekaite and Goodwin, 2012). In this paper, that shared focal activity is the testing situation.

It goes without saying that many tests take place in silence – though even in those cases, some participants (such as test administrators) have ratified talking roles. Nevertheless, there are numerous cases where talk is an essential component

and performative dimension of testing (see McNamara 1997, Shohamy, 1993). These include oral examinations involving demonstrations of competence such as language testing where talk is integral to the assessment process, and other types of assessment such as those discussed in this paper, where talk is present but is not the main 'mode' of assessment. Whether it is the primary mode or assessment, of a more peripheral role in test administration, the study of talk in tests provides a window into assessment response processes.

In this paper we discover talk in an unlikely place – large-scale, computer-based assessment. The OECD 'PIAAC' assesses the skills of adults in Literacy, Numeracy and Problem Solving in Technology Rich Environments (OECD, 2016a, 2016b). It offers respondents paper-based or computer-based versions depending on their experience and abilities. The computer-based version discussed that is the focus of this paper includes the module on problem solving that is not present in the paper-based version. PIAAC assessment is delivered in respondents' home and is administered by interviewers. The informal, 'low stakes', and household character of the assessment adds to the character of the PIAAC assessment as a social occasion meaning that the interviewer (and sometimes other family members) are sometimes drawn into the process to answer procedural questions or to encourage the respondent.

The PIAAC testing situation involves three-way human-computer-human interaction with distinctive and collaborative roles for the interviewer and computer. Talk is not essential to the computer-based mode of assessment, but the mere presence of the interviewer in the testing situation invites discussion - typically on procedural questions, respondent fatigue, about the duration of the assessment, or

self-evaluations of performance. A respondent who is struggling might ask searching questions, or make facial gestures as 'response cries' (Goffman, 1981) to elicit the support of the interviewer (see below).

In the respondents' home social roles, obligations and authority of interviewer, computer and respondent are accommodated and negotiated. Unlike a testing centre, a household is 'multi-purpose'. The testing situation may have to accommodate other things going on in the household – mealtimes, eating, drinking and childcare. It is not an ideal location. The telephone may ring. The respondent's siblings, relatives or partner may want to participate in the assessment or to comment about what is going on. The household provides a useful setting to observe human-computer interaction, and to obtain information about the test performance.

METHODOLOGY

The transcripts presented below are based on video-ethnographic observations of real-life PIAAC computer-based assessment conducted by the author. The use of video recording is an established technique in linguistic anthropology, enabling detailed analysis of in-situ talk and gesture in every-day events (Duranti, 1997). This supports the development of the transcripts of interviewer-respondent interaction that are presented below. The video recordings capture more than an observer could take in, such as changes in facial expression, gaze, and body posture that take place in a fraction of a second (Monkaresi et al., 2016).

The aim of video-ethnographic methods is similar to think-aloud protocols (cognitive interviews), since both attempt to observe normal assessment performance in a non-invasive manner (e.g. Ericsson and Simon, 1993; Pepper et al,

2016). The key difference is that while think-aloud protocols attempt to capture the inner speech of respondents, video-ethnographic observation aims to study the interactive and performative role of talk and gesture in testing situations. I.e. bringing the interviewer, and the context of the test into the frame.

The PIAAC assessment in Slovenia involved 5,331 respondents and of those 74.5% took the computer-based version (OECD, 2016b). The data presented in this paper involved a small number of randomly assigned video-recorded assessment events (*n*=10), the unit of analysis being individual assessments and short sequences of talk and gesture. There is no sense in which this is a representative sample. Rather, it is used to generate an in-depth understanding of interaction in individual testing situations and to inform the development of hypotheses that can be explored in other sources of data.

In a previous study, Ackerman-Piek and Massing (2014) discussed data from a large number (*n* 245) of audio recordings on administration of the PIAAC background questionnaire in Germany. They coded interview behaviour from the audio recordings and noted frequent departures from the standardised background questionnaire interview script as 'deviant behaviour' (p218) and as a threat to assessment validity. This paper differs from their work in a number of ways. The transcripts discussed below focus on the computer-based direct assessment rather than the background questionnaire, and on test taking behaviour in individual assessment events – hence its 'micro-analytic' orientation.

This paper does not view variation in interviewer behaviour and departures from 'standard' testing protocols as necessarily representing a threat to validity. Indeed, it considers how interviewer improvisations might perform as a source of

assessment quality – demonstrating their professional abilities and their sensitivity to the demands of the testing situation (Maddox, forthcoming). That is, it is not immediately clear whether the 'interviewer effect' is a threat to validity as an unwanted source of variance, or an essential component of assessment quality. Indeed, as the paper will show, interviewer improvisations are supported by interviewer guidance and training manuals (OECD, 2011). A skilled interviewer performs multiple roles (including quality control), and can respond quickly and sensitively to the demands of household based assessment (OECD, 2011). We should consider their distinctly human contribution, and whether a computer is able to take on and competently perform those roles.

REPAIR IN COMPUTER-BASED TESTING

Bruno Latour (1992, 1999) considers how human roles are replaced by non-human actors (e.g. speed bumps, automatic doors) and how that places certain constraints on human agency. Human actors are located in a socio-material assemblage involving the agency of human and technological artefacts. Such an assemblage is illustrated in the transcript below, which demonstrates how interviewers and respondents interact with each other and with the computer. The transcript takes place at the start of the assessment process as the respondent completes an initial 'orientation module', and focuses on a process of 'repair' (Frolich, Drew and Monk 1994), as the respondent makes an embarrassing procedural mistake and looks to the interviewer for guidance and face-saving support (i.e. maintenance of dignity). The interviewer (test administrator) sits nearby, close enough to be drawn into the repair process.

Key:

I = interviewer (test administrator)

R = respondent (test taker)

[square brackets indicate overlapping speech.

((double brackets indicate descriptions of what is going on.

Transcript 1. *'I blew it!'*

Line 1. I: We'll arrange it this way … We'll place it here, [so you're more comfortable.

((while she is talking, the interviewer is arranging the laptop and mouse))

Line 2. R: [I see, it's okay. That' it.

Line 3. R: So we are here and go down here ((looking at the screen)), I see. Good.

Line 4. R: Let's move on. ((R looks up from the screen directly at the interviewer)).

Line 5. R: Will you do the talking? ((R gestures as if typing in the air)).

Line 6. I: No, no, I'm not allowed to do anything.

Line 7. R: I just move on? [Yes, yes, yes, yes yes

Line 8. I: [You work independently, yes. That's it. ((R is looking at

Line 9. the screen, and the Interviewer is has leaned over momentarily to see the

Line 10. screen, i.e. shared gaze.))

Line 11. R: What about, I mean, what now? I've done this already. Choose a month..

Line 12. and I chose October – now what? Why in fact it again..

Line 13. I: Um, Um, choose May. ((the interviewer leans over to see the screen))

Line 14. R: I see! I did .. [no ((when the respondent realises what she has done she

Line 15. opens her eyes wide and gestures with her hand at 45 degrees to the

Line 16. screen. The interviewer leans over again to see the screen)).

Line 17. I: [The instructions are always at the top.

Figure 1. 'I blew it'

Line 18. R: I haven't read this at all, I read only 'select a month' and I chose the

Line 19. month we're in! I blew it! ((while saying this the respondent gestures

Line 20. indicating that she did something wrong, pointing at the screen,

Line 21. momentarily covering her face, and then looks directly at the interviewer

Line 22. and smiles))

Line 23. I: The instructions are always at the top. ((the interviewer leans over to see

Line 24. the screen and points to the top of the screen)).

Figure 2. 'Nothing works with Enter'

Line 25. R: Yes, yes, I have to take a look.

Line 26. I: Nothing works with the 'enter' ((the interviewer points to the enter key)), it

Line 27. always goes [here ((she points to the section on the screen and smiles)).

Line 28. R: [Yes yes yes, I [understand

Line 29. I: [OK. ((the respondent points her hand toward the

Line 30. screen, and nods with a serious face. The interviewer also nods then

Line 32. moves away, and the respondent continues with the assessment)).

The opening transcript illustrates how the assemblage of respondent, computer, and interviewer fit together in the testing situation and how the context of the testing situation is established through talk and gesture (on the production of 'context' through talk, see Duranti and Goodwin, 1992). The talk alone does not make sense without understanding the activity that is taking place, and the role of gesture, facial expressions and gaze. The questions about the roles of the

interviewer and respondent '*will you do the talking?*' (line 7) are typical of PIAAC testing situations as is recourse to an external point of authority for explanations of those roles '*no, no, I'm not allowed to do anything*' (line 8).

As the respondent makes a mistake we see dual roles of the interviewer in the 'repair process' – the first being the provision of information (which one might readily replace with a 'help' key on the computer). That information is provided in lines 14, 18, 24, 27 and 28. The other is emotional repair (i.e. of dignity and status). We see emotional repair taking place in talk, the expression of emotion and gesture (line 16 and line 20), and the response of the interviewer (lines 24, 25 and 28). The facial expressions go from one of concern (as the respondent opens her eyes and mouth and gestures concern with her hands and says 'I blew it!' (line 20), to a resolution that involves pointing at the screen and smiling (line 28). The whole interaction took place in around one minute and then the respondent continues to complete the assessment.

The process of repair takes place involving a three-way relationship between the interviewer, computer and respondent. The respondent does not consult the 'help' option on the computer, and instead, draws the interviewer into the problem solving strategy. Had the respondent consulted the computer 'help' option they may have resolved the problem without any need for human interaction. However, the contribution of the interviewer can be viewed as intimate, 'face saving' interaction. I.e. The tone of the talk, facial expression (smiling) and gesture (pointing at the machine) all contribute toward the repair of dignity – as affective response.

Before we consider removing the interviewer from the scene, we should ask if the computer can effectively replace those distinctly human characteristics. *Can a*

computer do that? I.e. which of the features of the 'crisis' could be avoided or resolved by improved affective design, and which relate to the distinctively human role and capacities of the interviewer?

In this case, with some rapid improvisations, it appears that the interviewer has correctly followed guidance relating to the affective, emotional state of the respondent. The interviewer manual identifies the role of talk and 'body language' in the assessment process and encourages the interviewer to observe the respondent and to practice active listening so that responses can be made in a sensitive manner:

> '*Active listening is key to understanding respondent's concerns. To listen actively, you must pay close attention to the words, the tone, and the body language of the respondent as he or she raises concerns or asks questions about the survey*' (OECD, 2011, p 47).

Furthermore, the manual identifies the significance of 'body language' as a reflection of the affective state of the respondent;

> '*Look for body language that might help you understand how the respondent feels*' (OECD, 2011, p47).

The interviewer is advised by the manual to monitor the respondent's progress in the assessment, and that interaction should be minimal (p82). But even there, the interviewer is expected to be alert to concerns:

> '*Throughout the Exercise portions of the interview, you will need to be sensitive to the respondent's reactions and concerns. This is especially true for disabled respondents or respondents uneasy about their reading or writing skills. As you monitor the administration of the exercise, sit close enough to the respondent so you can watch his or her progress on the screen or the booklet but not so close that he or she might feel you are trying to see what he or she is answering. Remain alert and pay attention to what the respondent is doing. Be ready to interact with the respondent if he or she has a question or concern*'. (OECD 2011, p85)

The interviewer in the opening transcript therefore appears to be performing their role particularly well, and in a distinctly human fashion. In doing so, the interviewer and respondent demonstrate what Goffman (1983b) called 'felicity conditions', that is, a meeting of minds based on a common understanding of what is going on. The interviewer sensitivity to what is going on is also demonstrated in the next example.

FATIGUE IN HUMAN-COMPUTER INTERACTION

In this example (from a different assessment event), a computer-based assessment was administered in the evening at the respondent's home. Like the opening example, there were several moments during the assessment where talk and gesture played a significant role in assessment performance. The respondent explained at the start of the assessment that he felt tired as he had been at work all day. The example illustrates the significance of fatigue, and how it plays out in assessment performance. The question of fatigue, and associated debates about

respondent resilience, self-efficacy, grit and 'ego depletion' have been much discussed in the research literature on education and assessment (e.g. Job et al. 2010; Eklöf, 2010). With those themes in mind, it is interesting to consider how interaction might shape respondent performance.

One of the things that distinguish the PIAAC computer-based test from computer games (and game based assessments), is that it pays little attention to stimulating respondent motivation and engagement. For example, as they complete the assessment, the computer does not encourage the respondent, provide feedback about their performance, or provide real-time information about how much time remains in the assessment. That lack of 'affective design' (Hudlicka, 2003; D'Mello and Kory, 2015, Monkaresi et al,2016) puts a particular burden on the interviewer to encourage respondents who are fatigued to encourage them to complete the assessment.

In this case, at 21 minutes into the assessment, the interviewer, sensing that the respondent is tired, interjects with some comments about the content of the test. The respondent considers that as the moment to go to the kitchen to pour himself a coffee. The transcript below occurs a few minutes later when he had returned with his coffee and had continued with the assessment.

Transcript 2. Can you tell me how far we are?

The respondent makes along and marked sigh with a long puff out of breath and an exaggerated puffing out of his cheeks. The blowing out breath lasts for 9 seconds.

Figure 3. 'Can you tell me how far we are?'

((The respondent turns in a slow and deliberate manner while also lifting his coffee cup to look directly at the interviewer. The respondent's facial expression and tone is slightly accusatory, though not unpleasant)).

Line 1. R: Can you tell me how far we are? ((He smiles slightly as he asks))

Line 2. I: No. ((The interviewer says with a slight shake of her head)).

Line 3. R: Ah. ((He smiles and turns away slightly as if to take a drink from his

Line 4. coffee)).

Line 5. I: You still have some exercises. You are over half, at two thirds. I'm

Line 6. speaking from [experience.

((As she speaks, the interviewer gestures with her hands left and right, makes an apologetic smile and opens her eyes wide and her head slightly bowed and tilted to one side. The respondent smiles in response, takes a sip of coffee and puts his cup down, smiles and nods)).

Line 7. R: [yeah,.[

Line 8. I: [But I cannot influence the computer's selection of exercises for you.

Line 9. So that … there can be a slight deviation.

((As the interviewer explains the respondent turns back to face the computer and continues with the assessment.))

In transcript 2, the interviewer and respondent are aware of the significance of fatigue in the assessment. The interviewer speak 'for' and collaborates 'with' the computer, making a tacit apology for the length of the assessment and the inability of the computer to communicate information on timing. Through talk and gesture the interviewer and respondent are able to not only communicate information, but also establish a kind of rapport (smiling, humour). The initial rebuff, 'no' (line 2), is followed by an ironic smile (line 3), and prompts the further discussion about timing (lines 5-7), and about the agency of the computer (lines 8 and 9).

As Sellar (2014, p5) argues, affect is located temporally and socially, and impacts on people's capacity to act. In this case, the short interaction highlights the dynamic character of respondent stance in the assessment process. That is, affective stance (i.e. engagement, motivation, will), is fluid and is influenced by interaction. The sensitive interviewer response and recognition of the fatigue helps to renew the respondent commitment to the test and the respondent continues to complete the test, concentrating for a further 20 minutes. At the end of the assessment the respondent clearly communicates a personal 'assessment' (Goodwin and Goodwin, 1992) of the testing process:

Transcript 3. We have finished.

Line 1. I: We have finished.

Line 2. R: Oooooh, finally.

Line 3. I: Was it a bit difficult?

Line 4. R: It was killing me, this now .. [but

Line 5. R: [OK, that's it.

Line 6. I: That's it.

Line 7. R: Damn, I thought it would never end …

The transcripts above show that the human interviewer is more sensitive and responsive to the affective state of the respondent than the computer. The PIAAC computer-based assessment is not able to respond to the signs of respondent fatigue, the lengthening response times or the long gap in the log files for the coffee break. In contrast, the interviewer can "smell the coffee", and departing from the standard testing protocol in the computer-based assessment the interviewer makes an ethical judgement (Smith, 2003) to communicate empathy in a way that encourages the fatigued respondent to continue.

PROCESS DATA AND VALIDITY PRACTICE

The final part of this paper considers how micro-analytic process data on talk and gesture should be integrated into validity practice.

Hubley and Zumbo (2017) note that process evidence is 'sorely neglected' in validity studies (p7), and suggest broadening the models and sources of response process and data available (also Li, Banerjee and Zumbo, 2017). However, this highlights a question about the different roles that process data can play within validity practice, and the justification for its inclusion.

In his discussion of different sources of validity evidence, Newton (2016) discusses the contributions of 'Micro- and Macro validation':

'*Macro-validation is akin to the customer's perspective on assessment: does the assessment procedure work in the way that it ought to work? Micro-validation is akin to the engineer's perspective on assessment: is the assessment procedure built in the way that it ought to be built? The critical point is that these two perspectives represent different kinds of inquiry. Macro-validation research tends to investigate outcome-related, or product-related questions; whereas micro-validation research tends to investigate input-related, or process-related questions*' (Newton 2016, p5).

'Micro-analytic' observations, of talk and gesture in tests can certainly inform validation practice as an aspect of 'process scrutiny' (Newton, p4), as it significantly expands the evidential base that can inform validation practice.

However, while Newton's (2016) 'Micro-Macro Validation Continuum' provides a general conceptual framework, this paper makes a more specific methodological claim regarding the contribution of micro-analytic data.

The micro-analytic approach described in this paper may be described as inductive as it seeks to observe social phenomenon, and to identify patterns that inform explanation, and support and hypotheses development. Those observations can then support pragmatic application, and the development of new theory, models and arguments.

The approach therefore supports the arguments made about validity as 'contextualised and pragmatic explanation' made by Zumbo (2009), and Stone and Zumbo (2016), as observations of testing situations provide information that support

practice oriented judgments on the valid use of assessment data (whether they are test scores or fine-grained process data).

The inductive micro-analytic approach contrasts with the more deductive orientation to 'process model interpretations' described by Kane and Mislevy (2017), and Kane's (2013, 2016) argument approach to validity. There is no reason why any source of behavioural process data (e.g. talk, eye tracking, gesture) should not be used to validate trait interpretations in real-world settings (see Kane and Mislevy, 2017), or be used for other forms of pragmatic explanation. However, the approach taken in this paper has not been to conform or refute existing arguments. Instead, it uses observational data on talk and gesture to identify new and unexpected inferences about assessment performance. Those observations lie outside the precisely specified evidence that we would associate with an argument-based approach (Kane, 2016).

Observing Disengagement

The concluding transcripts (below) illustrate the argument that micro-analytic observations identify features of assessment performance that are not captured by routine validation practice. In this final section we illustrate this point by looking at the theme of respondent engagement.

In routine validation practice, engagement in PIAAC is captured by test data on item-completion and response time data from log files (Goldhammer et al. 2016). PIAAC aims to assess the normal performance of respondents in every-day tasks, rather than rehearsed test performance. As a result, there is some variation in engagement as a feature of respondent resilience (Goldhammer, 2015; Goldhammer, Naumann and Greiff, 2015). However, the presence of large-scale,

construct irrelevant disengagement can be viewed as 'fundamental threat to validity' in low stakes assessment (Goldhammer et al. 2015, p6, also Wise, 2006), as an indicator of poor data quality.

In the transcript below, we join the assessment at 27 minutes, about halfway through a PIAAC assessment when the respondent looks up from the screen and indicates with a facial expression that she is struggling (i.e. a non-verbal response cry). The interviewer responds by moving from where she is sitting and leans over so that she can see the computer screen.

Transcript 4. Is there still much to go?

Line 1. I: *I can't see where ..* ((the interviewer looks at the computer screen))

Line 2: R: *yes, how far I am*

Line 3. I: *Yes*

Line 4. R: *Pardon?*

Line 5. I: *I don't know.*

Line 6. R: *A question .. the second exercise .. Is there still much to go?*

Line 7. I: *I never know the precise number of questions.*

Line 8. R: *I see* [

Line 9. I: [*so that* [

Line 10. R: [*yes, yes, okay*

Line 11. I: *Yes*.

Line 12. R: *Are you tired?*

Line 13. I: *Yes,. well, so so.*

In transcript 4 we observe the relationship between time and respondent fatigue (i.e. 'is there still much to go'). In this example, like the other transcripts presented in this paper, a 'solution' is found with interviewer-respondent interaction, and displays of empathy, i.e. physical proximity, joint gaze at the computer screen, enquiry into the wellbeing of the respondent (see also Maddox, Bayliss, Fleming, Engelhardt and Borgonovi, forthcoming). The respondent continues with the assessment, but a few minutes later (at 33 minutes and 46 seconds) the respondent looks directly at the interviewer, shakes her head and smiles. The non-verbal gesture initiates a second interaction:

Transcript 5. ..This is not for me

Line 1. I: *It won't go?*

Line 2. R: *No, this is not for me*

Line 3. I: *I see, if there are problems you can freely move on,*

Line 4. *there's no* [

Line 5. R: [*It's not so difficult, but I'm a bit..*

((she smiles, and gestures with her hand))

Line 6. I: *I see, it's long..*

The respondent is struggling – as indicated through non-verbal gestures, and through her verbal statement (i.e. '*this is not for me*'). The video-ethnographic observation captures an episode of respondent fatigue in talk and gesture. That is corroborated by the post-assessment interview:

Transcript 6. **Did you ever give up and skip?**

Line 1. Author: *When there was too much text, did you ever give up and skip?*

Line 2. *Just press advance?*

Line 3. R: *um, ..I think here, in this one..*

((the respondent is referring to a discussion of item C304B 'Contact Employer'))

Line 5. Author. *In that one yeh?*

Line 6. R: *In that one.*

Line 7. Author. *Any others?*

Line 8. R: *hm.. I don't know. Maybe I just… yeh, here, um..*

Line 9. Author: *Did you skip any others?*

Line 10. R. *I don't know. Maybe one. [Maybe one.*

Line 11. Author. *[Not so many?*

Line 12. R. *No, no.*

A single case is not sufficient to make significant validity judgements about the performance of a test. However, it does offer insights into assessment response processes and the character of respondent disengagement. By synchronising the video-ethnographic data and the assessment log file we can precisely identify the

sequences of test items that the respondent completed when she became disengaged. This enables comparison of different sources of validity data (e.g. test scores and log files).

The log files from the testing situation do not capture any of the talk and gesture in the testing situation, and provide little explanatory value into what was going on. However, they do capture data on item response times and key strokes. They also confirm the identity of the items that were subject to disengagement, and that may help to identify patterns of disengagement linked to test item design, test position or content. In this example, the log files captured relatively long response times that took place during the discussion between the interviewer and respondent, and some rapid response. However, the response times across the duration of the test are not sufficient for the respondent to be characterised as 'disengaged'.

Test scores also offer little insight into this example. Patterns of item non-response may be viewed as symptomatic of respondent disengagement. However, in this case non-response data is confounded by the respondent knowingly providing incorrect answers as disengagement behaviour. This is revealed in the post-assessment interview:

Transcript 7. In one or two cases..

Line 1. Author: You had to concentrate, maybe it wasn't so exciting, you still

Line 2 *concentrated and [*

Line 3. R: [*Yeah, I* tried*, but maybe in one or two cases I just … what's the word*

Line 4. *for označiti?*

Line 5. Author: *Highlighted?*

Line 6. R: *Yes, highlighted and I didn't care if it's true …*

Line 7. Author*:* Oh, I see, so you just finished?

Line 8. R: Yeah, yeah [

Line 9. Author*: [just highlighted a bit and then pressed advance.*

Line 10. R: *Yeah*.

For this respondent we can see clear evidence of an episode of disengagement from the video-ethnographic data, the evidence of talk and gesture, and from the post assessment interview. However, the disengagement behaviour is not captured by routine validation practice.

Sellar (2007) argues that despite the apparently overwhelming role of technology in assessment, we should be careful to recognise the place of human agency:

> '*It has become popular in our current moment of 'big data' hype to see 'the meat' (Land, 1995) – organic human bodies – as rapidly losing the game of performativity to algorithms that have revolutionised the speed and scale of data analytics. Clearly, processes of commensuration that operate on digital computing platforms are playing influential roles in both constituting and modulating new cultural formations, but they are not detached from analogue cognition and affective sense-making that form part of the growing data infrastructure in education..*' (Sellar, 2014, p8).

The transcripts on disengagement re-enforce Sellar's argument by capturing the significance of cognitive-affective decisions and response processes that take place under the radar of log files and routine quality assurance practices. While it would not be either practical or desirable to capture this level of detail on each respondent, these examples demonstrate the scope for micro-analytic observations to improve validity practices and to contribute to next generation test design.

CONCLUSIONS

The paper has described a micro-analytic approach to assessment, and demonstrated how information on talk and gesture can be integrated into judgements about test performance and validity. The focus on talk in tests signals, and is part of a wider transformation as multiple sources of process data are integrated into the analysis of assessment performance and validity (Ercikan and Pellegrino (Eds) 2017; Zumbo and Hubley (Eds.) 2017). That methodological pluralism involves an expansion of 'what counts as data' (Shear and Zumbo, 2014) in assessment, and transcends conventional distinctions between micro and macro, qualitative and quantitative (Newton, 20216).

If, as Hubley and Zumbo (2016, p300) argue, it is better to include more, rather than fewer sources of evidence on test validity, then the inclusion of micro-analytic data on talk and gesture significantly expands the types of data and inferences that are possible. The inclusion of interactive data on the testing situation proposed is not so much an 'intrusion of context' (Messick, 1989, p14-15), but a

recognition of ecological sources of variation in assessment performance, some of which may be considered as threats to validity, and others as evidence of quality.

The paper has shown that some performance characteristics of assessment are not effectively captured by log files or test scores. The micro-analytic approach described in the paper offers a way around that impasse. Observations of talk and gesture capture intimate characteristics of interaction, stance and affect in testing situations. Those areas of 'interaction' have been neglected in routine validity practice, despite strong arguments about their importance (e.g. McNamara and Roever, 2006, McNamara, 1997). However, with advances in CBT and digital technology, process data on talk, respondent gaze, facial expression and gesture promise to transform the analysis of test performance and validity (Orange, Gorin, Jia and Kerr, 2017; D'Mello and Kory, 2015).

Observations of PIAAC showed how interviewer-respondent interaction contributes to the quality of the assessment process, and the sensitivity of the interviewer to respondent affect have a performative role on respondent engagement. This suggests that it is necessary to re-think the idea of 'interviewer effects' as being synonymous with measurement error (Maddox, forthcoming).

The series of transcripts helped us to dig deep into the characteristics of human-computer-human interaction in the PIAAC assessment. The first example on 'repair' showed that the interviewer used talk and gesture to provide timely and sensitive information and face-saving support following a mistake. The second example, on 'fatigue', showed that the interviewer can speak *for* and collaborate *with* the computer, providing information on timing, and encouraging the respondent to continue with the assessment. The final series of transcripts shows what can be lost

if studies of assessment validity rely too much on log file data on key strokes and response times. By observing talk and gesture in testing situations, we can expand the information base of process data, and observe intimate features of interaction and affect that are not yet captured by computers.

REFERENCES

Ackerman-Piek, D. and Massing, N. (2014). Interviewer Behaviour and Interviewer Characteristics in PIAAC Germany. *Methods, Data, Analyses*. 8 (2), 199-222.

Cope, B., & Kalantzis, M. (2016). Big Data Comes to School: Implications for learning, Assessment and Research. *AERA Open*. 2 (2). 1-19.

DiCerbo, K. E. & Behrens, J. T. (2014). *Impacts of the Digital Ocean on Education*. London: Pearson.

D'Mello, S. and Kory, J. (2015). A review and meta-analysis of multi-modal affect detection systems. *ACM Computing Surveys*. 47 (3), 1-36.

Du Bois, J.W. and Karkkainen, E. (2012). Taking a Stance on Emotion: Affect, Sequence and Intersubjectivity in Dialogic Interaction. *Text and Talk*. 32 (4) 433-451.

Duranti, A. (1997). *Linguistic Anthropology*. Cambridge, Cambridge University Press.

Goodwin, C., & Duranti, A. (1992). Rethinking context: An introduction. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon* (pp. 1–42). Cambridge. Cambridge University Press.

Eklöf, H. (2010) Skill and will: test-taking motivation and assessment quality, Assessment in Education: Principles, Policy & Practice, 17:4, 345-356

Ercikan, K. & Pellegrino, J.W. (Eds). (2017). Validation of Score Meaning for the Next Generation of Assesments. Routledge.

Ericsson, K., & Simon, H. (1993). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Frohlich D, Drew P, Monk A. (1994) 'Management of Repair in Human-Computer Interaction'. *Human-Computer Interaction*, 9 (3-4), pp. 385-425.

Goffman, E. (1963). The neglected situation. *American Anthropologist*, 66, 133–36.

Goffman, E. (1981). *Forms of talk*. Philadelphia. University of Pennsylvania Press.

Goffman, E. (1983a). The interaction order. American Sociological Review, 48, 1–17.

Goffman, E. (1983b). Felicity's condition. *American Journal of Sociology*, 89, 1–53.

Goldhammer, F. (2015). Measuring Ability, Speed, or Both? – Challenges, Psychometric Solutions and What Can be Gained From Experimental Control. *Measurement: Interdisciplinary Research and Perspectives, 13*, 133-164.

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence, 3*, 21-40.

Goldhammer, F., Martens, T.; Christoph, G.and Lüdtke O. (2016), Test-taking engagement in PIAAC. OECD working papers. OECD. Paris.

Goodwin, C. (2007). Environmentally couples gestures. In S.D. Duncan, J. Cassell, and E.T. Levy (Eds.) *Gesture and the Dynamic Dimension of Language*. (pp195-212). Amsterdam. John Benjamins.

Goodwin, M.H., Cekaite, A. and Goodwin, C. (2012). Emotion as Stance. In M. L. Sorjonen, and A. Perakyla (Eds.), *Emotion in Interaction*. Oxford. Oxford University Press.

Goodwin, C., & Goodwin, M. H. (1992). Assessments and the construction of context. In A. Duranti & C. Goodwin (Eds.), Rethinking context: Language as an interactive phenomenon. (pp. 147–189). Cambridge: Cambridge University Press.

Greiff, A. Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioural data from computer-generated log files. *Computers in Human Behaviour*. 61. 36-46.

Hubley, A. M. and Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*. 103. 219-230.

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human-Computer Studies*. (59). 1-32.

Jeong, H. (2014). A comparative study of scores on computer-based and paper-based tests. *Behaviour and Information Technology*. 33 (4). 410-422.

John Jerrim (2016). PISA 2012: how do results for the paper and computer tests compare?, *Assessment in Education: Principles, Policy & Practice* (published on-line, volume and page numbers to follow).

Job, V., Dweck, C.S., and Walton, G.M. (2010). Ego depletion, is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological Science*, 21 (11), 1686-1693.

Kane, M.T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23:2, 198-211.

Kane, M.T. (2013).'Validating the interpretations and uses of test scores. *Journal of Educational Measurement*. 50. 1-73.

Kane, M.T. & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan, J. W. Pellegrino (Eds). (2017). Validation of Score Meaning for the Next Generation of Assesments. Routledge. pp 11-24.

Kendon, A. (2007). On the origins of Modern Gesture Studies. In S.D. Duncan, J. Cassell, and E.T. Levy (Eds.) *Gesture and the Dynamic Dimension of Language*. pp13-28.

Latour, B. (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*. pp. 225–258. Cambridge, Mass. MIT Press.

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

Lee, H. and Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education.* 2 (8). On-Line.

Lee, H. and Haberman, S.J. (2016). Investigating test-taking behaviours using timing and process data. *International Journal of Testing*. 16 (3). 240-267.

Li, Z., Banerjee, J., & Zumbo, B.D. (2017). 'Response time data as validity evidence: Has it lived up to its promise and, if not, what would it take to do so'. In B.D. Zumbo, and A.M. Hubley (Eds). *Understanding and Investigating Response Processes in Validation Research*. Springer. pp 159-178

Maddox, B. 'Interviewer-Respondent Interaction and Rapport in PIAAC: the OECD Survey of Adult Skills'. *Quality Assurance in Education*. (under review)

Maddox, B. (2015). 'The Neglected Situation: Assessment Performance and Interaction in Context'. *Assessment in Education: Principles, Policy and Practice*. 22 (4) 427-443.

Maddox, B. (2014) 'Globalising Assessment: An Ethnography of Literacy Assessment, Camels and Fast Food in the Mongolian Gobi'. *Comparative Education*. 50. 474-489.

Maddox, B., Zumbo, B.D., Tay-Lim, B. S-H., & Qu, D. (2015). 'An Anthropologist among the Psychometricians: Assessment Events, Ethnography and DIF in the Mongolian Gobi'. *International Journal of Testing*. 14 (2)291-309.

Maddox, B., Bayliss, A., Fleming, P. and Bourgonovi, F. The use of Eye Tracking Data in Large-Scale Assessment. *European Journal of Psychology in Education*. (under review)

McNamara, T. (1997). “Interaction” in second language performance assessment: Whose performance? *Applied Linguistics, 18,* 446–466.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Blackwell, Malden MA and Oxford.

McNeill, D. (1985). So you think gestures are non-verbal? *Psychological Review*. 92, 350-371.

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

Messick, S. (1998). Test Validity: A Matter of Consequences. *Social Indicators Research*. 45. 35-44.

Monkaresi, H. Bosch., N, Calvo., & R., D'Mello. S.K. (2016). Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*.

OECD. (2011). Interview Procedures Manual. PIAAC Main Study. 28th February, 2011. Paris, OECD.

OECD. (2016a). Skills Matter. Further Results from the Survey of Adult Skills. Paris, OECD.

OECD. (2016b). Technical Report of the Survey of Adult Skills (PIAAC), 2nd Edition. Paris, OECD.

Orange, A., Gorin, J., Jia, Y., & Kerr, D. (2017). 'Collecting, analysing, and interpreting response time, eye tracking and log data' in In K. Ercikan, J. W. Pellegrino (Eds). (2017). *Validation of Score Meaning for the Next Generation of Assesments*. Routledge. pp 39-51.

Pickard, R. (2003). Affective Computing: Challenges. *International Journal of Human-Computer Studies*. (59). 55-64.

Pepper, D., Hodgen, J., Lamesoo, K., Kõiv, P. & Tolboom, J. (2016): Think aloud: using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics, *International Journal of Research & Method in Education.*

Russell, M., Goldberg, A., and O'Connor, K. (2003). 'Computer-based testing and validity: A look back at the future. *Assessment in Education: Principles, Policy and Practice*. 10 (3) 279-293.

Shear, B., & Zumbo, B.D. (2014). 'What counts as evidence: A review of validity studies in Educational and Psychological Measurement. In Zumbo, B.D. & Chan, E.

K. H. (Eds.), *Validity and validation in social, behavioural, and health sciences*. Springer. pp91-112

Shohamy, E. (1993). The exercise of power and control in the rhetorics of testing. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 23–38). Tampere, Finland: Universities of Tampere and Jyväskylä.

Schegloff, E., & Sacks, H. (1973). Opening up closings. *Semiotica*, 8, 289–327.

Schegloff, E. A. (1988). Goffman and the analysis of conversation. In P. Drew & A. Wootton (Eds.), *Erving Goffman: Exploring the interaction order* (pp. 89–135). Cambridge: Polity Press.

Sellar, S. (2014). A feel for numbers: Affect, Data and Education Policy. *Critical Studies in Education*. 56 (1), 131-146.

Schegloff, E. A. (1988). Goffman and the analysis of conversation. In P. Drew & A. Wootton (Eds.), Erving Goffman: Exploring the interaction order (pp. 89–135). Cambridge: Polity Press.

Shear, B. and Zumbo, B.D. (2014). What counts as evidence : A review of validity studies in educational and psychological measurement. B.D. Zumbo and E.K.H Chan (Eds.). *Validity and Validation in Social, Behavioural, and Health Sciences*. Springer. 91-112.

Stone, J. & Zumbo, B.D. (2016). 'Validity as a pragmatist project: A global concern with local application. In V. Aryadoust and J. Fox (Eds), *Trends in language assessment research and practice*. pp 555-573. Newcastle, Cambridge Scholars publishing.

Thompson, G. (2016). Computer adaptive testing, big data and algorithmic approaches to education. *British Journal of Sociology of Education*.

Williamson, B. (2016). Digital Education Governance: Data Visualisation, Predictive Analytics and 'Real-Time' Policy Instruments. *Journal of Educational Policy*. 31 (2), 123-144.

Wise, S. (2006). 'An investigation of the differential effort received by items on a low-stakes computer based test'. *Applied Measurement in Education*, Vol. 19, pp. 95–114.

Zumbo, B.D. (2007). 'Three generations of differential item functioning (DIF) analyses: considering where it has been, where it is now, and where it is going'. *Language Assessment Quarterly*. 4, 223-233.

Zumbo, B.D. (2009). Validity as contextualised and pragmatic explanation, and its implications for validation practice', in R. Lissitz (Ed). *The concept of validity: revisions, new directions and applications*. pp 65-82. Charlotte, N.C. Information Age publishing.

Zumbo, B.D. and Chan, E. (2014). *Validity and Validation in Social, Behavioural and Health Sciences*. Springer.

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*, 23:2, 299-303.

Zumbo, B. D. and Hubley, A. M. (Eds.). (2017). Understanding and Investigating Response Processes in Validation Research. Springer.