Eur J Psychol Educ
https://doi.org/10.1007/s10212-018-0380-2

Observing response processes with eye tracking in international large-scale assessments: evidence from the OECD PIAAC assessment

Bryan Maddox[1] · **Andrew P. Bayliss**[1] · **Piers Fleming**[1] · **Paul E. Engelhardt**[1] · **S. Gareth Edwards**[1] · **Francesca Borgonovi**[2]

Received: 4 October 2017 / Revised: 27 March 2018 / Accepted: 4 April 2018

Abstract This paper reports on a pilot study that used eye tracking techniques to make detailed observations of item response processes in the OECD Programme for the International Assessment of Adult Competencies (PIAAC). The lab-based study also recorded physiological responses using measures of pupil diameter and electrodermal activity. The study tested 14 adult respondents as they individually completed the PIAAC computer-based assessment. The eye tracking observations help to fill an 'explanatory gap' by providing data on variation in item response processes that are not captured by other sources of process data such as think aloud protocols or computer-generated log files. The data on fixations and saccades provided detailed information on test item response strategies, enabling profiling of respondent engagement and response processes associated with successful performance. Much of that activity does not include the use of the keyboard and mouse, and involves 'off-screen' use of pen and paper (and calculator) that are not captured by assessment log-files. In conclusion, this paper points toward an important application of eye tracking in large-scale assessments. This includes insights into response processes in new domains such as adaptive problem-solving that aim to identify individuals' ability to select and combine resources from the digital and physical environment.

Keywords PIAAC · Response processes · Eye tracking · Large-scale assessments

✉ Bryan Maddox
b.maddox@uea.ac.uk

[1] University of East Anglia, Norwich, UK

[2] The organization for Economic Co-operation Development (OECD), Paris, France

Published online: 29 May 2018

This paper describes an exploratory study that used eye tracking to examine assessment response processes in the OECD's Programme for the International Assessment of Adult Competencies (PIAAC). We build on recent developments on the use of small-scale process data in large-scale assessment to support validation practice and test development (see Zumbo and Hubley, 2017; Ercikan and Pellegrino, 2017). For the purposes of this paper, we adopt Hubley and Zumbo's (2017) definition of response processes as '.. the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed score variation' (Hubley and Zumbo, p. 2).

There are multiple ways to investigate response processes in large-scale educational assessments. Those include the analysis of computer-generated log files on response times and key strokes (e.g. Goldhamer et al., 2015), video-ethnographic observation (e.g. Maddox 2017) and cognitive interviews (e.g. Padilla and Benitez (2017), Pepper et al. (2016) and Radišić and Baucal, (2018)). As discussed later in this paper, eye tracking further expands the information based on response processes, enabling fine-grained observation of respondent engagement at the item level (see Krstić, Šoškić, Ković and Holmqvist, 2018).

Eye tracking can provide important clues to visual attention (Ferreira and and Henderson 2004) as well as insightful data regarding assessment response processes. These techniques are extensively used in psychological research in areas such as reading, joint gaze and scene perception (e.g. Clifton et al. 2016; Liversedge et al., 2015; Lai et al., 2013). Their application is particularly influential in fields such as reading research, in which the use of eye tracking is a well-established methodological technique.

Eye tracking technology enables the researcher to observe and record detail of respondent's eye movements and the precise area of their attentional focus. They therefore capture detail that is not obtained from assessment log files, think aloud protocols or administrator observations of assessment. As a result, we wanted to examine what additional insights eye tracking would provide about assessment response processes.

Eye movements in reading and scene perception are extremely rapid and typically involve three to four fixations per second. Eye tracking enables observations of *saccades*, (rapid eye movements typically lasting approximately 100 ms) and *fixations* (temporary pauses of typically 200–300 ms). The underlying assumption for all eye tracking studies is that longer fixation durations and longer gaze duration (i.e. total viewing time) indicates more cognitive processing (Rayner 1998). Conversely, shorter more rapid fixations reflect shallower, less intensive processing. Analyses of fixation durations tend to focus on particular areas of interest (AOI), and so for standardised assessments, gaze durations can be examined for things such as question stems, visual illustrations or the response options in multiple choice questions (MCQs). A further type of dependent measure, which is provided by eye tracking data, is the possibility of examining *scan paths* that is examining the sequence of fixations to determine what respondents fixate on and when, and also the order in which they do so. As this paper highlights, eye tracker data on scan paths is especially useful for assessment research because it provides detailed information on response behaviours and sources of variation. That kind of data is not available from conventional think aloud protocols or from log files that are generated in computer-based testing. Finally, eye trackers also provide information on the size of the pupil, which is the best physiological measure of cognitive load/effort (Kahenman 1973; Beatty and Lucero-Wagoner 2000). Thus, pupil diameter provides a further type of data with regard to test engagement and the deployment of cognitive effort by test respondents.

Eye tracking and assessment

One of the legacies of the extensive research literature on reading and scene perception is that it has often sought to capture behavioural norms rather than sources of variation that one might associate with developmental processes and learning (see Liversedge, Schroeder, Hyönä and Rayner 2015). One significant exception is eye tracking research on dyslexic respondents (Hawelka et al., 2010; Benfatto et al. 2016). In contrast, the major concern in eye tracking research on educational assessment (and response process data more generally) is *variation* in response processes that can be attributed either to differences in the particular trait under investigation, or differential item functioning (DIF) that might be attributed to test design (Zumbo, 2007), or some additional source of variation that is present in the testing situation (Hubley and Zumbo 2017). An additional and overlapping contemporary concern is the use of eye tracking to understand how respondents engage with the complex and multi-modal assessment tasks that are used in 'next generation' computer-based testing (Oranje et al., 2017).

In the field of large-scale assessment, eye tracking has become increasingly used in processes of test item design and validation (e.g. Paulson and Henry 2002; Tai, Loeher and Brigham 2006; Oranje et al., 2017; and Krstić et al., 2018). However, the number of eye tracking studies remains quite limited, and most of those that do exist in this nascent field are exploratory in nature. For example, Solheim and Uppstad (2011) showed that student's achieving similar reading comprehension scores often had very different behavioural (eye movement) patterns, suggesting a much more nuanced picture of constructing meaning from a text. In particular, they differentiated what they referred to as 'task-oriented' readers from more 'effortful' readers. Task-oriented readers had a tendency to quickly scan the text stem and then proceed to the question and response options. After processing the question and response options, task-oriented readers returned to the text stem in order to search for the critical information. Effortful readers in contrast, adopted a strategy in which they careful processed the text stem first and then proceed to the question and response options.

In the studies by Tai et al. (2006) and Hu et al. (2017), participants were classified into a high- or low-performing group based on their test item accuracy, and then (post-classification), the researchers examined the eye movement patterns between the two groups (also, Krstić et al., 2018). Tai and colleagues concluded that the differences in eye movement patterns validate the notion that (1) different parts of a test item and (2) different sub-tasks are required to successfully answer an item. Furthermore, the differences in eye movement patterns are indicative of different types of information processing and problem-solving strategies. Their conclusions focus on how a deeper understanding of problem-solving behaviour is enhanced by the triangulation of evidence obtained from eye movements (i.e. attention allocation and processing effort). As noted above, pupillometry provides another key piece of evidence concerning processing effort, which would strengthen the idea of evidence 'triangulation'.

This study aims to examine the cognitive processes of respondents as they complete the PIAAC assessment. Eye tracking was used to give detailed information on the focus of attention over the duration of the assessment in combination with physiological measures as a proxy for engagement. While previous research has looked in detail at eye movements onscreen, we were also interested in off-screen behaviour which is not captured at all by computer log files, but might involve important interactions with proximal tools or people. Our research question was to examine the extent to which variability in cognitive processes and off-screen interaction is present and meaningful—even when the final test item response is the same.

The structure of the rest of the paper is as follows. We begin by describing our research methodology and design and then present our initial results. We highlight some of the most notable implications in terms of the insights that eye tracking can provide for large-scale assessment. In the "Results" section, we show how data on fixations and saccades provides detailed accounts of test item response strategies, enabling profiling of respondent engagement and response processes associated with successful performance. We conclude by discussing how eye tracking can contribute to large-scale assessments in the future.

Method

Eye tracking research in assessment typically takes place in the laboratory and involves a desk-mounted eye tracker connected to a computer. However, this is a rapidly developing field. As eye tracking and gaming technology advances, eye tracking is likely to be incorporated into computers, tablets and smart phones as standard equipment. This will increase the potential for eye tracking to get out of the lab, to inform analysis of respondent performance (see D'Mello et al., 2017). Desk-mounted eye trackers (such as the EyeLink 1000) offer high levels of precision and speed (i.e. sampling frequencies of 1000 Hz and 0.25–0.5 degrees of visual angle). They enable the analysis of eye movements for example, within single words and sentence comprehensions—albeit mediated by the algorithms and researcher decisions about saccade/fixation time thresholds for the capture of saccades and fixations, as well as blinks (Kennedy 2016). In this research, our primary interest was to observe scan paths to explore how respondents engaged with the test items rather than investigating the correlation between respondent ability and the duration and frequency of their saccades and fixations (for a contrasting methodological approach, see Krstić et al., 2018).

We chose to use less invasive eye tracking 'glasses' (SMI EG2). The glasses offer lower resolution, with a sampling frequency of 60 Hz. As a result, we were only able to accurately observe eye movements at the level of word and sentence. We could not observe finer detail of reading behaviour, such as word-landing positions and word revisits—aspects of reading that are used to distinguish proficient readers from those who have developmental disorders such as dyslexia (see Hawelka et al., 2010; Benfatto et al., 2016; Krieber et al. 2016). However, the eye tracking glasses have some important advantages. They do not use a chinrest and have the advantage of more closely mirroring 'normal' testing conditions. Moreover, they also allow researchers to capture 'off-screen' activity that typically takes place in assessments, such as the respondent's use of a pen and paper (or calculator), and any verbal interaction with an interviewer. In this research, those observations of off-screen activity were particularly informative.

PIAAC

The Programme for the International Assessment of Adult Competencies (PIAAC) is a large-scale assessment of the adult population, aimed at collecting comparable information on the information processing abilities of around 200,000 adults worldwide. PIAAC was implemented in 32 countries/national sub-regions. Data collection took place between 2011 and 2012 for 24 countries and a further 8 countries in 2015. The target population for the survey was the adult population, aged 16–65 years, residing in the country at the time of data collection, irrespective of nationality, citizenship or language status. The survey was administered in the

official language or languages of each participating country, and some countries gave respondents the possibility of participating in one of the widely spoken minority/regional languages.[1]

PIAAC has two main components: a background questionnaire and an assessment of literacy, numeracy and problem-solving in a technology-rich environment. The questionnaire was administered first in a CAPI format (i.e. computer-assisted personal interviewing) and response time ranged from 30 to 45 min. Upon completion of the questionnaire, respondents were expected to sit the cognitive assessment which took around 1 h to complete. Depending on their computer skills, the assessment was delivered either on a laptop computer or as a fill-in-paper booklet. Survey institutes involved in data collection in each participating country ensured that each respondent received sufficient information about the study and gave informed consent prior to participation. In the vast majority of cases, the PIAAC survey was administered in the respondents' own homes except in circumstances when respondents did not feel comfortable with completing the survey in their home. In these instances, the survey took place in a testing centre or other agreed location.

Country-specific sample sizes varied depending on the number of cognitive domains assessed and the number of languages in which the assessment was administered. Some countries boosted sample sizes in order to have reliable estimates of proficiency for the residents of particular geographical regions and/or for certain subgroups of the population, such as indigenous inhabitants or immigrants. The achieved national samples ranged from a minimum of 3892 persons in the Russian Federation to a maximum of 26,683 persons in Canada. The survey's technical standards and guidelines set a goal of a 70% unit response rate. Seven countries achieved this goal, while, for the most part, response rates were in the range of 50–60%.

Participants In this exploratory study, we wanted to examine how participants interacted with computer-based test items in the PIAAC assessment. Our lab-based study involved 14 voluntary adult participants (3 male, 11 female) who gave informed consent agreeing to participate. Participants were recruited from a panel of university staff and students in Norwich, UK. The panel includes adults, of between 20 and 45 years of age who were administrative and academic staff, postgraduate and undergraduate students who voluntarily respond to studies which they are interested to take part in for financial compensation. The study was advertised as 'computer-based learning research' involving 'questions regarding problem-solving, literacy and/or numeracy'. All participants had normal or corrected-to-normal vision. They were compensated £14 for their time.

Materials Since our aim was to provide insights into the behaviour or participants in large-scale assessments, we used the standard format of the PIAAC assessment (i.e. the identical version to the one used in the UK assessment in round 1 of PIAAC) delivered on a laptop computer. While participants in PIAAC can chose between paper-based and computer-based modes of assessment, all participants in our study used the computer-based version. The computer-based assessment delivered modules in literacy, numeracy and problem-solving in technology-rich environments (PSTRE). We were keen to reproduce conditions as similar as possible to the field study—though in this case, it involved assessment in a university lab under observation, rather than taking place in the respondents' household which would be the normal location for the PIAAC assessment. In other respects, the interviewer used standard

[1] See OECD 2013a and www.oecd.org/site/piaac for technical details.

testing protocols, including the completion of the background questionnaire prior to sitting the assessment.

Apparatus and procedure

Participants arrived at the laboratory and completed standard research paperwork including giving informed consent. They were then asked to complete some basic psychometric measures which are not covered in this paper. They then put on the eye tracking glasses. These were worn while completing the PIAAC background questionnaire but without recording while participants became accustomed to them. Physiological measurement equipment was then attached to the middle and index fingers of the participants' non-dominant hand. The equipment was then synchronised, recording started and participants began the PIAAC assessment modules.

The additional physiological measurement mentioned above was the collection of electrodermal activity (EDA) data through the use of 'Biopack MP150' an EDA100C EDA response amplifier and 'AcqKnoweldge' software as a measure of engagement. The EDA data was a secondary assessment to check convergent validity with the eye tracker observations and the pupil diameter data generated by the eye tracking glasses. We collected eye tracking data on each of the participants as they completed the entire computer-based assessment.

Results

We found that the eye tracking glasses captured the on-screen activity of participants as they completed the PIAAC assessment, and the content of off-screen behaviour, such as the use of pen and paper and the calculator. The eye tracker was also able to record and observe any verbal interactions that took place between the interviewer and the respondent. Occasionally, the respondents offered unprompted comments during the assessment including 'think aloud' type content as they completed numeracy items, and evaluative comments about their own performance.

In the remainder of the "Results" section, we highlight three key observations and then provide illustrative examples to support our key observations. Some of our results confirm the findings of existing work (e.g. Hu et al. 2017; Paul and Henry 2002; Tai et al. 2006), and some of our results (e.g. pupil data) are novel.

Key observation 1: scan paths and points of reference

The eye tracker allowed observation of 'scan paths' as respondents completed test items, which permitted descriptive analysis of the data. To do so, we made comparisons of how different respondents completed the same items. There was a precise and reliable relationship between paths, and the pupil data and electrodermal activity (EDA). The pupil data more closely mapped onto observations of response processes, whereas the EDA data showed slight delays that are expected with a skin-based conductance measure (Ren, Barreto, Gao and Adjouadi 2012). We observed which regions of the test item the respondent read, and in what order. Our observations had a confirmatory role in terms of validity, because they showed that respondents had to read the instructions and the test item before they were able to provide

answers. That relationship has not always been evident in assessment, as previous eye tracker studies have observed (Oranje et al. 2017). We also observed that respondents frequently re-fixated to the item instructions as they worked out their answers (see example below on the 'Bottles' item). This suggests that the item instructions operated like a point of reference, which is consistent with several lines of research, for example in picture comparison tasks (Gajewski and Henderson 2005) where participants will frequently return to the 'target' image while looking for the 'matching' image.[2]

Key observation 2: levels of engagement

Overall, in our study, the respondents exhibited high levels of engagement. This is not surprising as they were participating in a psychological study and knew that their behaviour was being recorded.

From observing the eye tracking video, we were not able to identify behaviours that indicated disengagement from the assessment tasks. However, we observed one case where a frustrated respondent did not fully read the text, and guessed the incorrect answer. Our analysis of EDA and pupil data as cognitive effort markers across the whole sample of respondents did not indicate any significant drop in cognitive load or arousal across the duration of the assessment. We compared these dependent measures across time-blocks just after the beginning and just before the end of each module. This suggests that for this cohort at least, 'disengagement' behaviour should not simply be attributed to fatigue associated with test duration. It would be very easy to identify disengagement behaviour on the eye tracking video, or behaviour that indicated confusion that might be associated with item difficulty or item misfit.

Key observation 3: response processes associated with successful performance

The eye tracking observations contributed to our understanding of the relationship between item difficulty and respondent ability (see Goldhamer et al., 2015; Goldhammer et al., 2014). More specifically, we were interested in (1) time considerations (i.e. how much time participants spent on each part of the item and (2) their different response strategies. We found that in most cases it was possible to use observations of eye movements to build a picture of how the respondent was able to tackle the item, as well as whether they were successful or unsuccessful. With respect to time considerations, we could observe why some respondents took longer than others. For example, respondents took longer when they double-checked calculations before submitting answers. Our choice of a well-educated sample meant that most of the participants were able to obtain the correct answers on the PIAAC items. Nevertheless, we observed significant variation between participants who obtained the correct answer, in terms of their response strategies and their response times.

In observations of the literacy items, we could identify which parts of the test item respondents read, and we could distinguish between careful systematic reading of all sentences from more hasty skimming, similar to the distinction of 'effortful' vs. 'task-oriented' readers made by Solheim and Uppstad (2011).

[2] The use of reference points has a key cognitive advantage of reducing the demands of working memory (i.e. participants do not have to fully encode, in our case, the instructions but instead can revisit).

In the numeracy items, we were able to observe respondent strategies for obtaining the correct answer, and what had gone wrong when they obtained the incorrect answer. This information was obtained by observing paper-based written calculations and participants use of the calculator. This type of information is only obtainable through the use of head-mounted eye tracking systems (as opposed to desk-mounted eye tracking systems), and is a key innovation of our study. In problem-solving, we were able to observe how people cycled through web-pages and how they read and identified information before submitting their answer, a process which revealed information on the strategies they adopted while processing and solving the test items. The following trial-level examples illustrate, more concretely, these insights.

Detailed examples

The PIAAC 'Bottles' item asks the respondents to make a simple counting calculation based on a visual stimulus. It is a 'Level 1' numeracy item, classified in the lowest proficiency band. Eye tracking observations of this item showed that while all respondents in our sample obtained the correct answer on this item, they differed significantly in the time they spent on the item and in the strategies that they used to obtain the correct answer. Crucially, information on response strategies is only available via eye tracking observations rather than other measures of engagement such as log file response times. The two examples below illustrate this point.

Example 1: observing the gaze path in rapid response behaviour

Participant 10 in our study demonstrated what appeared to be a high level of numeracy proficiency as she completed this item. The analysis of response times on log files normally considers the 'start' of the item the moment that the stimulus is loaded. However, in this case, the respondent started to read the item instructions 2 s before the stimulus appeared on the screen. She initially read the instructions for 4.43 s. Then, the respondent viewed the stimulus image. As discussed above, the respondent momentarily tracked back to the instructions from 5.25 until 6.45 (less than 1 s) and then back to the stimulus. She completed viewing the stimulus and shifted her attention to the keyboard at 10.06. She typed the correct answer and submitted the answer at 12 s. Her response was within the most rapid 5% of submitted answers observed in the PIAAC item histogram (Fig. 1).

Example 2: engagement in slow response behaviour

Participant 9 in our study demonstrated a much slower response on this item and less proficiency as viewed in the eye tracking data. After the instructions and stimulus appears, she spent 12.23 s carefully reading the instructions. By this time, participant 10 had already submitted the correct answer. She then spent a further 12 s viewing the stimulus, using the mouse cursor as a tool to help to count the bottles. On the scan path video, it is possible to see that the participant fixates momentarily on each of the bottles as she counted them. Then, at 24.58, her gaze moved away from the stimulus and the computer to look above the laptop. We might consider this as 'thinking' time away from the stimulus (Glenberg et al. 1998; Doherty-Sneddon and Phelps 2005). We observed

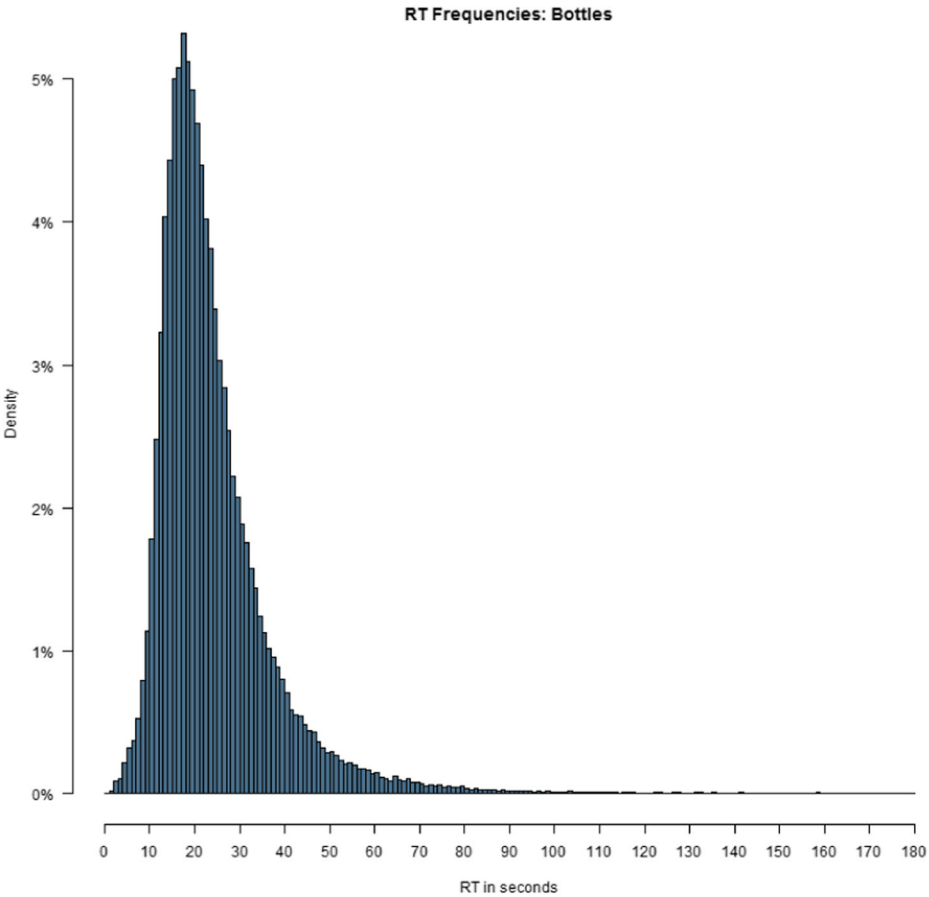

Fig. 1 Item response time histogram, for 'Bottles' problem (PIAAC), with density indicating percentage of total respondents against response times

five fixations at various 'off-screen' points, before she returned to look at the stimulus and the instructions again. At 39.09, she looked momentarily at the computer keyboard (keys 5 and 6). It appears that she was about to type the answer. Her finger hovered over key 4. Again, this type of information is only available via a head-mounted eye tracker. But then she paused and directed her gaze to the interviewer. The following verbal exchange took place for 19 s:

Respondent: Am I allowed to use that calculator?
Interviewer: Yep, sorry I should have .. okay, there's a calculator, paper, a pen and a ruler [
Respondent: [oh [
Interviewer: [and you are allowed to use them, um, as you wish.. (inaudible).. and the picture to the right hand side might be relevant for one of the questions[
Respondent: [Yeh
Interviewer: [and If it is it will be very obvious. It will say get the picture and do something with it.
Respondent: Yeh

Part way through this exchange, the respondent began to use the calculator (Fig. 2). She typed 4 × (multiplication key), then tracked back momentarily to the stimulus image, before continuing with the calculation. She then looked back again to the stimulus, paused and then typed her answer into the computer keyboard and submitted the answer at 42 s. She then said:

Respondent: That was scary! ((the respondent laughs quietly to herself)).

As this comparison demonstrates, the eye tracker data is able to provide a full record of the gaze and activities of the respondents as they complete the test item. That enables a comparison of their strategies and the time spent in each activity as they complete the item. We can see, for example, that it is plausible that respondent 10 was engaged and obtained a correct answer on this item within 12 s. We are also provided with substantive detail on what is going on 'between the clicks' on the computer as respondent 9 takes a much longer time to complete the item. Her response is among the longer response times indicated on the histogram in Fig. 1 (i.e. last quartile).

In the two examples, we can see that both respondents are 'engaged', and both successfully completed the item. However, as key funding 3 illustrates, we were able to observe some differences in ability across the participants. Those differences were observed in speed of calculations in numeracy items and in their reading behaviour in literacy items. While respondents may have obtained the same correct answer, the eye tracking data on scan paths can easily identify differences in their abilities and strategies through observations of item response processes. This kind of data could inform and corroborate the analysis of large-scale data from log files.

General discussion

There is much contemporary discussion about the contribution of various types of process data and how they can assist in the analysis of test and respondent performance and validity (e.g. Newton 2016; Ercikan and Pellegrino 2017; Zumbo and Hubley 2017). Our study has highlighted response process activity between the clicks that would not be captured by computer-generated log files. While log file data on response times and key strokes are easy and relatively inexpensive to capture, they are not sufficiently comprehensive or detailed on

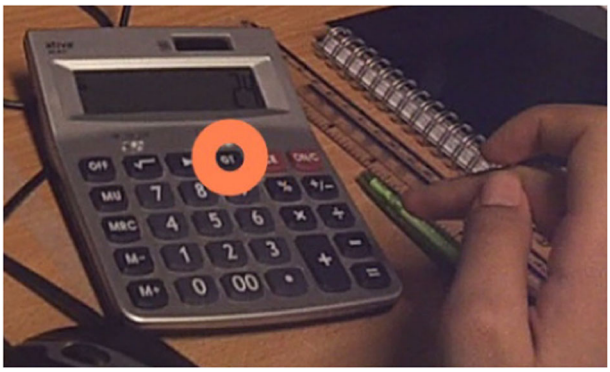

Fig. 2 Scan path image of respondent 9 using the calculator. The orange circle (produced by the eye tracker software) indicates the respondent's area of gaze

their own to capture and explain the substance of item response processes. This is contrary to the argument on log files put forward in the study by Oranje et al. (2017, p. 46). Instead, we might consider multiple sources of process data together (e.g. log files, eye tracking, emotion recognition, linguistic data and think aloud protocols) to inform our understanding of item response processes (e.g. D'Mello et al., 2017; Maddox and Zumbo 2017; Maddox 2017; Ercikan and Pellegrino 2017; Zumbo and Hubley 2017).

This exploratory study of the PIAAC assessment demonstrates the potential of using eye tracking to dig deeper into response processes in large-scale assessments and to identify patterns of behaviour (see also, Krstić et al., 2018).

We found that respondents were comfortable with wearing eye tracking glasses and that it did not interfere with their ability to complete the assessment. It is clear from our experience that eye tracking data can inform understanding of test item validity (e.g. by observing how respondents engage with items) and help to extend the test by differentiating more effectively between performance associated with higher or lower ability, or differences that may be attributed to developmental disorders such as dyslexia, or by unintended features of test item design or administration. As we have seen in this study, eye tracking data can help to identify behaviours that indicate higher or lower levels of respondent engagement, and to identify potential sources of disengagement (i.e. by identifying when disengagement occurs). Such methods could also be used to identify and profile behaviours that would indicate test fabrication (i.e. as departures from normal patterns of item engagement).

We have also shown that eye tracking observations are an accurate way of capturing respondent disengagement as they capture the substantive detail of disengaged behaviour such as identifying when respondents do not read the detail of the item question and stimulus. That kind of data, including statistical data on visits and timing of reading in areas of interest (AOI) such as item questions and key content in the stimulus could augment and inform the study of engagement in log file response time data.

Future directions and limitations

In this research, we followed standard PIAAC testing protocols and used the same virtual machine, laptop-type and sequence of test items that was used in the main PIAAC studies. However, the level of interaction between respondent and interviewer in our lab-based study was considerably less than in PIAAC assessments observed in the field. In ethnographic observations of household-based assessments in Slovenia observed by Maddox (2017), respondents frequently asked the interviewer questions about assessment procedures.

It is clear from this study that eye-tracking observations have the potential to fill in some of the missing information with regard to respondent engagement and 'off-screen' activity. Future eye tracking studies conducted in naturalistic 'in vivo' assessment environments (Maddox and Zumbo 2017) such as classrooms or households may therefore identify behaviours associated with respondent disengagement, 'mind wandering' and associated re-reading (Bixler and D'Mello 2016; Varao-Sousa et al. 2017). There is also scope for further research to investigate the impact on behaviour of wearing eye tracking glasses (see Risko and Kingston 2011), and the impact of variations in luminance on the measurement of pupil dilations.

A further theme relates to the diagnosis of the kinds of educational and developmental difficulties that the PIAAC programme might associate with low levels of ability (e.g. 'level 1'). Our study did not actively recruit respondents with low ability. However, it was able to observe some reading behaviours that would be typical of low performance and dyslexia.

Future eye tracking studies may be able to shed light on how well educational systems identify and intervene to remedy low reading performance and developmental disorders such as dyslexia, and to identify wider barriers to educational achievement.

Conclusions

This paper contributes to the recent expansion of interest in response processes in large-scale assessments (Zumbo and Hubley, 2017; Ercikan and Pellegrino, 2017). This exploratory study of the OECD PIAAC assessment has demonstrated how eye tracking can help to fill an explanatory gap, by providing detailed empirical observations on variation in item-level response processes that are not available from computer-generated log files, from conventional observation or from think aloud protocols.

The eye tracking observations contribute substantive information on respondent behaviour, including indicators of engagement. As we have shown, by profiling eye movement behaviour—e.g. scan paths, saccades and fixations, and by producing statistics on time spent reading AOIs, such as test item instructions or details of the stimulus, it is possible to produce nuanced models of behaviour associated with different levels of respondent engagement and respondent ability, and to identify respondents who exhibit behaviours associated with low performance, or with developmental disorders such as dyslexia.

The use of eye tracking techniques in large-scale assessment has until recently been limited to work on test item design and initial lab-based investigations into item performance (Paulson and Henry 2002; Tai et al., 2006; Solheim and Uppstad 2011; Hu et al. 2017; Oranje et al. 2017; and Krstić, 2018). However, the rise of research into 'processes data' in assessment (Ercikan and Pellegrino 2017; Zumbo and Hubley 2017) and advances in eye tracker technology (Bixler and D'Mello 2016; D'Mello et al., 2017) suggest that the application of eye tracking techniques can be extended to observations of how tested populations receive and engage with test items. The contribution of eye tracking in large-scale assessments relates to at least two distinct areas.

The first concerns the ability of eye tracking data to contribute to validity practice and validity judgements. They expand the information base on item response processes and what counts as data. This is in keeping with recent developments that expand the use of 'process data' in large-scale assessments (Shear and Zumbo 2014; Newton 2016). The value of such data should be judged not simply by its cost-effectiveness or its ease of collection but also by the ways that it can inform understanding of test performance. As we have demonstrated in this study, eye tracking offers very detailed and critical information about how respondents engage with test items. This includes answering questions about whether test items function as intended (Oranje et al. 2017) and about the quality of respondent engagement. Eye tracking data offers information on response processes 'between the clicks' that are not recorded by computer-generated log files, and as such can contribute to wider validity discussions about the way that process data such as log files, and other process data are used in large-scale assessment development.

A second (and related) question that this study raises concerns the potential for eye tracking data (and process data more generally) to operate as an extension of the test. As Oranje et al. (2017) have argued that process data from log files and eye tracking observations can be integrated into the design and analysis of performance in 'next generation' assessment. It is quite possible that with the rapid pace of technological development, eye-tracking data may become a mainstream part of assessment practice (D'Mello et al., 2017). As this paper has demonstrated, eye tracking data enables researchers to dig deeper into response processes and to investigate, profile and explain test-taker

performance. But, as process data (and para-data) becomes 'performance data', it is incumbent on test designers to research and understand the appropriateness, consequences and ethics of those decisions (Durrant and Kreuter 2013). Since there are currently no eye tracking studies that have observed large-scale assessment response processes in the wild, we are some way off having the necessary knowledge to properly inform such developments.

References

Beatty, J., & Lucero-Wagoner, B. (2000). Pupillary system. Chapter 6. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 142–161). Cambridge: Cambridge University Press.

Benfatto, N., Seimyr, G., Ygge J., Pansell T., Rydberg A., and Jacobson, C. (2016) Screening for dyslexia using eye tracking during reading. *PLoS ONE* 11(12): On-Line.

Bixler, R., & D'Mello, J. (2016). Automated gaze-based under-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction., 36*(1), 33–68.

Clifton, C., Jr, Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language, 86,* 1–19.

D'Mello, S., Diererle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist, 52*(2), 104–123.

Doherty-Sneddon, G., & Phelps, F. (2005). Gaze aversion: a response to cognitive or social difficulty? *Memory and Cognition., 33*(4), 727–733.

Durrant, G. and Kreuter, F. (2013). Editorial: The use of paradata in social survey research. *Journal of the Royal Statistical Society.* (2013) 176 Part 1, pp.1–3.

Ercikan, K., and Pellegrino, J.W. (2017) (Eds.) *Validation of score meaning for the next generation of assessments: the use of response processes*. Routledge.

Ferreira, F., & Henderson, J. M. (2004). Introduction to the interface of vision, language, and action. In J. M. Henderson and F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. ixxiv). New York: Psychology Press.

Gajewski, D. A., & Henderson, J. M. (2005). Minimal use of working memory in a scene comparison task. Visual Cognition: *Special Issue on Real-World Scene Perception*, 12, 979-1002.

Glenberg, A., Schroeder, J., & Robertson, D. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory and Cognition., 26*(4), 651–658.

Goldhamer, 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., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626.

Hawelka, S., Gagl, B., & Wimmer, H. (2010). A dual-route perspective on eye movements of dyslexic readers. *Cognition, 115*(3), 367–379.

Hu, Y., Wu, B., & Gu, X. (2017). An eye tracking study of high- and low-performing students in solving interactive and analytical problems. *Educational Technology & Society, 20*, 300–311.

Hubley, A.M., and Zumbo, B.D. (2017). Response processes in the context of validity: setting the stage. In B.D. Zumbo and a.M. Hubley (2017). (Eds.) *Understanding and investigating response processes in validation research*. Springer. pp 1–12.

Kahenman, D. (1973). Attention and effort. Englewood Cliffs, NJ: Prentice-Hall.

Kennedy. (2016). Eye tracking: a comprehensive guide to methods and measures. *Quarterly Journal of Experimental Psychology, 69*(3), 607–609.

Krieber, M., Bartl-Pokorny, K., Pokorny, F., Einspeiler, C., Langmann, A., Korner, C., Falck-Ytter, T., and Marchik, P. (2016). The relation between reading skills and eye movement patterns in adolescent readers: evidence from a regular orthography. *PloS One*, 11 (1), online.

Krstić, K., Šoškić, A., Ković, V., and Holmqvist, K. (2018). All good readers are the same, but every low-skilled reader is different: an eye-tracking study of the PISA data. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-018-0382-0 (in this issue).

Lai, M., Tsai, M., Yang, F., Hsu, C., Liu, T., Lee, S., Lee, M., Chiou, G., Liang, J., & Tsai, C. (2013). A review of using eye tracking technology in exploring learning 2000–2012. *Educational Research Review, 10*, 90–115.

Liversedge, S., Schroeder, S., Hyönä, J., & Rayner, K. (2015). Emerging issues in developmental eye-tracking research: insights from the workshop in Hannover, October 2013. *Journal of Cognitive Psychology, 27*(5), 677–683.

Maddox, B. (2017). Talk and gesture as process data. Measurement: Interdisciplinary Research and Perspectives, 15:3–4, 113–127.

Maddox, B., and Zumbo. (2017). Observing testing situations: validation as jazz. In B.D. Zumbo and A. Hubley (Eds.) *Understanding and investigating response processes in validation research*. Springer. pp 179–192.

Newton, P. (2016). Macro- and micro-validation: beyond the 'five sources' framework for classifying validation evidence and analysis. Practical Assessment, Research and Evaluation 21 (12), 1–13.

Oranje, Gorin, Jia and Kerr (2017). Collecting, analyzing, and interpreting response time, eye tracking and log data. In K. Ercikan, and J.W. Pellegrino (Eds.) *Validation of score meaning for the next generation of assessments: the use of response processes*. Routledge. pp 39–51.

Padilla, J. and Benitez, I. (2017). Cognitive interviewing and think aloud methods, in B.D. Zumbo and A.M. Hubley (Eds), *Understanding and investigating response processes in validation research*. Springer. pp 193–210.

Paulson, E. J., & Henry, J. (2002). Does the degrees of reading power assessment reflect the reading process? An eye-movement examination. *Journal of Adolescent and Adult Literacy, 46*, 234–244.

Pepper, D, Hogden, J, Lamesoo, K, Kõiv, P, and Talboom, J. (2016). Think aloud: using cognitive interviewing to validate the PISA assessment of student-efficacy in mathematics. International Journal of Research and Method in Education. 41 (1) pp. 3–16.

Radišić, J., & Baucal, A. (2018). Teachers' reflection on PISA items and why they are so hard for students in Serbia. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-018-0366-0 (in this issue).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Ren, P., Barreto, A., Gao, Y., & Adjouadi, M. (2012). Comparison of the use of pupil diameter and galvanic skin response signals for affective assessment of computer users. *Biomedical sciences instrumentation, 48*, 345–350.

Risko, E. F., & Kingston, A. (2011). Eyes wide shut: implied social presence, eye tracking and attention. *Attention, Percept and Psychophys, 73*, 291–296.

Shear, and Zumbo, (2014). What counts as evidence: a review of validity studies in educational and psychological measurement. In B.D. Zumbo and E. Chan (Eds). Validity and validation in social, behavioural, and health sciences.

Solheim, O. J., & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education, 4*, 153–168.

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research and Method in Education, 29*(2), 185–208.

Varao-Sousa, T. L., Solman, G. J., & Kingstone, A. (2017). Re-reading after mind wandering. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 71*(3), 203.

Zumbo, B. D. (2007). Three generations of DIF analyses: considering where is has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.

Zumbo, B.D. and Hubley, A. (2017). (Eds.) *Understanding and investigating response processes in validation research*. Springer.

Bryan Maddox. University of East Anglia. E-mail: b.maddox@uea.ac.uk

Current themes of research:

Dr. Bryan Maddox is a senior lecturer in education and international development at the University of East Anglia. He is a director of the Laboratory of International Assessment Studies and founder of Assessment MicroAnalytics ™. His research focuses on small-scale observations of assessment response processes using video-ethnographic methods and eye tracking.

Most relevant publications in the field of Psychology of Education:

Maddox, 2017). Talk and gesture as process data. Measurement: interdisciplinary research and perspectives. 15:3–4, 113–127.

Andrew P. Bayliss. University of East Anglia

Current themes of research:

Dr. Andrew Bayliss joined UEA in 2011. He gained his undergraduate degree and PhD in Psychology at the Bangor University and then took up postdoctoral fellowships from the ESRC, Leverhulme Trust, and the University of Queensland, Australia.

His research interests include social cognition, attention and action. He uses experimental approaches with methodologies including reaction time, eye tracking, motion capture (kinematics), electroencephalography and functional neuroimaging.

Most relevant publications in the field of Psychology of Education:

Edwards, S. G., Stephenson, L., Dalmaso, M., and Bayliss, A. P. (2015). Social orienting in gaze leading: a mechanism for shared attention. *Proceedings of the Royal Society: B. 282* (1812).

Piers Fleming. University of East Anglia

Current themes of research:

Dr. Piers Fleming is a Lecturer on the BSc Psychology programme. He joined the school in 2006 after 3 years working as a postdoc at the University of Nottingham. Dr. Fleming's work is focussed upon behaviour and judgments under uncertainty, including studies on risk perception, cooperative and altruistic behaviour. He is a cognitive psychologist, and a Chartered Psychologist (C.Psychol.) of the British Psychological Society (BPS). Dr. Fleming is an active researcher with expertise on risk judgements and decision-making and also works in the area of behavioural economics.

Most relevant publications in the field of Psychology of Education:

Fleming, P., Watson, S., Patouris, E., Bartholomew, K., Zizzo, D. (2017) Why do people file share unlawfully? A systematic review, meta-analysis and panel study. *Computers in Human Behaviour*, 72. pp. 535–548

Paul E. Engelhardt. University of East Anglia

Current themes of research:

Dr. Paul Engelhardt is a Lecturer in the School of Psychology at UEA. He completed a B.S. degree at the University of Nebraska Omaha, and MA and PhD degrees at the Michigan State University. After completing his education, Paul undertook a 2-year ESRC-funded postdoctoral research position at the University of Edinburgh. He then took a Lecturer/Senior Lecturer position at the University of Northumbria in Newcastle upon Tyne.

Most relevant publications in the field of Psychology of Education:

Engelhardt, P. E., Nigg, J. T., Ferreira, F. (2017). Executive function and intelligence in the resolution of temporary syntactic ambiguity: an individual differences investigation. *Quarterly Journal of Experimental Psychology*, 70 (1). pp. 1263–1281

S. Gareth Edwards. University of East Anglia

Current themes of research:

Dr. S. Gareth Edwards is a senior research associate (postdoc) at the University of East Anglia, currently researching the social cognitive processes relating to social eye gaze signals using a variety of behavioural and neuroscience methods. Gareth also has a keen interest in applying Psychology to Education, both practically and empirically, having focused his undergraduate and masters research projects of the 'testing effect'.

Most relevant publications in the field of Psychology of Education:

Dalmaso, M., Edwards, G., Bayliss, A. (2016). Re-encountering individuals who previously engaged in joint gaze modulates subsequent gaze cueing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 42 (2), pp. 271–284.

Francesca Borgonovi. University of East Anglia

Current themes of research:

Dr. Francesca Borgonovi is a senior analyst at the Organisation for Economic Cooperation and Development, Paris, with experience in comparative education, educational policy, educational psychology and educational assessment. She has extensive experience working on the OECD Programme for the International Assessment of Adult Competencies (PIAAC).

Most relevant publications in the field of Psychology of Education:

Borgonovi, F., & Pokropek, A. (forthcoming). Seeing Is Believing: Task-exposure specificity and the development of mathematics self-efficacy evaluations. J*ournal of Educational Psychology.*