International Journal of Research & Method in Education

ISSN: 1743-727X (Print) 1743-7288 (Online) Journal homepage: https://www.tandfonline.com/loi/cwse20

Think aloud: using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics

David Pepper, Jeremy Hodgen, Katri Lamesoo, Pille Kõiv & Jos Tolboom

To cite this article: David Pepper, Jeremy Hodgen, Katri Lamesoo, Pille Kõiv & Jos Tolboom (2018) Think aloud: using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics, International Journal of Research & Method in Education, 41:1, 3-16, DOI: 10.1080/1743727X.2016.1238891

To link to this article: https://doi.org/10.1080/1743727X.2016.1238891

Published online: 13 Oct 2016.

INTERNATIONAL JOURNAL OF RESEARCH & METHOD IN EDUCATION, 2018
VOL. 41, NO. 1, 3–16
https://doi.org/10.1080/1743727X.2016.1238891

Routledge
Taylor & Francis Group

Think aloud: using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics

David Pepper[a], Jeremy Hodgen[b], Katri Lamesoo[c], Pille Kõiv[c] and Jos Tolboom[d]

[a]School of Education, Communication & Society, King's College London, London, UK; [b]School of Education, University of Nottingham, Nottingham, UK; [c]Centre for Educational Innovation, University of Tartu, Tartu, Estonia; [d]The Netherlands Institute for Curriculum Development (SLO), Enschede, Netherlands

ABSTRACT

Cognitive interviewing (CI) provides a method of systematically collecting validity evidence of response processes for questionnaire items. CI involves a range of techniques for prompting individuals to verbalise their responses to items. One such technique is concurrent verbalisation, as developed in Think Aloud Protocol (TAP). This article investigates the value of the technique for validating questionnaire items administered to young people in international surveys. To date, the literature on TAP has focused on allaying concerns about *reactivity* – whether response processes are affected by thinking aloud. This article investigates another concern, namely the completeness of concurrent verbalisations – the extent to which respondents verbalise their response processes. An independent, exploratory validation of the PISA assessment of student self-efficacy in mathematics by a small international team of researchers using CI with concurrent verbalisation in four education systems (England, Estonia, Hong Kong, and the Netherlands) provided the basis for this investigation. The researchers found that students generally thought aloud in response to each of the items, thereby providing validity evidence of responses processes varying within and between the education systems, but that practical steps could be taken to increase the completeness of concurrent verbalisations in future validations.

ARTICLE HISTORY
Received 12 November 2015
Accepted 22 July 2016

KEYWORDS
Cognitive interviewing; think aloud; validation; assessment; PISA; international; self-efficacy; mathematics

Introduction

The Organisation for Economic Cooperation and Development (OECD) Programme for International Student Assessment (PISA) administered questionnaires with self-efficacy in mathematics (SEM) items to students aged 15 in 41 education systems in 2003 and 65 education systems in 2012. Self-efficacy is 'a judgment of one's capability to accomplish a certain level of performance' (Bandura 1986) and it is thought to 'determine the goals people set for themselves, how much effort they expend, how long they persevere in the face of difficulties, and their resilience to failures' (Bandura 1995, 8). Self-efficacy is therefore considered fundamental to personal motivation (Bandura 1997, 2010). As in PISA, assessments using self-report questionnaires have been the norm in self-efficacy studies (Bandura 2006). The AERA/APA/NCME (1999, 2014) *Standards for Educational and Psychological Testing* asserted the importance of integrating validity evidence of response processes into comprehensive validations of assessments. The OECD has not subjected PISA mathematics – student questionnaires to such

CONTACT David Pepper david.pepper@kcl.ac.uk School of Education, Communication & Society, King's College London, Waterloo Bridge Wing, Franklin-Wilkins Building, London SE1 9NH, UK

validations. This is of concern because of the political influence of PISA (Pereyra, Kotthoff, and Cowen 2011; Breakspear 2012) and the OECD's (2015) claim that: 'A sense of self-efficacy is essential if students are to fulfil their potential'.

Given the continued influence of PISA and the potential significance of SEM but the general lack of validity evidence of response processes for SEM assessments, there is a clear need for a more comprehensive validation of such PISA assessments. We report on part of an exploratory validation that investigated the potential of cognitive interviewing (CI) for gathering validity evidence of response processes for the PISA SEM items. CI involves a range of techniques for prompting respondents to verbalise their response processes for questionnaire items. We demonstrate how one technique, namely concurrent verbalisation or 'think aloud', can contribute to the validation of questionnaires administered to young people in diverse education systems through international assessments such as PISA. Our experience of using the technique should be of practical use to other researchers interested in investigating response processes for questionnaire items. Our use of this technique can be located within a wider methodological literature concerned with 'testing situations', incorporating analyses not only of item characteristics but also student characteristics, the educational context and the wider social context (Zumbo et al. 2015).

Rather than seeking to allay concerns about the reactivity of task performance to thinking aloud, which has been the focus of the literature (see, e.g. Fox, Ericsson, and Best 2011), we focus on the completeness of students' concurrent verbalisations. In this context, 'completeness' refers to the extent to which individuals verbalise their response processes while responding to items. Although concurrent verbalisations will to some extent always be incomplete, some completeness is important for producing validity evidence of response processes. Since little is known about response processes for SEM items or how the processes might vary between education systems, we provide examples of the validity evidence that could be produced in a comprehensive validation of the PISA assessment of SEM, or other international surveys of adolescents, using CI.

Background

Cognitive interviewing and Think Aloud Protocol

Survey respondents sometimes qualify their answers to questionnaire items, or even offer unsolicited feedback on them. This may be evident in verbal comments made during an interviewer-administered questionnaire or in comments written on a self-administered questionnaire. Such comments are incidental to questionnaire surveys and are generally disregarded, not least because they often come too late to influence the wording of items. However, we argue respondents' verbalisations can provide valuable insights into the functioning of questionnaire items.

CI is a method of pre-testing draft versions of questionnaires with sub-samples of respondents from target populations before final versions of questionnaires are administered to full samples (Willis 2005; Miller 2011). CI offers a range of techniques for systematically eliciting and recording individuals' comments on items when they respond to questionnaires. The major techniques include concurrent probing (questions asked during each item response), retrospective probing (questions asked after all item responses), and concurrent verbalisation ('think aloud' during each item response). The goal is to identify whether items function as survey designers intended. CI therefore provides a source of evidence for the confirmation, revision, deletion, or replacement of items.

Healthcare researchers such as Willis (2005) have been at the forefront of developments in CI and have influenced education researchers such as Karabenick et al. (2007). These researchers have used the technique of concurrent probing to identify problems with questionnaire items. Probes are generally used at three strategic points during each item response: when respondents have read an item (to check their comprehension); when they have reflected upon the item (to check what information they deem relevant); and, when they have responded to the item (to explain their response).

Although cognitive interviewers have generally preferred concurrent probing, the origins of CI are in the methodology of Think Aloud Protocol (TAP) which uses concurrent verbalisation instead (Willis 2005). Ericsson and Simon (1993) developed a theoretical model for TAP supported by their review of empirical studies. In this model, 'think aloud' refers to the concurrent verbalisation of thoughts as they become a focus of attention. Subjects are instructed to think aloud and their verbalisations are transcribed as a 'protocol', which is analysed to gain insights into cognitive processes involved in the performance of problem-solving tasks.

Ericsson and Simon (1993) preferred concurrent verbalisation because, according to their model, concurrent probing is more disruptive to task performance. Willis (2005) acknowledged that concurrent probes may produce 'local reactivity' (where probes about an item encourage respondents to identify spurious problems with the item) and 'extended reactivity' (where probes about one item encourage respondents to identify spurious problems with other items). Indeed, Ericsson and Simon (1993) found that studies requiring subjects to explain their approach to a task resulted in a more analytical approach to subsequent items.

Ericsson and Simon (1993) also preferred concurrent verbalisation to retrospective probing which, according to their model, is more reliant on respondents' recall. However, Leighton (2004) found that retrospective probes could clarify concurrent verbalisations. Willis (2005) suggested that retrospective probes could state: 'It seemed like you paused for a while on this question; could you remember what you were thinking?' but this highlights the reliance on respondents' recall. To be effective, particularly with younger respondents, retrospective probes should therefore immediately follow the administration of a set of items (Leighton 2004; Willis 2005).

Concurrent probing was originally developed for use with the interviewer-administered questionnaires in healthcare research. By contrast, self-administered questionnaires do not involve any interaction with the respondent. In this respect, self-administered questionnaires are more comparable than interviewer-administered questionnaires to the problem-solving tasks used in TAP studies. Concurrent verbalisation may therefore be more appropriate for identifying any problems with items in self-administered questionnaires. Furthermore, as Willis (2005) has noted, there is an argument for concurrent verbalisation when CI will inform not only the development but also the interpretation of survey results and it is consequently important for interviewers to avoid interceding in item responses.

Instructions to think aloud should not be complex or restrictive. Practice tasks unrelated to the target tasks, however, should be initially administered to subjects in order to reinforce the think aloud instructions. If subjects stop thinking aloud, interviewers need to avoid using prompts which unduly influence cognitive processing (Greene, Robertson, and Croker Costa 2011). Simple prompts to 'think aloud' or to 'keep talking' should be used; prompts such as 'tell me what you are thinking' are more likely to induce explanation (Ericsson and Simon 1993). When necessary to performance of the task, either because the task is demanding (such as for a novice) or because it is automated (such as for an expert), however, some completeness of verbalisations is sacrificed to avoid any reactivity of task performance to thinking aloud. Thinking aloud is therefore subjugated to the performance of the task (Ericsson and Simon 1993).

In respect of reactivity, Ericsson and Simon (1993) found that, when existing studies had used think aloud instructions which conformed to their model of TAP, the effect of concurrent verbalisation on the performance of a wide range of task types was not statistically significant. Fox, Ericsson, and Best (2011) updated the evidence base with a meta-analysis of 64 studies incorporating 94 effect sizes. The authors found that the effect of concurrent verbalisation on task performance was not statistically significant when think aloud instructions conformed to the Ericsson and Simon model of TAP. Lastly, although there were not enough existing studies to draw conclusions about the effect of different procedures or task types on response times, verbalisation generally appeared to prolong responses. According to the Ericsson and Simon model, this is attributable to the additional time required for verbalising information. To prevent reactivity, time constraints on tasks performed while thinking aloud should therefore be avoided.

Ericsson and Simon (1993) reported that existing studies had found some incompleteness of concurrent verbalisations, which was attributable to: reading aloud or attempting to understand written problems; experiencing difficulties in solving a problem; or, the presence of 'mediating steps' which were not the focus of the individual's attention and therefore not available for verbalisation.

Completeness of verbalisations and self-efficacy items

Both Messick (1995) and Kane (2006) proposed asking respondents to think aloud while responding to pilot versions of assessment instruments in order to gather evidence of their response processes. The AERA/APA/NCME (1999) Standards identified five strands of evidence for the validity of assessments: test content; response processes; internal structure; relations to other variables; and, the consequences of the test.[1] Castillo-Díaz and Padilla (2013) argued that the Standards offered too little guidance on methods of collecting validity evidence of response processes and found that this strand of validity is rarely sought but that CI could address this.

Bandura (2006) responded to the lack of guidance on assessment or validation in Bandura (1997) but was not informed by the Standards and did not call for validity evidence of response processes. Later, Usher and Pajares (2009) called for more information about the validity of assessments of self-efficacy. Indeed, the validation of self-efficacy assessments to date has generally been limited to evidence of relations to other variables, internal structure, and instrument content.

The translation of PISA SEM items into several languages provides a basis for gathering evidence of response processes in a diverse range of education systems. In the English language, the PISA mathematics – student questionnaire asks: 'How confident do you feel about having to do the following Mathematics tasks?'. Under this question, there are eight SEM items with a four-point response scale labelled *Very confident*, *Confident*, *Not very confident*, and *Not at all confident*. Each item comprises a reference to a task, such as 'Using a train timetable to work out how long it would take to get from one place to another', 'Understanding graphs presented in newspapers' or 'Calculating the petrol consumption rate of a car'. The items were apparently developed to emphasis problem-solving in real-world situations, as per the PISA mathematics framework.

The PISA mathematics – student questionnaire, including the SEM items, was piloted in 'a few participating countries to look at qualitative as well as some quantitative aspects of item responses' (OECD 2005, 40). It is not clear whether the piloting included evidence of student response processes and the results of the piloting have not been published. Hopfenbeck and Maul (2013) similarly identified a lack of validity evidence of student response processes for the PISA science – student questionnaire and used CI with concurrent probes for the PISA self-regulated learning in science items among students in Norway. Yet, as Miller (2011) has noted, evidence of response process produced by CI may be most valuable when multiple versions of questionnaires sensitive to the characteristics of diverse target populations are necessary. We therefore argue that PISA student questionnaires require more comprehensive validation – including validity evidence of responses processes.

Methodology

Sample

In this exploratory validation, which we conducted independently of the OECD, 30 students participated interviews in 2012, with 10 students in 2 schools in England, 10 in 2 schools in Estonia, and 10 in 2 schools in Hong Kong. Eleven students (boys and girls with a range of attainment) had already participated in pilot interviews in 2011–2012, with four taking place in England, three in the Netherlands, two in Estonia, and two in Hong Kong as we iteratively trialled and refined the interview procedure in each education system.

We asked mathematics teachers acting as points of contact in the schools to select students according to their age so that, as in PISA, study participants were 15–16-year-olds. We also asked the teachers to try to ensure a gender balance (except in one of the two Hong Kong schools, which was for girls), resulting in a sample of 18 girls and 12 boys. Lastly, we asked the teachers to try to ensure a broad range of prior attainment in mathematics was represented among the study participants, which resulted in an achieved sample of 10 students who teachers rated 'high', 11 they rated 'moderate' and 9 they rated 'low'.

The sample of students is small but the resource-intensive nature of CI means that typical samples are of 5–15 respondents (Willis 2005). Our inclusion of the four education systems where we had access to schools, represents only some of the diversity likely within these education systems and among the many participating in PISA. Future studies with more resources could therefore interview larger samples of students within each of the participating education systems.

Procedure

Since evidence from our CI was intended to inform not only the future development of PISA SEM items but also the interpretation of existing PISA SEM data, it was particularly important to minimise any reactivity of response responses to CI techniques. Given that an extensive literature indicated that the structure of response processes is not reactive to thinking aloud, our CI therefore involved concurrent verbalisation consistent with the Ericsson and Simon (1993) model of TAP. Since the literature suggests using retrospective probes for clarification of concurrent verbalisations as soon as a set of items is completed, particularly with young people, we also employed this technique.

We conducted interviews in the language of instruction of schools, on the basis of a protocol we developed in English, Dutch, and Estonian. PISA is conducted in Chinese in Hong Kong but the PISA manager in Hong Kong was unable to provide access to the SEM items used locally. With the assistance of colleagues at The University of Hong Kong, we therefore recruited two schools with English as a language of instruction. The lead author conducted the interviews in England and Hong Kong, the third and fourth authors conducted the interviews in Estonia, and the fifth author conducted the interviews in the Netherlands.

The purpose of the protocol was to facilitate a consistent approach across the cognitive interviews. The first part of the protocol included: our ethical procedure; a script to prepare students for the cognitive interview; instructions to use prompts to 'keep talking' or 'keep thinking aloud' if students were silent during any item response and retrospective probes to clarify verbalisations (at interviewers' discretion since we could not anticipate what would require clarification). The second part of the protocol included two sets of practice questionnaire items developed during the pilot interviews and subsequently used in England, Estonia, and Hong Kong. The third part of the protocol included the PISA SEM items in English available from the OECD website or versions in Dutch or Estonian made available by PISA national project managers.

The script to prepare students for the cognitive interview included our instructions about thinking aloud:

> We're interested in what you're thinking while you're responding to the questions. So, I'm going to ask you to think aloud as you work through the questions in each section. You can just say what you're thinking as you read and respond to the questions, as if I'm not here.

To encourage the students to think aloud without monitoring or explaining their verbalisations we added that:

> You might think aloud about things other than the questions, like what you're doing afterwards. That's fine. Don't worry if what you say to me makes sense or not. When you get to the end of each section, we can talk about your responses.

Finally, just before administering a set of items to the students, we asked them to:

> Work through this section and think aloud as you go along. If you get stuck on anything, don't worry, move onto the next thing. At the end of the section we'll discuss your responses.

In keeping with the PISA SEM items, our practice items were questionnaire items with a four-point Likert response scale. We designed the content of the practice items to be unrelated to the PISA SEM items, and undemanding. The practice items asked students how often they engaged in various activities during their free time, such as watching television or going to the cinema.

There were two sets of practice items: Practice Items A comprised four items administered to all students and Practice Items B comprised a further four items held in reserve for any students who did not think aloud during their responses to Practice Items A. This procedure enabled us to give students feedback on their thinking aloud at the end of Practice Items A and then administer either Practice Items B followed by further feedback, or the PISA SEM items. An example of interviewer feedback to a boy in England with low prior attainment in maths after Practice Items B (items e, f, g, and h) follows:

(1) **Student:** Uhuh. Very often [e]. Quite often [f]. Never [g]. Very often [h].
(2) **Interviewer:** Okay, and can you tell me why you selected those responses?
(3) **Student:** I selected this one because I listen to music every day. I read the
(4) book like when I have time. I hardly go shopping unless I'm going to buy
(5) myself something, which is never really, I just ask my mum to get it for me
(6) … well I have to give her the money … And then I play sport every week.
(7) **Interviewer:** Okay. And were you thinking about those things as you were
(8) ticking the boxes?
(9) **Student:** Yeah I was thinking about my football training today and then on
(10) Friday. Shopping I was thinking about the last time I went, which was way
(11) before Christmas … Then reading my book was like a couple of days ago. And
(12) the music was like lunch time actually.
(13) **Interviewer:** Okay that's great. I'm going to give you some more questions,
(14) could you try thinking aloud with those details as I give those to you as well,
(15) so you could tell me about all those things you're thinking about.

In line 1, the student concurrently verbalises his selected responses to the four items in Practice Items B (I listen to music, I read a book, I go shopping, I play a sport). This is a case of a student performing some reading aloud but not thinking aloud more broadly. His initial utterance of 'Uhuh' suggested he was not verbalising his thoughts even though they were in the focus of his attention. In line 2, the interviewer therefore probes for the student's reasons for selecting these responses and in lines 3–6, the student responds with his reasons. In lines 9–12, in response to the interviewer's query in lines 7–8, the student elaborates further. Finally, in lines 13–15, the interviewer asks the student to verbalise any such details during the PISA SEM items that follow. This is a case where a third set of practice items or more elaborated feedback after the first set of items might have been beneficial.

At their individual discretion, the interviewers administered Practice Items B to 18 of the 30 students in England, Estonia and Hong Kong (Table 1). The practice items were administered to more

Table 1. Number of students administered each set of practice items.

Practice items	A only	A and B
England	4	6
Estonia	6	4
Hong Kong	2	8
Total	12	18

students in Hong Kong than in England or Estonia, to more girls than boys and to more students with lower prior attainment in maths.

Data preparation and analysis

We audio-recorded students' verbalisations and video-recorded their item responses, which aided transcription. Transcripts were translated into English from Estonian and Dutch, and directly imported from MS Word into ATLAS.ti (Version 5.0) qualitative software for coding and analyses. Two types of coding provide the basis of the analysis. One was document coding, where the lead author assigned four 'families' to each interview, comprising the student's education system, school, gender, and prior attainment. The other was quotation coding, where each interviewer coded the transcripts of their interviews by questionnaire item, student item response, not thinking aloud (if a student did not verbalise anything during an item response, i.e. it was incomplete), and reading aloud (where the verbalisation was limited to reading aloud the item, i.e. partially incomplete). The lead author manually checked for a consistent approach to transcription and quotation coding during the pilot phase and the co-authors checked the document coding prior to collation and analysis of quotations.

The lead author used ATLAS.ti to collate the quotations for each of the practice and SEM item in separate Word documents, with the interview 'families' automatically attributed to each quotation. A first wave of analysis was suggestive of variation in response processes between students in the four education systems. The next step was to use the software to collate the quotations for each item in separate Word documents for each education system. A second wave of analysis confirmed variation between the education systems. The next section presents some results of these analyses.

Results

Quantity of verbalisations

Our analysis of the quantity of verbalisations during PISA SEM item responses is based on the 30 study interviews in England, Estonia, and Hong Kong with the finalised interview protocol.89% of the students' 240 item responses (30 students × 8 items) were accompanied by verbalisation (Table 2) but there was some variation between England and Estonia (93% and 94%, respectively) and Hong Kong (80%). Thus, verbalisations generally accompanied item responses in each education system, but to a greater extent in England and Estonia than in Hong Kong.

There was also variation in the extent of verbalisations between individual students' item responses. Eighteen students thought aloud during eight item responses and another seven students thought aloud during seven of their item responses (Table 3). Thus, 25 students thought aloud during at least 7 of their eight item responses.

The five students who did not think aloud during two or more of their item responses comprised boys and girls in England, Estonia, and Hong Kong with a range of prior attainment in mathematics (Table 4). The interviewers administered each of the five students Practice Items B, indicating that none of the five followed the think aloud procedure during Practice Items A. Four of the five students did not follow the think aloud procedure during their responses to Practice Items B either. Two of these students did not think aloud during each of their Practice

Table 2. Number and percentage of item responses with concurrent verbalisation.

Location	England	Estonia	Hong Kong	Total
Item responses	80	80	80	240
Verbalisations	74	75	64	213
Percentage	93	94	80	89

Table 3. Students' numbers of item responses with concurrent verbalisation.

Item responses	England	Estonia	Hong Kong	Total
8	8	7	3	18
7	1	2	4	7
6	0	0	2	2
5	0	1	0	1
4	0	0	0	0
3	1	0	0	1
2	0	0	0	0
1	0	0	0	0
0	0	0	1	1
Total	10	10	10	30

Items B responses (a low-attaining boy in England and a high-attaining boy in Hong Kong). The other three did think aloud but one did so without ticking any responses (a low-attaining girl in Hong Kong) and one stated that he was 'not accustomed' to thinking aloud (a low-attaining boy in Estonia).

Quality of verbalisations

Our analysis of the quality of concurrent verbalisations during SEM item responses distinguishes narrower, less complete 'reading aloud' from broader, more complete 'thinking aloud'. We quantify the extent of these two types of verbalisation in the study interviews in England, Estonia, and Hong Kong. We then give examples of thinking aloud from our CI in England, Estonia, Hong Kong, and the Netherlands, and show the value of retrospective probes for clarification.

In this study, students' concurrent verbalisations generally involved some reading aloud of the question ('How confident are you … '), the tasks (such as 'Understanding graphs … ') or the response categories (such as 'Very confident'). Seventeen per cent of item responses accompanied by verbalisations were limited to such reading aloud (Table 5). Reading aloud produced useful data about response processes in two ways. Firstly, it provided indications as to whether there were any problems with students' comprehension of items. This was evident, for example, when students paused and queried the meaning of 'timetable'. Secondly, reading aloud drew attention to problems with translations of items, and provided evidence of implications for item responses. For example, the official PISA translation of understanding 'graphs' in English was understanding *joonised* in Estonian, meaning 'figures' such as pictures (rather than *graafikud* meaning graphs), which confused several students, thus reducing their confidence.

Table 4. Students repeatedly not verbalising during self-efficacy item responses.

Country	Gender	Maths prior attainment	Practice Items A and B administered?	Practice Items B with verbalisation?	Self-efficacy items with verbalisation
Hong Kong	Boy	High	Yes	No	6
Hong Kong	Girl	Moderate	Yes	Yes	6
Estonia	Boy	Low	Yes	Yes	5
England	Boy	Low	Yes	No	3
Hong Kong	Girl	Low	Yes	Yes	0

Table 5. Number and percentage of item responses limited to reading aloud.

Location	England	Estonia	Hong Kong	Total
Item responses	80	80	80	240
Verbalisations	23	10	7	40
Percentage	29	13	9	17

Table 6. Number and percentage of item responses not limited to reading aloud.

Location	England	Estonia	Hong Kong	Total
Item responses	80	80	80	240
Verbalisations	51	65	57	173
Percentage	64	81	71	72

Table 7. Students' numbers of item responses with concurrent verbalisation not limited to reading aloud.

Item responses	England	Estonia	Hong Kong	Total
8	2	4	1	7
7	3	3	5	11
6	1	0	2	3
5	0	2	0	2
4	1	0	0	1
3	1	0	0	1
2	0	1	1	2
1	1	0	0	1
0	1	0	1	2
Total	10	10	10	30

Students' concurrent verbalisations generally involved more than reading aloud the items. Seventy-two percent of item responses were accompanied by verbalisations that involved more than reading aloud the question, the tasks or responses categories (Table 6). Seven of the eight item responses of 18 students were accompanied by more than just reading aloud the question, task or response categories (one half or more of the students in each education system; as reported in Table 7). In fact, 24 students verbalised more than just the text of the items for at least half their item responses (7 students in England, 8 students in Hong Kong and 9 students in Estonia).

Thinking aloud provided information about what Ericsson and Simon (1993) referred to as 'mediating steps' (which we refer to as 'mediating processes'). This information included the demands students inferred from the type of task referred to in each of the items. Contrasting task demands were evident in the concurrent verbalisations of students between education systems for the *Petrol* item and the *Newspapers* item. While responding to *Petrol*, one student participating in a pilot interview in the Netherlands outlined a simple procedure for calculating the consumption rate of a car, verbalising how:

> … you observe how many litres petrol it used on a certain distance, and then you do the certain distance divided by the number of litres petrol. Until you come out on a nice number …
>
> [Petrol – Very confident; NL Field School, Pilot Interview 1 – boy, moderate attainment].

This contrasted with some students in Hong Kong who inferred additional demands, resulting in a more complex task. One such student verbalised that:

> First I have to define what car this is, or maybe it is a Toyota for a family, or maybe it is a racing car … it depends … the rate is very fast if you are attending a contest like you are driving a racing car. So … because there are many factors …
>
> [Petrol – Not very confident; HK Peninsula School, Interview 5 – boy, high attainment]

The different task demands these two students inferred from *Petrol* are reflected in their item responses. While the student in the Netherlands (moderate attainment) selected *Very Confident*, the student in Hong Kong (high attainment) selected 'Not very confident'. Although it does not necessarily follow that students with high prior attainment respond confidently and students with lower prior attainment do not, it is noteworthy that these two students' item responses contrast with their prior attainment but are consistent with the different task demands they each inferred from the item.

Contrasting task demands inferred by students were also evident in the concurrent verbalisations of students in the same education system. Indeed, not all students in Hong Kong inferred more complex demands from the type of task referred to in each item. Responding to the Newspapers item, one boy in Hong Kong Peninsula School reasoned that newspapers need to be clear so that their readership can understand them without difficulty, and selected *Confident*:

> I think it's confident because most of the readers are … maybe some of them are adults, maybe some of them are old people and they can't use so complicated graph, or otherwise they can't really understand what it's all about.
>
> [Newspapers – Confident; HK Peninsula School, Interview 5 – boy, high attainment]

By contrast, one girl in Hong Kong Island School expected graphs in newspapers to be intentionally misleading, meaning that it would make it difficult for her to understand the 'true' situation, so she selected *Not very confident*:

> I think usually the graphs presented in newspaper may be … there will be misleading data, so that company will use [graphs] to mislead us.
>
> [Newspapers – Not very confident; HK Island School, Interview 1 – girl, high attainment]

Again, the contrasting task demands the two students inferred from the item, evident in their concurrent verbalisations, are reflected in their item responses.

Contrasting task demands were sometimes, however, not evident in concurrent verbalisations but in comments prompted by retrospective probes for clarification. For example, one student in England verbalised: 'Yeah, it's just … … yeah, confident', during her response to *Newspapers*. The interviewer noted the pause and, once the student had responded to the set of items asked:

> … And how about that graphs one … . I think you just said yes and ticked it, so I was wondering about the reasons.

In response, and echoing the reasoning of the boy in Hong Kong, the student commented:

> Yeah 'cos normally in newspapers … 'cos they're just for like the general public, they need to be quite simple, so that everyone can kind of read them quite easily … [Newspapers – Confident; EN Hill School, Interview 6 – girl, high attainment]

We therefore found that a combination of concurrent verbalisation and retrospective prompts resulted in more complete information about responses processes, which contributed to our finding that the only students who inferred more complex demands from the type of task in each item were in Hong Kong.

Discussion

Practice and prompts

Our international team conducted CI on the basis of a protocol, which helped us to standardise our interview procedure. In accordance with advice in the existing literature, our finalised protocol included practice items and prompts to think aloud. While 25 of the students thought aloud during at least 7 of their 8 SEM item responses, the remaining 5 students did not think aloud during 2 or more of their SEM item responses.

The characteristics of these five students (education system, gender, and prior attainment) were varied. However, four of the five students did not follow the think aloud procedure during both of the two sets of practice items we administered. Two of these students repeatedly did not think aloud, one thought aloud without ticking any responses and one expressed discomfort with thinking aloud. A question arises as to whether we could encourage or enable these and other students to follow the think aloud instructions. Since over-zealous use of prompts to think aloud could result in the reactivity of response processes to think aloud procedure, our discussion focuses on the practice items.

In our analysis of the quality of verbalisations, we distinguished narrower, less complete reading aloud of items from broader, more complete thinking aloud about mediating processes. We used this type of thinking aloud in response to *one half or more items* as a benchmark, which 24 of the 30 cognitive interviews met or exceeded. In the remaining six interviews, the students therefore thought aloud about mediating processes in response to less than one half of the items. It may be that no mediating processes were a focus of students' attention during this items or that interviewers could enable more students to think aloud about mediating processes for more items.

To attempt to increase the completeness of concurrent verbalisations, three steps could be taken. Firstly, interviewers could note that difficulties with a think aloud procedure can manifest not only as an absence of thinking aloud during item responses but also as: thinking aloud but not providing actual item responses (in this case, not ticking boxes); or, thinking aloud but making a more or less explicit verbal report of discomfort. Secondly, interviewers could have not just one or two but three sets of practice items available for administration to what is likely to be only a minority even of younger respondents. Each practice set need not contain many items; indeed, our use of four items per set meant that we could give timely feedback to students on their cooperation with our think aloud procedure. Thirdly, correspondingly, the protocol could help interviewers to ensure that appropriate feedback is always given on completion of a set of practice items by providing interviewers with bespoke prompts and probes in response to instances of thinking aloud, reading aloud or the absence of either type of verbalisation.

Although we administered two sets of practice items to more than one half of the students, one set was adequate for the remainder of the students, who thought aloud during at least seven of their eight item responses. While the benefit of administering at least one set of practice items to all respondents is therefore clear, we advocate a flexible approach to the administration of additional sets of practice items. This approach has the benefit of minimising the resource-intensity of CI while maximising the completeness of verbalisations.

Validity evidence and validation

With some exceptions (notably Karabenick et al. 2007; Hopfenbeck and Maul 2013), CI respondents are generally adults (Willis 2005). We are not aware of any existing studies using the concurrent verbalisation technique with adolescents for the development of questionnaire surveys, whether national or international.

Although we cannot be sure whether verbalisations included all thoughts in the focus of attention, we can identify absences for individual item responses. In our study interviews, verbalisations accompanied nearly 90% of item responses, with less than 20% of item responses involving only reading aloud and more than 70% of item responses involving more than just reading loud. When students simply read aloud items, we nonetheless identified issues with comprehension or with translation. When students thought aloud more completely, we identified students inferring different demands from the type of task in each item, notably students in Hong Kong inferring more complex tasks from the Newspapers and Petrol items. These issues are problematic for PISA because it could result in the under- or over-estimation of student SEM in different education systems.

We cannot be sure why verbalisation was less complete for a smaller proportion of responses in Hong Kong (80%) than England or Estonia (over 90%). This difference in completeness could reflect random error arising from our small sample or the effort of concurrently verbalising in English thoughts that were held in Chinese.[2] Indeed, Ericsson and Simon (1993) showed that respondents tend to stop verbalising during periods of cognitive intensity. According to Ericsson and Simon's model, in cases where respondents continued to think aloud, simultaneous translation would extend response times as students recoded from language to language. Response times for each item could be explored in future analyses between education systems. While such analyses would help to evaluate the Ericsson and Simon model, future studies seeking to validate assessment instruments should be resourced to sidestep this potentially confounding language issue. This issue should not affect our finding that students in Hong Kong inferred different demands from the types of tasks in each of the item, which appeared to be a function not of their language competence but of their mathematical competence (or at least their mathematics curriculum).

In our project, CI exemplified an approach to collecting validity evidence of student item response processes specifically for PISA SEM items in four education systems. The OECD could undertake CI with a larger sample of students in each education system to account for variation within each context. The OECD could also undertake CI in all participating education systems to account for variation between each context. Indeed, issues specific to each education system that were evident in our results indicate that secure inferences about response processes cannot be made from one education system to another. It is therefore not enough to pilot the PISA mathematics – student questionnaire in just 'a few participating countries' (OECD 2005, 40) or to conduct CI with tests but not questionnaires (see OECD 2005, 2014).

While acknowledging the limitations of our sample, we note that evidence of response processes even from our CI for a small sample of students provides an initial basis for interpreting Differential Item Functioning (DIF) in PISA data sets. Combining CI and DIF analyses would be consistent with 'Third Generation DIF' and its 'ecological approach', which shifts the focus of DIF analyses from 'test items' (as in PISA; OECD 2005) to the wider 'testing situation' (Zumbo et al. 2015). However, we also acknowledge that a different method such as ethnographic observation could be combined with DIF for an ecological approach focusing on testing situations (see Maddox et al. 2015).

With the notable exception of Karabenick et al.'s (2007) study of response processes for SEM items in the U.S.A., the validation of self-efficacy questionnaires has generally been limited to three of the other five strands of validity evidence (relations to other variables, internal structure, instrument content). The fifth strand of validity evidence, the consequences of the assessment, remains lacking from validations of self-efficacy assessments, including the PISA SEM assessment. However, per the argument-based approach to validation (Kane 2006, 2013), an explicit statement of the intended interpretations and uses of the assessment – currently absent from PISA documentation – is a prerequisite for an integrated evaluation of validity across all strands of validity evidence for the assessment.

Conclusion

This article contributes to the literature on the completeness of concurrent verbalisations during tasks, and quantifies the completeness in terms of more limited 'reading aloud' and more expansive 'thinking aloud'. In our CI conducted as part of an independent validation of SEM items from the PISA mathematics – student questionnaire, reading aloud proved useful in identifying issues with translation and comprehension. Thinking aloud was more valuable in identifying response processes which mediated reading an item and selecting a response option. Retrospective probes for clarification were valuable in illuminating response processes in instances when students only read aloud (~20% of item responses) or did not read or think aloud (~10% of item responses). Consistent with Leighton (2004), we therefore argue that concurrent verbalisation should be paired with retrospective probes.

The article also contributes to the literature on using CI as a method of collecting validity evidence of responses processes for assessment items, particularly questionnaire items administered to young people in international surveys. We showed how CI can provide evidence of variations in responses processes within and between education systems. Only a small minority of the 15-year-olds who participated in our CI in England, Estonia, Hong Kong, and the Netherlands repeatedly did not follow the think aloud procedure but we have recommended some modifications to the practice phase of the CI procedure with the aim of greater completeness. Although we sought to engage all of the students as 'experiential experts' on their item responses (Maddox et al. 2015), we cannot be sure whether these students were unwilling rather than unable to think aloud. Indeed, self-reported motivation to participate in international surveys varies between students (Eklöf 2010). Therefore, while we argue that CI is an effective method of collecting validity evidence of adolescents' response processes for questionnaires items, future studies could bear out whether the modifications result in greater completeness. As we have noted, CI is one of a number of methods of collecting validity evidence of response processes and different sources could be combined to provide more complete pictures of assessment situations.

Lastly, the article contributes to the literature on validating PISA assessments based on questionnaires, specifically the PISA assessment of SEM. We presented examples of the response processes for PISA SEM items which CI could produce if conducted as part of a more comprehensive validation within each participating education system. We argue that the OECD should conduct such a validation of the PISA SEM items, and indeed other PISA questionnaire items, ahead of their use in future PISA cycles. In the meantime, the variation in response processes we found between education systems suggests cautious re-interpretation of SEM results from PISA 2003 and 2012 is warranted. Until such time as these results are re-interpreted, policy-makers should reserve judgement and postpone action on communications such as OECD (2015).

Notes

1. The references to 'tests' reflect the predominance of standardised tests in educational assessment but the Standards are intended to refer to assessments more generally.
2. Although English was the language of instruction in the two Hong Kong schools which participated in this study, staff and students used Chinese in informal exchanges.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the UK Economic and Social Research Council under Grant [ES/H025448/1].

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *US: Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bandura, Albert. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, Albert. 1995. *Self-efficacy in Changing Societies*. Cambridge: Cambridge University Press.

Bandura, Albert. 1997. *Self-efficacy: The Exercise of Self-Control*. New York: W.H. Freeman.

Bandura, Albert. 2006. "Guide for Constructing Self-Efficacy Scales." In *Self-Efficacy Beliefs of Adolescents*, edited by F. Pajares and T. Urdan, 307–337. Greenwich, CT: Information Age.

Bandura, Albert. 2010. "Self-Efficacy." In *The Corsini Encyclopedia of Psychology*, edited by Irving B. Weiner and W. Edward Craighead, 1534–1536. Hoboken, NJ: John Wiley.

Breakspear, S. 2012. "The Policy Impact of PISA." OECD Education Working Papers. Paris: OECD.

Castillo-Díaz, Miguel, and José-Luis Padilla. 2013. "How Cognitive Interviewing Can Provide Validity Evidence of the Response Processes to Scale Items." *Social Indicators Research* 114 (3): 963–975.

Eklöf, Hanna. 2010. "Skill and Will: Test-taking Motivation and Assessment Quality." *Assessment in Education: Principles, Policy & Practice* 17 (4): 345–356.

Ericsson, K. Anders, and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT.

Fox, Mark C., K. Anders Ericsson, and Ryan Best. 2011. "Do Procedures for Verbal Reporting of Thinking Have To Be Reactive? A Meta-Analysis and Recommendations for Best Reporting Methods." *Psychological Bulletin* 137 (2): 316–344.

Greene, Jeffrey Alan, Jane Robertson, and Lara-Jeane Croker Costa. 2011. "Assessing Self-Regulated Learning Using Think-Aloud Methods." In *Handbook of Self-Regulation of Learning and Performance*, edited by B. J. Zimmerman and D. H. Schunk, 313–328. Abingdon, Oxon: Taylor & Francis.

Hopfenbeck, Therese N., and Andrew Maul. 2013. "Examining Evidence for the Validity of PISA Learning Strategy Scales Based on Student Response Processes." *International Journal of Testing* 11 (2): 95–121.

Kane, Michael T. 2006. "Validation." In *Educational Measurement*, edited by R. L. Brennan, 17–64. Westport, CT: Praeger.

Kane, Michael T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50 (1): 1–73.

Karabenick, Stuart A., Michael E. Woolley, Jeanne M. Friedel, Bridget V. Ammon, Julianne Blazevski, Christina Rhee Bonney, and Elizabeth De Groot. 2007. "Cognitive Processing of Self-Report Items in Educational Research: Do They Think What We Mean?" *Educational Psychologist* 42 (3): 139–151.

Leighton, Jacqueline P. 2004. "Avoiding Misconception, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing." *Educational Measurement: Issues and Practice* 23 (4): 6–15.

Maddox, Bryan, Bruno D. Zumbo, Brenda Tay-Lim, and Demin Qu. 2015. "An Anthropologist among the Psychometricians: Assessment Events, Ethnography, and Differential Item Functioning in the Mongolian Gobi." *International Journal of Testing* 15 (4): 291–309.

Messick, S. 1995. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50 (9): 741–749.

Miller, K. 2011. "Cognitive Interviewing." In *Question Evaluation Methods: Contributing to the Science of Data Quality*, edited by J. Madans, K. Miller, A. Maitland, and G. Willis, 49–75. Hoboken, NJ: Wiley.

OECD. 2005. *PISA 2003 Technical Report*. Paris: OECD.

OECD. 2014. *PISA 2012 Technical Report*. Paris: OECD.

OECD. 2015. *How Confident Are Students in Their Ability to Solve Mathematics Problems?* Paris: OECD.

Pereyra, M. A., H. G. Kotthoff, and R. Cowen. 2011. *PISA under Examination: Changing Knowledge, Changing Tests, and Changing Schools*. Rotterdam: Sense.

Usher, Ellen L., and Frank Pajares. 2009. "Sources of Self-Efficacy in Mathematics: A Validation Study." *Contemporary Educational Psychology* 34 (1): 89–101.

Willis, G. B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. London: Sage.

Zumbo, Bruno D., Yan Liu, Amery D. Wu, Benjamin R. Shear, Oscar L. Olvera Astivia, and Tavinda K. Ark. 2015. "A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding." *Language Assessment Quarterly* 12 (1): 136–151.