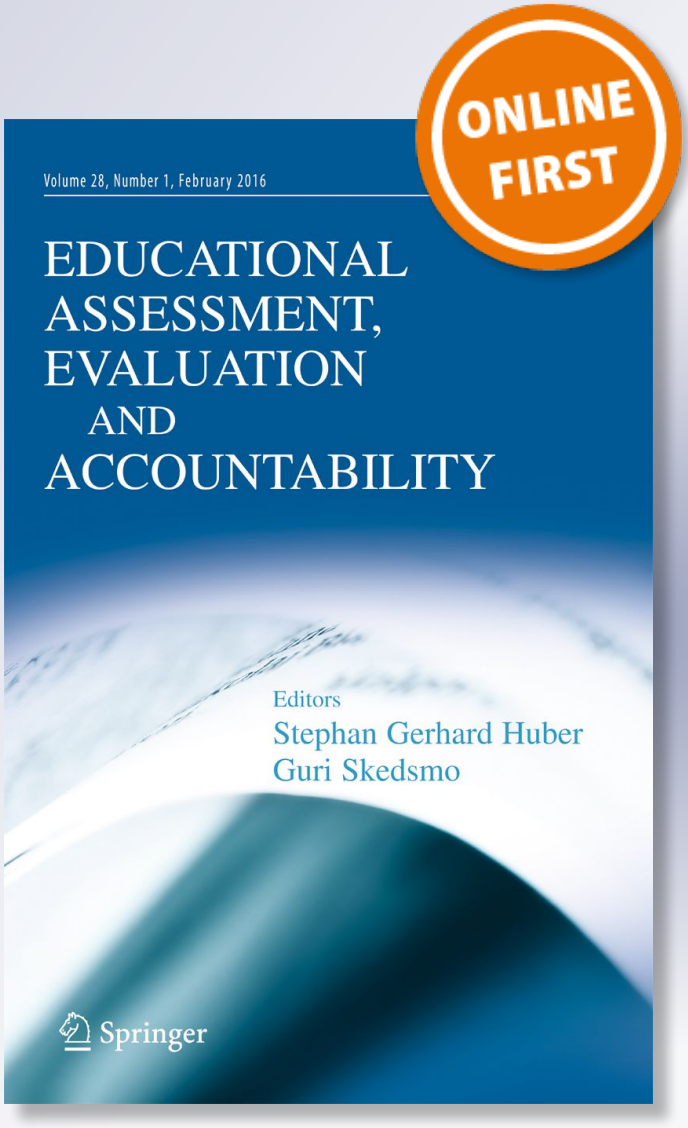
Instruments for school self-evaluation: lost in translation? A study on respondents' cognitive processing

Jerich Faddar, Jan Vanhoof & Sven De Maeyer

Educational Assessment, Evaluation and Accountability

ISSN 1874-8597

Educ Asse Eval Acc
DOI 10.1007/s11092-017-9270-4



Springer

Educ Asse Eval Acc
DOI 10.1007/s11092-017-9270-4

Instruments for school self-evaluation: lost in translation? A study on respondents' cognitive processing

Jerich Faddar[1] · **Jan Vanhoof**[1] · **Sven De Maeyer**[1]

Received: 16 August 2016 / Accepted: 11 September 2017

Abstract School self-evaluation (SSE), as an important leverage for quality assurance, often relies on surveys among staff members to collect information on the schools' functioning. The extent to which respondents cognitively process items as developers intended them determines the cognitive validity of SSE results. However, it is unclear what problems occur in respondents' cognitive processes which lead to cognitively invalid SSE results and how respondents' positions in the school affects these cognitive processes. This study draws on cognitive interviews conducted with 20 teachers and principals to understand their thinking process while answering an SSE survey. Cognitively invalid results were analysed using a content analysis to identify problems in respondents' cognitive processes. Findings showed that respondents experience semantic and syntactical issues when interpreting items. While elaborating, problems were found regarding items' topic and focus, particularly concerning whom to make a statement about. Issues also emerged in the response stage, especially that the 'don't know' option was not used as intended. Respondents' positions influence their understanding about whom a statement is required and how self-evident some items are to them. These problems should be taken into account by developers of SSE surveys and other instruments that intend to measure organisational characteristics.

Keywords School self-evaluation · Cognitive processing · Instruments · Survey design · Validity

✉ Jerich Faddar
jerich.faddar@uantwerpen.be

[1] Department of Training and Education Sciences, Faculty of Social Sciences, University of Antwerp, Gratiekapelstraat 10, Room 2.04, 2000 Antwerp, Belgium

Published online: 22 September 2017

1 Problem statement and conceptual underpinning

Over the past decades, school self-evaluation (SSE) has gained a prominent position in many educational systems in the evaluation of schools, being an important leverage for quality assurance and school improvement (McNamara and O'Hara 2005; McNamara et al. 2011; MacBeath 2005). SSE can be described as a process, in large part initiated by the school, whereby highly eligible participants systematically describe and judge the functioning of the school in order to make decisions or adopt initiatives within the framework of school development (Vanhoof and Van Petegem 2010). The procedure of SSE distinguishes between a description of the functioning of the school on the one hand, and a judgement of this on the other.

In order to create a description of the school as an organisation, there is a need to measure constructs at an organisational level. However, since schools cannot literally speak for themselves, this sets a methodological challenge. Therefore, SSE tends to rely on staff members to describe the school in which they are working with regard to well-considered aspects of the school's functioning. School staff have, as it is argued, a good insight into the functioning of the school from their day-to-day experiences (MacBeath and McGlynn 2002). Several instruments have been developed externally to schools and made available in order to facilitate the process of capturing such organisational characteristics, often grounded in school effectiveness literature (Vanhoof 2007; MacBeath et al. 2000). These instruments often ask staff members to fill in survey questions. By doing so, this method accounts for a multilevel approach as staff members (i.e. lower-level units) are providing information at the organisational level (i.e. a higher-level unit) (Kozlowski and Klein 2000; Bliese 2000).

The formulation of items can differ in design, while being appropriate to overcome a multilevel approach (Chen et al. 2004). Both a consensus design and a referent-shift design are commonly used (van Mierlo et al. 2009). The consensus design starts from the perspective of an individual making statements on collective properties (e.g. "I have a clear view of the job descriptions of other school staff") and the responses of all individuals in the organisation are subsequently aggregated onto the organisational level. The referent-shift design tends to capture organisational characteristics from an overarching perspective by asking respondents to make statements about the organisation which they are part of (e.g. "In this school everyone has a clear view of the job descriptions of other school staff"). While both item designs constitute a multilevel approach, as individuals are intended to generate statements about the organisational level, the referent-shift design requires respondents to think not only about themselves or their own behaviour but also about the organisation as a whole.

The use of surveys as a methodology reveals an ambition to identify a true reflection of respondents' perception on the schools' functioning (Guba and Lincoln 1994; Patton 2002; Cohen et al. 2011). Starting from that point of view, the challenge is to obtain data that reflect this perception with the least error as possible. Notwithstanding the frequent use of surveys, literature has already pointed to several problems that might be lurking beneath the surface as different kinds of errors might occur and data might be distorted as a result (Groves et al. 2009). Furthermore, it has already been questioned whether SSE instruments are sufficiently underpinned methodologically (Hendriks 2000). This raises a fundamental concern about the validity of the results from SSE surveys which are seen as necessary conditions (Kane 2006; Hofman et al. 2005), in

particular when schools rely on these results as a source of information for policy decisions and actions which can have a large impact on school processes and its outcomes (Scheerens and Bosker 1997; Scheerens 2000; Hofman et al. 2005).

One crucial element in obtaining valid SSE results is how items are cognitively processed by respondents (Bateson 1984; O'Muircheartaigh 1999). In other words, it is important to know how respondents are interpreting and reasoning while filling in SSE survey questions. Cognitive theories distinguish different crucial stages during the processing of items which conceal an interplay between the items and the respondents' memory (Schwarz 2007; Tourangeau et al. 2000; Karabenick et al. 2007). The different stages of this process are shown in Fig. 1. While, from a theoretical perspective, cognitive stages are ordered in the presented sequence, in reality respondents can shift from every stage to another (Ryan et al. 2012).

First, respondents have to be able to read and interpret the item, involving semantics, syntactics and pragmatics (Lenzner et al. 2010; Tourangeau and Bradburn 2010; Tourangeau et al. 2000). Semantics refers to the respondents' knowledge of words or technical terms, while syntactics refers to the grammatical complexity of sentences or syntactic ambiguity (i.e. items mapped onto multiple underlying representations). Pragmatics can yield problems, for example, when stylistic elements or other items near the items of interest hamper respondents in deducing the intent of the item.

Secondly, respondents have to retrieve relevant information from their memories (Karabenick et al. 2007; Schwarz 2007; Tourangeau et al. 2000). This search can consist of experiences, feelings, thoughts or perceptions that are stored in the autobiographical memory which are cognitively processed at the moment of survey administration (Karabenick et al. 2007). Two aspects herein are of importance. On the one hand, respondents have to connect with the *content* of the item. The *context* of an elaboration, on the other hand, refers to the level (e.g. an individual, a team or management) on which statements are asked to be made and the reference period respondents need to consider in their statement.

Lastly, respondents are expected to generate a judgement, based on the preceding cognitive stages. This judgement is then formulated into a response. Survey developers provide different formats in which respondents need to pronounce their judgement. Items can be designed in such a way that respondents get the opportunity to write down what they want, known as an open-ended format. By contrast, closed-ended formats are characterised by forcing respondents to make use of predefined answer options (Fowler

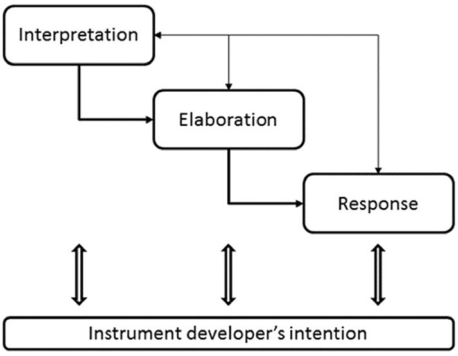


Fig. 1 Framework of cognitive validity

2014; Krosnick and Presser 2010). In this regard, each response option intends to catch the different statements a respondent might be willing to express.

The extent to which a respondent performs the cognitive processes of interpretation, elaboration and response in line with how the instrument developer intended them is referred to as cognitive validity (Karabenick et al. 2007; Koskey et al. 2010). The three critical stages in a cognitive process as described above can be used to assess the extent of cognitive validity.

Based on literature across different contexts, it can be argued that the way in which respondents interpret the task of filling in a survey depends on the underlying assumptions, values and knowledge they have available on the occasion with which respondents look upon it (e.g. Babik et al. 2015; Mohammed and Ringseis 2001; Cannell et al. 1981; Tourangeau et al. 2000). From a purely cognitive standpoint, these underlying assumptions, values and knowledge vary among individuals, resulting in an individual point of reference or mental model (Johnson-Laird 1983). Asking respondents to make a description or even a judgement on a statement implies that they make an appeal to this point of reference or mental model (Tourangeau et al. 2000; Cannell et al. 1981). They cognitively process information which is stored in their memory and make this explicit in the presented answering options.

It is often pursued in the context of SSE that opinions of different people are heard. As a consequence, participants holding different positions in a school such as principals, middle managers, teachers, administrative and technical personnel are asked to participate in the survey. Because of these different positions and roles in relation to the school's organisation, participants might have different points of reference (Edwards et al. 2006; Kozlowski and Ilgen 2006). This draws on the perspective that respondents with another function in the school have another background, different experiences and a different expertise that influence this point of reference. Ultimately, these differences can generate complementary information on the organisation's functioning (MacBeath et al. 2000; MacBeath 2005; Kyriakides and Campbell 2004). It is expected, for example, that middle managers or school leaders are more familiar with educational policy, management and administration in comparison to teachers due to additional training and experience (Day et al. 2009, 2010; OECD 2014). However, how these differences affect the cognitive process of respondents in answering an SSE survey is still unknown.

It is argued that arriving at accurate responses demands a lot of cognitive effort from respondents, and it has already been demonstrated that specific item formulations can increase the cognitive burden placed on respondents (Belson 1981). Moreover, literature shows that item complexity can lead to biases and distortions in responses (Fowler 1992; Knäuper et al. 1997; Lenzner 2012; Krosnick 1991). Next to the aspect of item design, it could be argued that differences between respondents can cause distorted survey results (Krosnick 1991). Nonetheless, it is unknown what problems in cognitive processes can lead to cognitively invalid results in the particular context of SSE surveys. SSE survey's complexity level can be increased by at least two aspects. First, with regard to item design, respondents are expected, in case of referent-shift items, to think both about themselves and about the school as a whole. Second, SSE surveys include an educational vocabulary which brings abstract and complex concepts that can increase the cognitive burden on respondents (Koskey et al. 2010). At the respondent level, it is readily assumed that SSE participants with a different function or position in

a school process the SSE items in the same way. However, it is unclear in what way this position can affect the cognitive processes a respondent executes when filling in an SSE survey. Despite methodological concerns, users and developers of SSE surveys seem to pass over the issue of cognitive validity rather readily, leading to a collective glossing over of this issue. This study aims to broaden the fundamental knowledge of how respondents cognitively process SSE surveys. In addition to theory development, these insights are beneficial in the identification of issues for the improvement of existing instruments and for the development of new instruments. Altogether, this study aims, by identifying possible flaws in respondents' cognitive processes, to increase the chance that items are cognitively validly processed as an important question for valid interpretations of SSE results. The current study focuses on the following research questions:

1. What problems can be identified during the interpretation, elaboration and response stage of the answering process of SSE items?
2. How does the position of individual respondents in the organisation influence their cognitive processes when answering SSE items?

2 Methods

2.1 Approach and technique

Given the exploratory nature of the research questions, this study draws on a qualitative approach. Gaining insights into cognitive processes sets a methodological challenge, and cognitive interviewing is found to be an exceedingly suitable technique to unfold respondents' thoughts underlying survey items (Presser et al. 2004; Ericsson and Simon 1993; Willis 2005; Ryan et al. 2012). The technique is based on many studies where respondents verbalise thoughts that occur in their working memory (where information from short-term and long-term memory are brought together) (Bradburn 2004; Conrad and Blair 2009). As advocated in different studies, this study applies a hybrid model of cognitive interviews, which means that both a think-aloud protocol and a concurrent systematic probing technique are used (Beatty and Willis 2007; Blair and Brick 2009; Collins 2003). As participants may experience problems in making their thoughts explicit during a think-aloud protocol (Royston 1989), they were given a brief introductory training in thinking-aloud, consisting of two exercises (Ericsson and Simon 1993). Respondents were asked to describe everything they see while mentally walking through their house and counting the windows. Secondly, they were asked to execute a multiplication out loud.

2.2 Instrument

Items of two exemplary scales from a real SSE survey were used as a case in the cognitive interview. One scale concerns the latent construct 'integrated policy in schools', the other 'reflective capacity of schools' (Vanhoof et al. 2011). Each item was formulated both in a consensus design (e.g. "I have a vision that exceeds my own job responsibilities") and in a referent-shift design (e.g. "In this school everyone has a

vision that exceeds one's own job responsibilities"). Respondents had to fill in the SSE survey by means of paper and pencil and were expected to indicate an answer option on a commonly used 4-point scale or a 'don't know' option. The full list of items is provided in Appendix 1. It is noteworthy that several steps were taken to safeguard the quality of the SSE survey when it was developed: a panel of experts within the domain of educational policy and evaluation made a critical review of the instrument, and a pilot test was undertaken with participants from the field.

2.3 The instrument developers' intentions: an illustration

In order to uncover what the items aim to tap into, the instrument developers were asked, prior to the conduct of this study, to formulate a precise description of their intention with each of the items. This was done by asking them to complete a written form for each of the items on the three cognitive stages of interpretation, elaboration and response that are critical in assessing cognitive validity, as suggested by Woolley et al. (2006). This resulted in cognitive validity criteria that were brought into the analysis as a point of reference.

This section outlines the instrument developers' intentions with the designed items for two exemplar items which will be referred to, among others, in the section "Results". The first example is formulated in a consensus design but can be treated in the results in its referent-shift form. In that case, the contextual aspect of the elaboration stage changes, as shown in example item 2.

2.3.1 Example item 1: I have a vision that exceeds my own job responsibilities

How the item should be interpreted:

> The respondent strives to think about his daily activities for an organisational objective that goes beyond her/his own core activities. One starts from the belief that working together in an integrated manner is necessary to achieve a greater purpose which would not have been realised as an individual within the organisation.

How respondents should elaborate on the item:

- With regard to the content: Respondents should adhere to a broad and integrated view of one's profession. As a school leader, this means that one exceeds the administrative aspect; as a teacher, this means that one exceeds the instructional aspect. Vision is to be understood as a realistic ideal that respondents can have. Such an ideal can be rather limited and involve only their own core activities. In this case, this vision does not exceed their own job responsibilities. It does when the individual's ideal and her/his actions meet a higher (organisational) purpose. A respondent should refer to examples such as a teacher who sees one's own communication about the school with family members as a form of marketing or a teacher who maintains discipline in a way that connects with broader policies concerning pupils' social and emotional guidance.
- With regard to the context: Respondents should refer to the current state of affairs (at the moment of filling out the survey) as it applies to him- or herself, by means of concrete examples.

How respondents are expected to use the provided answer options:

> The use of one of the 4-point scale options with labels ranging from 'totally disagree', over 'disagree' and 'agree', to 'totally agree' should reflect the extent to which a respondent agrees with the above-mentioned criteria. The 'don't know' option should only be used when the respondent has insufficient information about the current situation (e.g. when the respondent has recently joined the school, or when she/he could not provide any example). Doubt about the choice of an answer option when the respondent has relevant information should not lead to a 'don't know' answer.

2.3.2 Example item 2: In this school determining points for improvement is not seen as a threat

How the item should be interpreted:

> The item aims to detect an openness towards naming problems or matters that run sub-optimally. Respondents are not put off by (self-)reflection, and determining points for improvement is not seen as a personal failure.

How respondents should elaborate on the item:

- With regard to the content: Respondents name their doubts, difficulties and failures. The respondent is not afraid of identifying own weaknesses, and one dares to discuss with others what could be improved regarding their activities.
- With regard to the context: Respondents should refer to the current state of affairs (at the moment of survey administration), applied to the whole staff of their school (i.e. a pedagogical entity which is meaningful to the respondent), by means of concrete examples.

How respondents are expected to use the provided answer options:

> The use of one of the 4-point scale with labels ranging from 'totally disagree', over 'disagree' and 'agree', to 'totally agree' should reflect the extent to which a respondent agrees with the above-mentioned criteria. The 'don't know' option should only be used when the respondent has insufficient information about the current situation (e.g. when the respondent has recently joined the school, or when she/he could not provide any example). Doubt about the choice of an answer option when the respondent has relevant information should not lead to a 'don't know' answer.

2.4 Participants

Four primary schools participated in the study and were selected on the basis of a purposive sample in terms of school size (i.e. number of teachers). In each school, a cognitive interview was conducted with four randomly chosen teachers (where possible, someone in the middle management replaced one teacher) and the principal. In

total, 20 participants cooperated in our study. Respondents in each school were randomly allocated into two groups. One group started to think-aloud on one half of the items, continuing with the systematic probing on the second half. The other group started thinking aloud on the second half of the items, continuing with the systematic probing on the first half. With this interview design, all participants did not have to process all items with both protocols, which reduced the participants' workload by half. Furthermore, both groups started with the think-aloud protocol to avoid distortion of their spontaneous thinking which could occur due to the probing questions.

2.5 Analysis

Data consists of 400 observations; 20 participants verbalised their cognitive process on 20 SSE items. The analysis of the data was performed in two stages of coding. Firstly, all observations were coded for their cognitive validity based on the criteria for each of the critical cognitive stages (i.e. interpretation, elaboration and response) resulting in 1200 units of coding. A cognitive validity rating is allocated to each observation, ranging from 'cognitively invalid' (i.e. respondent says nothing in line with the developers' intention), through 'partially cognitively valid' (i.e. respondent does say at least one thing that is in line with the developers' intention), to 'cognitively valid' (i.e. everything the respondent says is in line with the developers' intention). In order to ensure the reliability of the cognitive validity coding, a second researcher independently recoded 13.5% of all observations. This resulted in a Cohen's Kappa of 0.62, which indicates a substantial level of agreement (Landis and Koch 1977). Next, a discussion led to a consensus upon the allocated codes. As a result, the analysing process could be continued. As this study aims to identify problems in respondents' answering process of SSE items, the analyses focus on respondents' verbalisations that are coded as 'cognitively invalid' and 'partially cognitively valid'. Altogether, 360 units of coding were included in the analysis.

This study does not pursue a representation of to what extent problematic issues in respondents' cognitive processes manifested themselves. The intention with this study is to identify recurrent problems in the respondents' cognitive processes across the different items under review, and therefore, the consecutive part of this study draws on the methodology of content analysis (Krippendorff 2012). This methodology is appropriate as it allows the possibility to create codes or themes which emerge from the data, next to the use of pre-existing categories or themes from the existing literature (Krippendorff 2012). Both deductive and inductive coding approaches were applied in this study. Each of the observations was searched in-depth for problems that made a cognitive process not to be in line with the instrument developers' intention. Codes were clustered in themes which correspond with problematic respondent behaviour in terms of cognitive validity. To ensure the reliability of the coding, a second researcher independently recoded the data during the analysis for respondent behaviour, resulting in a Cohen's Kappa of 0.66.

3 Results

This part of this section provides an answer to the first research question which searches for problems in respondents' cognitive processes when answering SSE items.

3.1 Interpretation stage

Different problems occur when respondents are trying to make an interpretation of items, leading to a discrepancy between the intention of the instrument developer and the actual interpretation.

3.1.1 Lost in giving meaning to words

First, at the semantic level, respondents experienced difficulties with terms or concepts that are formulated in the items. Despite the fact that these terms or concepts are meaningful, and important for an accurate understanding of the items, it happened that respondents were unfamiliar with them. This was the case not only for specific educational terms but also for some rather common terms. As a result, respondents did not know the term while trying to make sense of the whole item. With example item 1 "In this school everyone has a vision that exceeds one's own job responsibilities" (see "The instrument developers' intentions: an illustration"), it is readily assumed that respondents understand the central concepts 'vision' and 'job responsibilities'. However, a proper interpretation of these concepts turned out to be not self-evident. The meanings of these terms were explicitly doubted by some respondents. The same problem was experienced by respondents while processing other items. Terms such as 'collective reflection' or 'administrative activities' also lead to a problematic interpretation stage for some respondents.

In this school everyone has a vision that exceeds one's own job responsibilities.

> Phew, a vision that exceeds one's own job responsibilities. Well, I think that's a rather vague question. (…) Um, it, I don't know, it means like that you do more than your job description says? And what are those job responsibilities exactly? (…) (Respondent B, teacher, school 3)

I have a positive attitude towards collective reflection.

> That is a difficult question. I am not sure what is meant by collect, collective reflection. So I'm being asked if I have a positive attitude towards it but I cannot respond to it as I don't know. What is collective reflection? Talking about something in a group? Does it mean writing something together, formulating a reflection about a certain topic?(Respondent B, teacher, school 1)

I know about the administrative activities.

> I don't understand. I'm not going to answer this. I don't know what is meant by administrative activities. I'm just going to put a question mark next to it, I'll… I don't know if I should fill it in. I'm going to leave it open and just put a question mark next to it because I don't understand this question. (Respondent A, principal, school 1)

It also happened that respondents recognised the concept but were not able to give the word a meaning that fitted the context in which it was used. One item asked respondents whether they collect data on their own functioning. A respondent could,

possibly, refer to student achievement results or feedback of the school leader on their instructional competence. However, the task of trying to find out the intention of the term 'data' was perceived as a difficult one, and could only be guessed upon by some respondents. One respondent referred to gossip and opinions that had been spread about the functioning of other colleagues as an interpretation of 'data'.

In this school initiatives are taken to collect data on one's own functioning.

> Data? Er, things that are being said by others? Data, well … er, yes, I'm thinking. (…) those data … something they say about me, about me… what they experience of me? I don't know … It's a word that has different meanings, isn't it? Data … No, I read more into it than just what is being said about about oneself because (…) (Respondent E, remedial teacher, school 4)

Furthermore, respondents do not always manage to interpret the appropriate scope of a phrase. At first sight, the item "I have a clear view of the job description of other school staff" is not expected to create many difficulties for respondents. Still, an appropriate interpretation of the phrase 'other school staff' is often violated by respondents. While the instrument developers were aiming for respondents to think broadly about the school team, most respondents were only thinking about staff members who are involved in the school's primary processes. Administrative and technical staff were not considered to be part of the school team.

3.1.2 Mixing up the sentence structure

Next to semantic problems, respondents were confronted with problems at a syntactic level. Syntactic issues refer to problems where respondents struggled with the sentence structure of items. For example, some respondents had problems with an item that was formulated with a relative clause by means of a relative pronoun. While reading the item, the relative pronoun seemed to be ignored by some respondents. As a result, some respondents were found to be mixing up different parts of the sentence. While the instrument developers were, with example item 1, aiming to collect information about whether a respondent had a vision that went beyond their own job responsibilities, some respondents thought it wanted information about a vision on a too extensive workload.

In this school everyone has a vision that exceeds one's own job responsibilities.

> In this school um, oh my … That's a difficult one. I believe, yes, a vision … In this school everyone has a vision that exceeds one's own job responsibilities. In this school most people think … that um, work, job responsibilities um, have significantly increased compared to years before. (Respondent B, teacher, school 4)

3.2 Elaboration stage

After having completed an interpretation of the item, a respondent continues his/her cognitive tasks by making an elaboration. Problems in respondents' elaborations are categorised in two aspects: content and context.

3.3 Content-related problems

Although items are manifest indicators which represent a latent construct (e.g. reflective capacity), each item taps into a specific aspect of that construct with a different content.

3.3.1 Expanding

Concerning the content of an item, it was found that some respondents were thinking of information that was broader than what the instrument developers were aiming for. The information on which they were relying in order to make a judgement was related to the item's content, but they also based their judgement on information that was not. These respondents were expanding the scope of the item by including information which that specific item was not probing for. One example is found with regard to example item 1, where a respondent also pointed to the effects of determining points for improvement. The respondent mentioned that determining points for improvement leads to a more extensive workload and that changes are not embedded in the school in the longer term, whereas the item wanted to know whether there is an openness to determining them, regardless of their effects or whether they are embedded long term.

I do not experience determining points for improvement as a threat.

> Er… phew, sometimes I think by myself: Oh my, well um … that will bring a lot of extra work and um … Or for example I sometimes think by myself: Yes, now is OK, but the next month this feeling will already have watered down. Then, well sometimes it is threatening in the sense that it once more involves a lot of, a lot of er… additional work. (Respondent E, teacher, school 2)

3.3.2 Narrowing down

Some cases were characterised by respondents who did not fully cover the content of an item while elaborating on it. While providing information on the topic of the item, crucial elements regarding the specific item were not taken into consideration. The item's topic was narrowed down, which resulted in partial information where respondents' judgments were based on. For example, developers of the item "In this school everyone has a clear view of the job descriptions of other school staff" wanted to know whether respondents knew who to approach with certain questions. For the instrument developers, it suffices that there is an informal allocation of tasks among the staff members for respondents to agree with the statement. However, some respondents thought only about a formal document that states their individual tasks and responsibilities in order to (not) agree with the statement. Another item "I take initiatives to collect data on my own functioning" included a respondent who only considered whether or not she/he made the effort to reflect on her/his own functioning. Individual reflections were the only source of information on which the respondent based his/her judgement on the item, while the instrument developers' intention was broader. Other sources of data could also have been taken into consideration such as student achievement results or feedback from colleagues or students.

In this school everyone has a clear view of the job descriptions of other school staff.

> A clear view of the job descriptions of other school staff? There are no job descriptions in this school. Our management has never undertaken this and everyone knows it from each other, so... No. (Respondent A, teacher, school 4)

I take initiatives to collect data on my own functioning.

> So, me taking the effort to write down um, well, to reflect on myself and to keep track of it, that's what I think. And that I do, yes that's about class practice again. Um, in my lesson plans I write things that are open to improvement. That is to say, I write my own tips for the next time I'll be giving that lesson. (Respondent B, teacher, school 3)

3.3.3 Elaborating out of scope

The next problem that emerged from the data was an elaboration that was out of scope. The information instrument developers were hoping to obtain with an item was not captured. Some respondents retrieved information from their memories which did not relate to the item's topic. Instead, respondents provided information concerning other constructs which could characterise their working environment or organisation. With regard to example item 1, it was demonstrated that a respondent extensively considered the way in which points for improvement are dealt with within the school team, while nothing was mentioned about the attitude of the school team regarding the determination of points for improvement in itself. On the item "In this school everyone observes other people's performance", one respondent elaborated with information about a school climate wherein every team member kept a close, controlling, watch on others' activities. The instrument developers, however, were aiming for information about the extent to which staff members were visiting and observing each other's lessons as a means for learning from each other.

I do not experience determining points for improvement as a threat.

> I have to add to this that er I also needed to learn it the hard way. That I cannot quickly carry out small improvements in passing, but that I have to put them on the agenda of a meeting. And then let people who want to say something about it have their say. Rather than discussing, defining and establishing a structure. (Respondent C, principal, school 3)

In this school everyone observes other people's activities.

> In this school everybody is watching everybody. Haha. I, um, fully agree with this one. Everyone, no wait, not everyone but most people here know other people's schedule by heart, perfectly know how many minutes of surveillance everyone has done or how much surveillance they have not done. So yeah, people are very er, yes, alert to er, unfair practices, or that is to say, not necessarily unfair but... what feels like unjust to them. (Respondent A, teacher, school 4)

3.3.4 Making assumptions

Some cases were characterised by respondents who did not rely on facts or experiences from their autobiographical memories to base their judgement on. In these situations, respondents were making assumptions on their behaviour and were relying on them in their elaboration. For example, a middle manager, when answering example item 1, stated that her/his judgement was based on the assumption that she/he reflects on matters that were not her/his responsibility. The middle manager did not make any reference to a concrete example to support this assumption. Another respondent, when answering the item "In this school everyone has a clear view of the job description of other school staff", was making a judgement on having the opportunity to consult the job description of others as these had been sent by the school leader by e-mail.

I have a vision that exceeds my own job responsibilities.

> Disagree. I didn't take totally disagree because there are probably things which I reflect upon that are not my responsibility. (Respondent E, middle manager, school 1)

In this school everyone has a clear view of the job descriptions of other school staff.

> Yes, I fully agree, that job description, I can do it perfectly now, we all received them from each other. So I know, I could trace perfectly, er, what the job descriptions are of all my other colleagues. (Respondent A, teacher, school 3)

3.3.5 Relying on preceding elaborations

While every administered item was intended to investigate specific content, respondents did not always notice the specificity of each item. It was found in a few cases that respondents did not discern any difference from other items, and were not engaging in a search for relevant information on that particular item. They relied on the elaboration made for a preceding item and were basing their statement thereon.

In this school everyone has a clear view of the job descriptions of other school staff.

> Oh, well, to me that's the same question as the first one, so yes, I agree. (Respondent C, middle manager, school 4)

3.4 Context-related problems

As argued in the conceptual underpinning of this study, referent-shift design items, in contrast to consensus design items, place a larger burden on the cognitive processes of respondents. Respondents are expected to make a statement about a higher-level unit (e.g. the school as a whole), which requires them to think not only about themselves but also on a higher level.

3.4.1 Referent-level problem

Results show that in some cases respondents made statements on another referent than those intended. A referent-level problem refers to a phenomenon where respondents were mistaken about what level of the organisation a particular item was asking for a judgement about. For example, if an item aimed for a judgement at the school level, the respondent ended up thinking only about or him- or herself in order to make that judgement. It also happened that respondents, although focussing on the school level, redirected the scope to the level of the management which did not meet the intentions of the instrument developer either.

In this school everyone has a critical attitude towards their own actions.

> Well, I still am rather reflective, but maybe not that fierce as I used to be (Respondent A, teacher, school 3)

In this school determining points for improvement is not seen as a threat.

> Yes, in my opinion that is a question that is rather difficult to interpret. (…) Um, that reminds me of the questions we had on our performance appraisal. It contained a few elements that I experienced as threatening. I don't know if this was intended. But, yes, I experienced it that way. (Respondent B, teacher, school 3)

In this school everyone has a critical attitude towards their own actions.

> Um.. Does that means the management or the team? I assume the principal, and then I say 'disagree'. (Respondent B, teacher, school 1)

With regard to the referent, another restriction was found in the cognitive processes of respondents when they were elaborating on referent-shift items. Respondents did not seem to consider a school team in the broad sense as intended by the instrument developers. Respondents' cognitive processes only focussed on the staff members involved with the primary processes at school. Administrative and/or technical staff was not considered to be part of the school team by respondents when filling in the SSE survey.

In this school everyone has a vision that exceeds one's own job responsibilities.

> Yes, we've also had the Inspectorate's evaluation. We have to rewrite this manual of World Studies all by ourselves, so to speak. We now also have to make and develop and write out our art lessons ourselves. So we are already writing two manuals, actually. So I really do know what my, my colleagues' and my school's opinion is about, yes, that we have an extensive workload. (Respondent C, teacher, school 1)

3.4.2 Reference-period problem

While items in this study were only intended to describe the current state of affairs, it happened at times that a response was based on information beyond the

intended timeframe. This problem is referred to in this study as a *reference-period problem*. Respondents were thinking about actions or experiences that took place in the distant past.

In this school everyone has a clear view of the job descriptions of other school staff.

> Yes, now I'm going to select 'agree', not 'totally agree', because there has been a conflict once. There has been an argument with the maintenance staff. (Respondent A, teacher, school 2)

3.5 Response stage

3.5.1 Don't know

With regard to respondents' task of formulating a response, a series of problems occurred. The first problem was related to the use of the 'don't know' option provided. Its use was only intended for instances where a respondent appeared not to have any relevant information on the topic of the item (e.g. when a teacher had only been working in the school for a few months and consequently did not have enough relevant information to make a judgement). While assuming that the formulation 'don't know' would satisfy this intention, it happened that in some cases respondents selected the 'don't know' option while there was clear evidence that they indeed had been making a judgement in their mind based on relevant information. Another problem that occurred was that respondents who did not succeed in interpreting the question relied on the 'don't know' option.

In this school everyone has a vision that exceeds one's own job responsibilities.

> Has a vision that exceeds one's own job responsibilities? Huh? What? A vision that exceeds one's own job responsibilities? Yes, I know our vision and it exceeds, huh I don't get it. I'm going to indicate 'don't know'. That's a little too abstract for me. (Respondent E, middle manager, school 1)

3.5.2 Lacking an answer option

Some respondents experienced a lack of specific answer options. As the instrument developers did not insert a neutral category, respondents experienced difficulty in expressing their mental judgements. Often, respondents had arguments both in favour of and against the phrasing of the item. Another issue found in respondents' verbalisations was the lack of an open-ended answering box. Next to the predefined response options, they wanted to provide additional information which would have been relevant to their opinion. Nevertheless, respondents were forced to indicate a closed-ended response option provided by the instrument developers which did not truly mirror their mental judgement.

In this school determining points for improvement is not seen as a threat.

> I would have loved being able to write some extra explanation here. To write that they, that they don't perceive it as threatening when, um, it comes from themselves. (Respondent A, principal, school 1)

3.5.3 Divisive element

It also happened that in some cases more than one answer option was selected, which meant the answer became uninterpretable. One principal identified two distinct groups in his school team and said that the item applied to one group but not the other. Consequently, the principal was not able or willing to reflect this conclusion in one response option. As a result, the respondent indicated two contradictory options: agree and disagree.

> In this school everyone has a critical attitude towards their own actions.
>
> One does, the other does not. Has a critical attitude towards their own actions. Er, I'm going to select agree and disagree. (Respondent C, principal, school 3)

3.5.4 Using answer options in reverse

A final phenomenon was the selection of an answer option that reflected the opposite of the spirit of a respondent's elaboration. Moreover, the opposite use of predefined answer options which are supposed to reflect respondents' judgements contradicts the instrument developers' intentions. While everything a respondent was saying would fit an answer that would affirm the application of the item (i.e. *agree* or *totally agree*), the selected response option represented the opposite. Possibly, the response options were, erroneously, used in reverse to the manner intended by the instrument developers.

3.6 On the respondent's position in the school

The following paragraphs focus on the second research question, which aims to identify how respondents' cognitive processes can be influenced by their position in the school. Results show that there is indeed an influence, either in a positive or negative way. However, not every aspect in the cognitive processing of items is affected by this position.

3.6.1 No particular influence

Although respondents held different positions within their schools, no particular influences were found with regard to some issues like having difficulties in giving meaning to words. It could be expected from principals, for example, that they would be familiar with a more extensive vocabulary with regard to policy, management and school administration. However, the data shows that they had difficulties in interpreting items just like other respondents holding other positions in the school. Furthermore, in some cases, it was found that principals struggled with difficult phrases as well.

3.6.2 Self-evident items

Results show that certain answers were felt to be self-evident by some respondents because of their position. They did not therefore think about examples that would support the statement they made in response to the item. Instead, they only

elaborated on items by referring to their position in the school or the larger structures in which they operated. Consequently, they concluded that the content of the item was self-evident.

I have a vision that exceeds my own job responsibilities.

> I have a vision that exceeds my own job responsibilities. Um, yes as a principal you don't leave at four o'clock. Of course you have to support your school, and uphold the vision of your comprehensive school, which can be … I'm also a member of a task force on vision and vision development and implementation. So I agree with this one. (Respondent B, principal, school 2)

3.6.3 Excluding the self

As principals and/or middle managers were asked to make statements about a higher-level unit, some of them tended to make only a judgement on the team they manage. Although principals and middle managers were viewed as a part of the higher-level unit, and therefore should also consider themselves in making a statement, they excluded themselves from their judgement. For example, one item focused on the extent to which the respondent takes initiative to collect data on their own functioning. A principal mentioned that no staff members except for one had asked for a performance appraisal. This principal did not consider to what extent she/he collected data on her/his own functioning, although the principal is also part of the school.

In this school initiatives are taken to collect data on one's own functioning.

> No, as in the previous question, I can say that we provide feedback regularly. But there is almost no one, there is one person out of 55, there is only one person that actually asked for a performance appraisal. And the others, they get stressed out because of it. So actually they don't take initiatives towards me. They never ask us, I mean the management, for a performance appraisal. So, in this school initiatives are taken…. So fully disagree. (Respondent A, principal, school 1)

3.6.4 Overlooking entities

Problems arose across different referent levels as well. Even within one level, there may arise some problems in appropriate interpretation based on respondents' structural or physical position within a school. A proper conceptualisation of what is understood by the word *school* was not unambiguous. For some respondents whose school consisted of two campuses that operated independently as separate entities, the concept of his/her school was not that self-evident. As a middle manager working on both campuses, he/she was answering the items for both campuses together. The teachers that were interviewed, by contrast, worked on only one campus, and consequently made statements about their own campus.

> They are two different schools, but in reality it is actually one school. I see it as one school. When filling in the questions I was thinking of both schools. Yes, yes

> of course. I don't like to talk about "we and them", because that is very general. (Respondent E, middle manager, school 3)

> Especially in our case. I mean, it is about … yes, we do know less about … yes, we are in this school, aren't we. So we do have meetings together and we hear stories and ideas from the other school. But we have short and rather informal meetings in this school as well. But I think that this is better to answer this about one's own school, because I could absolutely not say about the others… (Respondent D, teacher, school 3)

4 Discussion

The findings of this study demonstrate that several problems arise during the answering process of respondents to items in an SSE questionnaire. First, when interpreting items, there are predominantly linguistic issues that generate cognitively invalid results. Second, during the elaboration stage, there is a problem for respondents to stay on topic and refer to the appropriate time frame. Furthermore, it happens they are mistaken about whom a statement is required. Third, when selecting an answering option, respondents do not always use the predefined options as intended or lack an option that reflects their mental judgement. Finally, the position the respondents hold in the school influences some aspects of their answering process.

Based on this study, it could be argued that sound conclusions of SSE results are not self-evident. Like the study by Koskey et al. (2010), the current study demonstrates how difficult it is for respondents to reflect a complex reality in SSE survey items (Shum and Rips 1999). An important lesson we take from this study is that instrument developers should be careful in making assumptions about the underlying thought processes of respondents. Several problems arise in each stage of respondents' cognitive process which threaten the validity of SSE results. It could be discussed that some problems are more severe than others and could, consequently, complicate a proper interpretation of SSE survey results in a differential way. This article did not, however, attempt to generate a ranking of problems in terms of severity in threatening cognitive validity.

The insights gained in the course of this research are cruxes for SSE instrument developers in making improvements when developing and revising SSE instruments. Despite a pilot test and a critical review of a panel of experts to ensure the tested SSE instruments' quality, this study illustrates the importance of making instrument developers' intentions explicit for every item (Willis 2005). The cognitive stages are a suitable framework to that end. Furthermore, instrument developers should conduct a thorough cognitive pretest for which such an explicit framework is an important and advantageous tool. Changes in SSE survey design based on cognitive interview findings lead indeed to a more accurate understanding of items and yield more valid conclusions (Desimone and Le Floch 2004; Ryan et al. 2012; Madans et al. 2011).

Based on the findings of this exploratory study, a next step is to test interventions or improvements to SSE instruments (Cohen et al. 2011). This could for instance involve formulating and testing well-defined hypotheses about the formulation of items where definitions of terms are provided when respondents scroll over them (Peytchev et al. 2010). This will broaden and deepen our understanding what impact interventions in the design of SSE instruments have on the cognitive processes executed by respondents

while filling in an SSE instrument (Willis 2005; Madans et al. 2011; Jobe 2003). Next, this would also enable identifying the effect of these interventions on obtained results in terms of data quality such as the number of substantive responses or the level of item nonresponse (e.g. Knäuper et al. 1997; Lenzner 2012; Lenzner et al. 2010).

Considering an argument-based approach to validity (Kane 2006), there are some important conclusions to be drawn. This approach advocates, on the one hand, a clear outline concerning the measurement procedures, which turn an attribute into a measurement instrument with a set of indicators (items). On the other hand, it suggests building an argument for the conclusions drawn from the scored instrument. Obviously, these two aspects should be congruent with each other in order to make valid conclusions with regard to the attribute. This study takes the first important step in studying this whole issue by examining the (mis)match between the intention of the instrument's items and how respondents cognitively process them. However, in the view of an argument-based approach, there should also be an examination of the congruence between the way respondents cognitively process SSE items and how SSE results are interpreted by SSE users. As such, a cognitive construal of items deviant from the instrument developers' intention does not necessarily create a validity problem, at least if the conclusions drawn from the data fit the way respondents construed the items (Kane 2013). However, since we detect in our study a whole range of problems occurring within the nexus between the instrument developers' intentions and the answering process of respondents, this could indicate that some problems may also occur when SSE users are interpreting the data. Although this particular issue has not been addressed yet, literature on data use within the framework of school feedback has already demonstrated a lack of know-how to accurately interpret information (Saunders 2000; Kerr et al. 2006; Williams and Coles 2007).

The findings of this study also demonstrate the need to consider ways of adequate use of survey instruments in the SSE process. Implementing surveys as a means to describe and evaluate the functioning of schools requires an accurate understanding of the instrument and its consequent results. In order to facilitate this, enhancing the (self-)evaluation capacity of participants and schools should be given priority at policy level. This demands efforts both at the individual participant level and at organisational and leadership level (Preskill and Boyle 2008). Investing more in the development of instruments that enable participants to probe for more information on what is being asked, or where a common understanding of items is created, is advisable. Instruments that are (more) responsive for participants' expertise or function, by means of an adjusted vocabulary for instance, might be desirable, so is equipping schools with more resources in terms of staff so that enough time can be spent on properly implementing an SSE. At least one member of the staff, preferably with an expertise in quality assurance, should be able to really create ownership with regard to the instrument used in the SSE process and who could support other members of the staff. Such an advisory role is crucial to guarantee substantive and valuable SSE results that make sense to the SSE users. Providing an adequate (external) guidance, that could take up the role of a critical friend, has proven to be helpful as well to generate a valid picture of a school's functioning (O'Brien et al. 2017). Especially in a context where much emphasis is put on the importance of evidence-informed decision-making in schools, these interventions seem to be vital to avoid SSE turning into senseless efforts by generating distorted representations of schools (OECD 2007; Schildkamp et al. 2013).

Despite our efforts to train participants to think aloud, the used think-aloud protocol delivered a limited amount of usable data on respondents' cognitive processes. Although a

hybrid model of cognitive interviewing is stimulated for good reasons (Ryan et al. 2012), it seems to be more appropriate to make use of a systematic probing technique when aiming to uncover problems in respondents' cognitive processing of SSE survey items.

This study addresses the issue of cognitive validity in the context of SSE surveys and it demonstrates the importance of this topic as it reveals several problems. More research on this topic should be undertaken and awareness should be raised as it delivers also crucial insights for other research domains where information on organisations' functioning is collected by means of surveys.

5 Conclusion

This article focusses on what problems can be identified that cause cognitively invalid answers on school self-evaluation items (RQ1). Cognitively invalid means that the cognitive tasks a respondent performs in order to answer an item (interpretation, elaboration and response) are not in line with how the instrument developer intended them. In addition, this study searches for how the position of the respondents influences their cognitive processes (RQ2).

Results show that respondents have problems with the interpretation of some specific terms. Giving words or concepts appropriate meaning is a problematic issue when answering SSE surveys. Next to semantic issues, respondents also struggle in some cases with the sentence structure of items. These findings connect with existing literature on survey questions, where linguistic issues have already been addressed some decades ago (Belson 1981; Fowler 1992). It may be concluded that SSE instrument developers should pay more attention to linguistic issues.

Also, with regard to respondents' elaborations (i.e. the search in their memory for relevant information on the specific topic), issues emerge in the data. Results show that respondents have difficulties staying on topic. It happens that their thoughts stray and they consider information which is not relevant, or, conversely, think that they are not covering the broader scope of the item. Possibly, we could link these phenomena to the way our memory operates. It is generally assumed that memory search includes progressively more specific cues, which could lead to a broader or a more narrow focus in respondents' thoughts (Tourangeau 2000; Karabenick et al. 2007). Regarding the contextual aspect of an elaboration, it is found that respondents can mistake the appropriate time frame and think not of a current state of affairs but bring up outdated information. Results also show respondents are mistaken about whom a statement is required. While an elaboration can be seen as a result of an invalid interpretation (e.g. Koskey et al. 2010), the current study demonstrates the importance of a clear focus throughout the elaboration stage, being crucial for cognitively valid SSE results.

The current study also demonstrates problems in the response stage. The intention of the predefined 'don't know' option, provided only for respondents who have no relevant information, is violated in different ways. For example, we found respondents selecting 'don't know' due to an item's complexity, which is supported by earlier research (Lenzner 2012; Krosnick 1991). Furthermore, respondents sometimes lack an answering option that reflects their mental judgement. Findings regarding the use of predefined answer options connect with earlier research within the field of survey methodology. Choosing between open and closed questions and what options to provide is a fundamental consideration that should be made by the instrument developer (Cohen et al. 2011; Schwarz and Hippler 2004).

With regard to the position of respondents in the school (RQ2), we conclude that their position does indeed influence how SSE items are cognitively processed, but not in every aspect. Some principals or middle managers also have difficulties with the interpretation of items, just like their fellow teachers. Others found some items to be self-evident, just because of the position they hold, which could be linked to different experiences and underlying mental models they have (Day et al. 2009; Johnson-Laird 1983). Furthermore, in some cases, they exclude themselves from their judgement and make a statement only on the school team they manage. Next to this finding, results show that the answers of a principal and middle manager encompass two campuses of their school as they operate in both campuses. Consequently, their responses were based on a consideration of both campuses, whereas teachers, who were only working in one of the two campuses, answered the items only for their own campus.

Appendix 1

Table 1 School self-evaluation items

INTEGRATED POLICY
In this school...
...everyone has a clear view of the job descriptions of other school staff.
...everyone has a vision that exceeds one's own job responsibilities.
...the management informs the team about administrative activities.
...everyone gives due consideration to the activities, ambitions and aspirations of other …school staff in what I do.
...everyone believes in the value of mutual coordination.
I...
...have a clear view of the job descriptions of other school staff.
...have a vision that exceeds my own job responsibilities.
...know about the administrative activities.
...give due consideration to the activities, ambitions and aspirations of other school staff in what I do.
...believe in the added value of mutual coordination.
REFLECTIVE CAPACITY
In this school...
...determining points for improvement is not seen as a threat.
...everyone has a reflective attitude towards their own actions.
...everyone has a positive attitude towards collective reflection
...everyone observes other people's performance.
...initiatives are taken to collect data on one's own functioning.
I...
...do not experience determining points for improvement as a threat.
...have a reflective attitude towards my own actions.
...have a positive attitude towards collective reflection.
...observe other people's performance.
...take initiatives to collect data on my own functioning.

References

Babik, D., Singh, R., Zhao, X., & Ford, E. W. (2015). What you think and what I think: studying intersubjectivity in knowledge artifacts evaluation. *Information Systems Frontiers*, 1–26. https://doi.org/10.1007/s10796-015-9586-x.

Bateson, N. (1984). *Data construction in social surveys* (Vol. Vol. 10, Contemporary social research). London: George Allen & Unwin.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. [Article]. *Public Opinion Quarterly, 71*(2), 287–311.

Belson, W. A. (1981). *The design and understanding of survey questions*. Hampshire: Gower Aldershot.

Blair, J., & Brick, P. (2009) Current practices in cognitive interviewing. In *64th Annual Conference of the American Association for Public Opinion Research (AAPOR), Hollywoord, Florida,* (pp. 5691–5700).

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Mutlilevel theory, research, and methods in organizations. Foundations, extensions, and new directions* (pp. 349–381). San Francisco: Jossey-Bass.

Bradburn, N. M. (2004). Understanding the question-answer process. *Survey Methodology, 30*(1), 5–15.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological Methodology, 12*, 389–437. https://doi.org/10.2307/270748.

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multi-level construct validation. In F. J. Yammarino, & F. Dansereau (Eds.), *Multi-level issues in organizational behavior and processes* (Vol. 3, pp. 273–303, Research in Multi Level Issues). The Netherlands: Elsevier Ltd.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (Seventh edition ed.). London: Routledge.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Research, 12*(3), 229–238. https://doi.org/10.1023/a:1023254226592.

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly, 73*(1), 32–55. https://doi.org/10.1093/poq/nfp013.

Day, C., Sammons, P., Hopkins, D., Harris, A., Leithwood, K., Gu, Q., Brown, E., et al. (2009). *The impact of school leadership on pupil outcomes. Final report.* London: The National College for School Leadership.

Day, C., Sammons, P., Leithwood, K., Hopkins, D., Harris, A., Gu, Q., et al. (2010). *Ten strong claims about successful school leadership*. Nottingham: The National College for School Leadership.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1–22. https://doi.org/10.3102/01623737026001001.

Edwards, B. D., Day, E. A., Arthur Jr., W., & Bell, S. T. (2006). Relationships among team ability composition, team mental models, and team performance. *Journal of Applied Psychology, 91*(3), 727.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data* (Revised Edition ed.). Cambridge: MIT-press.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly, 56*(2), 218–231. https://doi.org/10.1086/269312.

Fowler, F. J. (2014). *Survey research methods* (5ed., Applied social research methods series). Los Angeles Sage publications.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology.* Hoboken: John Wiley & Sons.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (Vol. 2, pp. 105–117). London: Sage.

Hendriks, M. (2000). *Kwaliteitszorg voortgezet onderwijs. Instrumenten en organisaties.* Utrecht: VVO/Q5, project kwaliteitszorg voortgezet onderwijs.

Hofman, R. H., Dukstra, N. J., & Hofman, W. H. A. (2005). School self-evaluation instruments: an assessment framework. *International Journal of Leadership in Education, 8*(3), 253–272. https://doi.org/10.1080/13603120500088802.

Jobe, J. B. (2003). Cognitive psychology and self-reports: models and methods. *Quality of Life Research, 12*(3), 219–227. https://doi.org/10.1023/a:1023279029852.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness* (Vol. 6). Cambridge: Harvard University Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4, pp. 17–64). Westport: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., et al. (2007). Cognitive processing of self-report items in educational research: do they think what we mean? *Educational Psychologist, 42*(3), 139–151. https://doi.org/10.1080/00461520701416231.

Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*(4), 496–520. https://doi.org/10.1086/505057.

Knäuper, B., Belli, R. F., Hill, D. H., & Herzog, A. R. (1997). Question difficulty and respondents' cognitive ability: the effect on data quality. *Journal of Official Statistics, 13*(2), 181–199.

Koskey, K. L. K., Karabenick, S. A., Woolley, M. E., Bonney, C. R., & Dever, B. V. (2010). Cognitive validity of students' self-reports of classroom mastery goal structure: what students are thinking and why it matters. *Contemporary Educational Psychology, 35*(4), 254–263. https://doi.org/10.1016/j.cedpsych.2010.05.004.

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7*(3), 77–124. https://doi.org/10.1111/j.1529-1006.2006.00030.x.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations. Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations. Foundations, extensions, and new directions*. San Francisco: Jossey-Bass.

Krippendorff, K. (2012). *Content analysis: an introduction to its methodology*. Los Angeles: Sage.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (Vol. 2, pp. 263–314). Bingley: Emerald Group Publishing.

Kyriakides, L., & Campbell, R. J. (2004). School self-evaluation and school improvement: a critique of values and procedures. *Studies in Educational Evaluation, 30*(1), 23–36. https://doi.org/10.1016/S0191-491X(04)90002-8.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310.

Lenzner, T. (2012). Effects of survey question comprehensibility on response quality. *Field Methods, 24*(4), 409–428. https://doi.org/10.1177/1525822x12448166.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: a psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020. https://doi.org/10.1002/acp.1602.

MacBeath, J. (2005). *Schools must speak for themselves: the case for school self-evaluation* (2nd ed.). London: Routledge.

MacBeath, J., & McGlynn, A. (2002). *Self-evaluation: what's in it for schools?* London: Routledge.

MacBeath, J., Schratz, M., Meuret, D., & Jakobsen, L. (2000). *Self-evaluation in European schools: a story of change*. London: RoutledgeFalmer.

Madans, J., Miller, K., Maitland, A., & Willis, G. (2011). *Question evaluation methods: contributing to the science of data quality* (Vol. 567). Hoboken: John Wiley & Sons.

McNamara, G., & O'Hara, J. (2005). Internal review and self-evaluation—the chosen route to school improvement in Ireland? *Studies in Educational Evaluation, 31*(4), 267–282. https://doi.org/10.1016/j.stueduc.2005.11.003.

McNamara, G., O'Hara, J., Lisi, P. L., & Davidsdottir, S. (2011). Operationalising self-evaluation in schools: experiences from Ireland and Iceland. *Irish Educational Studies, 30*(1), 63–82. https://doi.org/10.1080/03323315.2011.535977.

Mohammed, S., & Ringseis, E. (2001). Cognitive diversity and consensus in group decision making: the role of inputs, processes, and outcomes. *Organizational Behavior and Human Decision Processes, 85*(2), 310–335. https://doi.org/10.1006/obhd.2000.2943.

O'Brien, S., McNamara, G., O'Hara, J., & Brown, M. (2017). External specialist support for school self-evaluation: testing a model of support in Irish post-primary schools. *Evaluation, 23*(1), 61–79. https://doi.org/10.1177/1356389016684248.

OECD. (2007). *Evidence in education: linking research and policy (knowledge management)*. Paris: OECD.

OECD. (2014). *TALIS 2013 results: an international perspective on teaching and learning*. Paris: TALIS, OECD Publishing.

O'Muircheartaigh, C. (1999). CASM: successes, failures, and potential. In M. G. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 39–63). New York: Wiley & Sons, Inc..

Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks: SAGE Publications, inc..

Peytchev, A., Conrad, F. G., Couper, M. P., & Tourangeau, R. (2010). Increasing respondents' use of definitions in web surveys. *Journal of Official Statistics, 26*(4), 633–650.

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation, 29*(4), 443–459. https://doi.org/10.1177/1098214008324182.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., et al. (2004). Methods for testing and evaluating survey questions. *The Public Opinion Quarterly, 68*(1), 109–130. https://doi.org/10.2307/3521540.

Royston, P. (1989). Using intensive interviews to evaluate questions. In F. J. Fowler Jr. (Ed.), *Health survey research methods*. Washington, DC: U.S. Government Printing Office.

Ryan, K. E., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small- and medium-scale evaluations. *American Journal of Evaluation, 33*(3), 414–430. https://doi.org/10.1177/1098214012441499.

Saunders, L. (2000). Understanding schools' use of 'value added' data: the psychology and sociology of numbers. *Research Papers in Education, 15*(3), 241–258. https://doi.org/10.1080/02671520050128740.

Scheerens, J. (2000). *Improving school effectiveness* (Fundamentals of educational planning, Vol. 68). Paris: UNESCO International Institute for Educational Planning.

Scheerens, J., & Bosker, R. J. (1997). *The foundation of educational effectiveness*. Oxford: Pergamon Press.

Schildkamp, K., Lai, M. K., & Earl, L. M. (Eds.). (2013). *Data-based decision making in education. Challenges and opportunities (Studies in educational leadership)*. Dordrecht: Springer.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*(2), 277–287. https://doi.org/10.1002/acp.1340.

Schwarz, N., & Hippler, H.-J. (2004). Response alternatives: the impact of their choice and presentation order. In P. Biemer, R. M. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 41–56). Hoboken: John Wiley & Sons, Inc.

Shum, M. S., & Rips, L. J. (1999). The respondent's confession: autobiographical memory in the context of surveys. In M. G. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research*. New York: John Wiley & Sons Inc..

Tourangeau, R. (2000). Remembering what happened: memory errors and survey reports. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report. Implications for research and practice* (pp. 29–48). Mahwah: Lawrence Erlbaum Associates.

Tourangeau, R., & Bradburn, N. M. (2010). The psychology of survey response. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (Second ed.). Bingley: Emerald Group Publishing Limited.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods, 12*(2), 368–392. https://doi.org/10.1177/1094428107309322.

Vanhoof, J. (2007). *Zelfevaluatie binnenstebuiten. Onderzoek naar zelfevaluaties in scholen*. Mechelen: Plantyn.

Vanhoof, J., & Van Petegem, P. (2010). Evaluating the quality of self-evaluations: the (mis)match between internal and external meta-evaluation. *Studies in Educational Evaluation, 36*(1–2), 20–26. https://doi.org/10.1016/j.stueduc.2010.10.001.

Vanhoof, J., Deneire, A., & Van Petegem, P. (2011). *Waar zit beleidsvoerend vermogen in (ver)scholen? Aanknopingspunten voor zelfevaluatie en ontwikkeling*. Mechelen: Plantyn.

Williams, D., & Coles, L. (2007). Teachers' approaches to finding and using research evidence: an information literacy perspective. *Educational Research, 49*(2), 185–206. https://doi.org/10.1080/00131880701369719.

Willis, G. B. (2005). *Cognitive interviewing. A tool for improving questionnaire design*. London: SAGE Publications.

Woolley, M. E., Bowen, G. L., & Bowen, N. K. (2006). The development and evaluation of procedures to assess child self-report item validity. *Educational and Psychological Measurement, 66*(4), 687–700. https://doi.org/10.1177/0013164405282467.