SCHOOL EFFECTIVENESS and SCHOOL IMPROVEMENT

An International Journal of Research, Policy and Practice

School Effectiveness and School Improvement

An International Journal of Research, Policy and Practice

ISSN: 0924-3453 (Print) 1744-5124 (Online) Journal homepage: http://www.tandfonline.com/loi/nses20

School self-evaluation instruments and cognitive validity. Do items capture what they intend to?

Jerich Faddar, Jan Vanhoof & Sven De Maeyer

To cite this article: Jerich Faddar, Jan Vanhoof & Sven De Maeyer (2017) School self-evaluation instruments and cognitive validity. Do items capture what they intend to?, School Effectiveness and School Improvement, 28:4, 608-628, DOI: 10.1080/09243453.2017.1360363

To link to this article: http://dx.doi.org/10.1080/09243453.2017.1360363

Published online: 03 Aug 2017.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT, 2017
VOL. 28, NO. 4, 608–628
https://doi.org/10.1080/09243453.2017.1360363

ARTICLE

School self-evaluation instruments and cognitive validity. Do items capture what they intend to?

Jerich Faddar, Jan Vanhoof and Sven De Maeyer

Department of Training and Education Sciences, Faculty of Social Sciences, University of Antwerp, Antwerp, Belgium

ABSTRACT

School self-evaluation (SSE) often makes use of questionnaires in order to sketch a picture of the school. How respondents cognitively process questionnaire items determines the validity of SSE results. Still, one readily assumes that respondents interpret and answer items as intended by the instrument developer (referred to as cognitive validity), but it remains unclear whether they do. This study tested an exemplary SSE instrument by focusing on the extent to which SSE results are cognitively valid, and on the extent to which differences in cognitive validity can be attributed to respondents and/or items. Cognitive interviews with 20 participants made respondents' answering processes manifest. Results show that, overall, fewer than 50% of respondents' processes of interpreting and elaborating on items are cognitively valid. Cross-classified multilevel analyses indicate that various hierarchical levels, respondents and items, are significant in explaining differences in cognitive validity, but not for all stages of the answering process.

ARTICLE HISTORY
Received 27 December 2015
Accepted 21 July 2017

KEYWORDS
School self-evaluation; cognitive validity; questionnaires; item design

Problem statement

In efforts to enhance the quality of education, many educational systems expect schools to monitor and improve the quality of what they deliver themselves. School self-evaluation (SSE) is a mechanism that schools use to meet this expectation (MacBeath, 1999; McNamara, O'Hara, Lisi, & Davidsdottir, 2011). SSE can be defined as a process by which highly eligible participants describe and judge the functioning of the school in a systematic way, in order to inform school policies and suggest actions that should be undertaken (Vanhoof & Van Petegem, 2010). Often, school processes such as effective communication or distributed leadership are within the scope of SSE, as such processes are found to have considerable impact on a school's outcomes (Scheerens, 2008). Indeed, when school policies and actions draw on the information obtained from SSE's, it is of utmost importance that the description provided by participants in an SSE is accurate.

Often, the description of a school's functioning is mapped out by administering surveys with school staff, leading to a picture of the school as an organisation

CONTACT Jerich Faddar jerich.faddar@uantwerpen.be

(MacBeath, Schratz, Meuret, & Jakobsen, 2000; Meuret & Morlaix, 2003; Schildkamp, Visscher, & Luyten, 2009). When respondents are asked to make statements about organisational characteristics, two different questionnaire designs can be identified: a consensus design and a referent-shift design (Bliese, 2000; Chan, 1998; Chen, Mathieu, & Bliese, 2004). In a consensus design, respondents are asked to make statements about themselves, yet with a focus on collective properties (e.g., "I cooperate on a daily basis with my colleagues from different grades") (e.g., Hendriks & Bosker, 2003). Afterwards, the results are aggregated onto the organisational level (Bliese, 2000; Gisev, Bell, & Chen, 2013; Mathieu & Chen, 2011). Another frequently applied design is referred to as the referent-shift design, in which respondents report on characteristics of a higher level unit (i.e., the school) (e.g., "In this school, we cooperate on a daily basis with colleagues from different grades") (e.g., Hendriks & Bosker, 2003; Maslowski, 2001; Vanhoof, Deneire, & Van Petegem, 2011; Van Petegem, Cautreels, & Deneire, 2003). The latter design requires multilevel thinking of respondents, since they need to think of their school as a whole instead of exclusively of themselves as individuals.

However, literature on survey methodology has already pointed out that problems such as errors and distortions may influence survey results (Alwin, 1991, 2010; Groves et al., 2009). Moreover, in the context of SSE it is argued that applied measurement instruments are lacking a methodological and psychometric underpinning (Hendriks, 2000). In obtaining quality data for reliable and valid interpretations of SSE results, there is a vital role for respondents and the way in which they cognitively process items (Bateson, 1984). It is known that answering survey items demands several cognitive tasks that require a high level of cognitive effort from respondents (Krosnick, 1991). Respondents are, for example, expected to be able to read and interpret the survey items, and to have access to relevant information on the subject under review (Bateson, 1984; O'Muircheartaigh, 1999). Moreover, one assumes that respondents cognitively process items similarly to the intention with which the items are administered. Cognitive processing of items means that respondents are interpreting items, elaborating on them by retrieving relevant information from their memory, and answering them congruently with the instrument developers' intentions (Karabenick et al., 2007). The extent to which respondents process items as intended determines the degree to which results of surveys are cognitively valid (Karabenick et al., 2007; Koskey, Karabenick, Woolley, Bonney, & Dever, 2010).

How respondents interpret, elaborate, and respond to items determines what information they provide with regard to the school's functioning to those who analyse and use the SSE results (Karabenick et al., 2007; Tourangeau & Bradburn, 2010). When respondents interpret items incorrectly and do not think about the information that is asked for, the SSE results do not reflect what they are supposed to. This may lead to distortions in the results and, consequently, generate problems when drawing sound conclusions based on the results (Kane, 2013). The cognitive process that respondents are going through should be in line with what the measurement instrument intends to measure. In this respect, cognitive validity contributes to the concept of content validity (Lissitz & Samuelsen, 2007). High-quality data and being able to draw valid conclusions based on SSE results is of key importance to schools. This is especially the case in a context in which schools are expected to safeguard their own quality, and a bigger emphasis on schools' capacity in terms of evidence-informed decision-making is in place (Schildkamp, Lai, & Earl, 2013).

Although the cognitive validity of SSE survey results is an important aspect of overall validity, it seems as if there is a collective glossing over with regard to this issue. Nevertheless, in the context of SSE surveys it can be argued that two major concerns threaten the cognitive validity. First, respondents must meet the requirement of multi-level thinking, as they are asked to make statements on the level of themselves and/or on the level of the school as a whole (Chen et al., 2004; Kozlowski & Klein, 2000). If respondents fail to grasp the proper level on which they are asked to make a statement, an invalid interpretation of their answer is most likely. Second, the educational context can bring abstract and complex terms into the survey, which respondents or users are supposed to be able to read and/or interpret without any difficulty and in line with the intended meaning of the instrument (Koskey et al., 2010). Given these two concerns in the particular context of SSE, more research is highly needed. This study aims to examine to what extent SSE embedded questionnaire items are processed in a cognitively valid way. In order to be able to do so, a deeper insight into the respondents' cognitive processes is provided in the following paragraphs.

Cognitive validity framework

Literature determines three critical steps (see Figure 1) in assessing the cognitive validity of survey responses (Karabenick et al., 2007). First, it should be checked whether respondents comprehend the item and interpret it as intended. Next, an inquiry should be made of respondents' coherent elaboration on the item interpretation. In other words, what information do they retrieve from their memory and are they relying on while making a judgment? Finally, it should be examined whether the selection of a response option is congruent with the intended use and the item interpretation and elaboration. Although these steps are presented sequentially, respondents can shift from

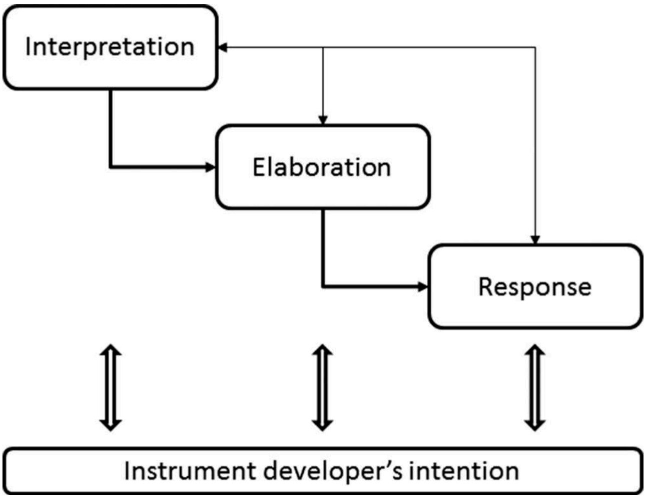


Figure 1. Framework of cognitive validity.

one stage to another (Collins, 2003; Ryan, Gannon-Slater, & Culbertson, 2012). The next paragraphs will elaborate more on the details of each of these critical steps.

Item interpretation comprises several tasks in itself. Respondents should not only be able to read and understand the words of the item, but are also required to come to an interpretation of the item analogous to the intention of the instrument developer (Belson, 1981; Fowler, 1992; Schwarz, 2007; Tourangeau, Rips, & Rasinski, 2000). Several issues on a semantic, syntactic, and pragmatic level may emerge and be problematic for an accurate item interpretation (Lenzner, Kaczmirek, & Lenzner, 2010; Tourangeau & Bradburn, 2010; Tourangeau et al., 2000). Semantic problems are caused by unclear terms, technical concepts, or words that respondents may not know, while syntactic problems may arise as respondents encounter items of high grammatical complexity, or syntactic ambiguity (i.e., items map onto multiple underlying representations). Furthermore, pragmatic problems refer to issues as failing to deduce the intent of an item through stylistic elements or other items near to the items of interest. Within the context of SSE, it can be argued that educational concepts are often abstract and complex, which may have consequences for the understanding and interpreting of survey items (Koskey et al., 2010).

Based on the interpretation of an item, the next critical step of elaboration is for respondents to retrieve information from their memory that is relevant for answering the item. This information from the memory can comprise experiences, thoughts, feelings, or perceptions that are cognitively processed at the moment of survey administration (Karabenick et al., 2007). A first important aspect of a coherent elaboration is the *content* of the item. Respondents need to think about and report on instances of behaviour and/or attitudes related to that particular item. For example, when respondents are asked to make a statement on the item “In this school, one has a clear view on the job description of others in the school”, they should provide examples that show they know what responsibilities or tasks others in the school have, and who they should consult regarding particular questions. Another element in a coherent elaboration is *context*. During their elaboration, respondents should refer to an appropriate level on which they are making statements (such as individuals, management, or groups) and within an appropriate reference period (Koskey et al., 2010). In the aforementioned example, this means that respondents elaborate by giving instances referring to all staff of the school (including administrative or supporting staff), and which are still relevant at the moment of item administration. It is especially necessary that they refer to the appropriate level in the case of an SSE survey, which can be characterised by multilevel thinking, where respondents are asked to make statements about an organisation or individual characteristics.

A final task for respondents is to formulate a response (e.g., Schwarz, 2007; Tourangeau et al., 2000). This is a critical task for respondents, as they have to hold the item interpretation, retrieved information, and the possible answer options in their working memory (Karabenick et al., 2007). Moreover, the provided response options can be designed in different ways, each asking a different level of cognitive effort of respondents (Galesic, Tourangeau, Couper, & Conrad, 2008; Krosnick, 1991). It is necessary that a respondent’s answer reflects its preceding item interpretation and elaboration (Karabenick et al., 2007). In order to make a valid interpretation of the observed score, a respondent needs to select a pre-defined answer option as intended by the instrument developer. For example, if a response option *I don’t know* is provided with the aim of capturing cases in which a respondent has no relevant knowledge about an item, it should not be used when a respondent does not know what the item is asking for.

This study

The previously described insights, bridging the fields of survey methodology and cognitive psychology, provide a valuable perspective to study the answering process of respondents. Up till now, little research has focused on aspects of the answering process of respondents on SSE questionnaires in general and the cognitive validity of the results in particular. It is unknown whether respondents process SSE questionnaire items as intended by the instrument developer or not. Therefore, this study adopts this framework within the specific context of SSE and focuses on the following research questions:

(RQ1) To what extent are results of school self-evaluation surveys cognitively valid?

(a) To what extent are respondents interpreting the items as intended;
(b) to what extent are respondents coherently retrieving information from their memory; and
(c) to what extent are congruent answer options chosen?

Literature points to a crucial role of respondents when cognitively processing questionnaire items. However, cognitive processes are embedded in one individual. Respondents may differ, for example, in cognitive ability or motivation to fill in survey questions, which can influence the results (Krosnick, 1991). Furthermore, it is argued that differences between items, such as differing answer options or negative formulations, can influence the cognitive processes of respondents (e.g., Fowler, 1992). Within the context of SSE, it is still unclear whether respondents and items actually do have an influence on the cognitive validity. Consequently, this study also aims to examine the following research question:

(RQ2) To what extent can differences in cognitive validity ratings be attributed to the level of respondents and/or items?

In the domain of SSE questionnaires, two item designs are frequently used: a consensus and a referent-shift design. Although it has been argued that the latter design would be more complex to cognitively process, evidence is lacking. Therefore, the third research question focuses on the impact of item design on cognitive validity.

(RQ3) To what extent does item design (a referent-shift vs. a consensus design) have an effect on the cognitive validity of its results?

Methods

Approach and technique

To explore the extent to which results of SSE surveys meet the precondition of cognitive validity, a qualitative approach for data collection was chosen. It enables us to gather in-depth and ample information on respondents' cognitive processes (Cohen, Manion, &

Morrison, 2011; Collins, 2003; Ericsson & Simon, 1993). To map out the cognitive process of respondents while filling in an SSE survey, a commonly used technique is cognitive interviewing (Beatty & Willis, 2007; Collins, 2003; Conrad, Blair, & Tracy, 1999; Ericsson & Simon, 1993; Willis, 2005).

A hybrid model of cognitive interviewing was applied, which consists of a combination of a think-aloud protocol and the probing-technique (Karabenick et al., 2007; Ryan et al., 2012). The think-aloud technique is valuable because of its open-ended format. Little interviewer bias is imposed, and the respondent is free to provide answers that may not have surfaced in other formats. The disadvantages are that respondents might stray from the task, or have difficulties expressing their thoughts (Royston, 1989). In this study, however, respondents were given a brief introductory training in thinking aloud, as advised by Ericsson and Simon (1993). Anticipating the disadvantages of the think-aloud protocol, the verbal probing technique (i.e., short and direct questions) was also applied. This technique enables probing for relevant information about the cognitive processes respondents go through when filling in items (Karabenick et al., 2007; Willis, 2005), thus implying a smaller burden on the respondent and larger control for the interviewer as the interview is more guided. The verbal probes posed in the interview consisted of general probes and specific probes. General probes were systematically used for every question, while specific probes were formulated for one particular question (DeMaio & Landreth, 2004).

In order to obtain rich information on respondents' cognitive processes, a hybrid model was adopted in the study. This implied that respondents could not be asked to process the items twice. Processing an item while thinking aloud would prime and influence the respondents' own thinking when they were asked to process the item using the systematic probing technique afterwards. In addition, the cognitive interviewing literature shows that cognitive interviewing places a large cognitive burden on respondents (Willis, 2005). In particular, the think-aloud protocol is considered burdensome. In order to make optimal use of the hybrid model of cognitive interviewing, and to reduce the cognitive workload of the respondents, we randomly allocated the respondents to two groups. Group 1 had to fill in only half of the items while thinking aloud; the other items were administered by means of the systematic probing technique. Group 2 also started with the think-aloud protocol, but did so with the items from Group 1's systematic probing technique. Both groups began with the think-aloud protocol so that their spontaneous thinking was not distorted by the probing questions that were provided during the verbal probing technique.

Instrument

In the current study, it was important to test an existing instrument that met well-established criteria. First, the instrument had to tap into processes at the organisational level and needed to be well embedded in the local context. Next, the instrument developers had to be accessible to elicit what they intended with the items. Furthermore, the items needed to have the capacity to be formulated both in a consensus and referent-shift design. Considering all these criteria, this led to the selection of an instrument that taps into the construct of the policy-making capacity of schools. Within this instrument, two exemplary scales measuring "integrated policy" and

Table 1. Item design examples.

	Item design	Example item
Ex. 1	Consensus design	*I have a clear view on the job description of others in the school.*
Ex. 2	Referent-shift design	*In this school, one has a clear view on the job description of others in the school.*

"reflective capacity" were selected randomly to serve as testing subjects in this study (Vanhoof et al., 2011). These scales, next to six others such as innovative capacity or effective communication, are widely used in the local context. In terms of design and linguistic complexity, the two selected scales are similar to other scales that are part of the instrument.

To answer RQ3, two relevant variations of all items were adopted: a referent-shift design variation with the stem "In this school …", and a consensus design variation with the stem "I …". Table 1 shows an example of each design. For each item, the same response options were provided: a 4-point scale ranging from *totally disagree* to *totally agree*, accompanied by a *don't know* option.

Participants

Four primary schools participated in the study and were selected on the basis of a random sample, controlling for school size in terms of number of teachers. In each school, the principal, a middle management officer, and a selection of three teachers performed a cognitive interview. When a school did not have a middle management position, an extra teacher was sampled. In total, 20 participants from primary education cooperated in our study.

Outcome measure and analysis

In determining to what extent results from SSE surveys are cognitively valid, it was crucial to identify how the instrument developers intended the items. With this goal in mind, the instrument developers were asked to extensively describe their intentions with regard to each of the items. This was done by means of a written form that addresses the interpretation, elaboration, and response stage for each item. These descriptions served as criteria for cognitive validity as suggested by Woolley, Bowen, and Bowen (2006). Regarding the elaboration stage, special attention was paid to the content of the item and the appropriate context to which the item is referring. An example of the cognitive validity criteria is attached in Appendix 1.

The outcome measure in this study, cognitive validity, is generated by making a comparison between how instrument developers intended the items, and how respondents actually cognitively processed them. The researchers coded the respondents' verbalised cognitive processing, based on the cognitive validity criteria. The same coding scheme was used throughout the analysis for all three cognitive stages (see Appendix 1). When the answer of a respondent did not correspond to the intention of the instrument developer, a rating of "0" was granted to the answer. If a respondent mentioned elements that both did and did not match the intention of the developer, a rating of "1" was assigned. A rating of "2" was allocated to those answers that were fully congruent with the developer's intention. All interviewees' verbalisations were coded using the NVivo 10 software. In order to guarantee the reliability of the cognitive validity ratings, a second researcher independently rated 13.5% of all observations based on the cognitive validity criteria, which resulted in a Cohen's Kappa of 0.62. Consequently, cognitive validity codings were considered to be sufficiently reliable for further analysis (Fleiss, 1981; Landis & Koch, 1977).

The data consist of 400 observations; 20 respondents verbalised their cognitive process by answering an SSE instrument consisting of 20 items. Each of the observations was coded for the three critical steps in determining cognitive validity: item interpretation, elaboration, and response. This means that there are 1,200 coding units, of which one half was collected by means of a thinking-aloud protocol, the other half by means of the systematic probing technique. A closer look at the data revealed that, during the interpretation stage, 45.50% of the observations across items and respondents were not appropriate for coding because they did not contain enough information for coders to make a judgment on their cognitive validity (i.e., "insufficient prompt"). During the elaboration stage, 22.50% of the observations appeared to be insufficient to allocate cognitive validity ratings. With regard to the response stage, 2.75% of all observations were insufficient for cognitive validity coding purposes. A cross-table analysis (see Table 2) shows that most insufficient prompt data were administered by means of the think-aloud protocol. As the data with the code "insufficient prompt" do not give us any insights into respondents' cognitive processes, these were excluded from the analyses. All other observations were taken into account by the following analyses.

In order to tackle RQ2, which aims to examine the extent to which differences in cognitive validity ratings can be attributed to the level of respondents and/or items, an explanatory analysis should take into account that the data represent a multilevel design: Cognitive validity ratings are nested in respondents on the one hand and in items on the other hand. Moreover, the multilevel design is cross-classified since cognitive validity ratings are nested within every respondent and every item. A statistical model to analyse the data should be able to model this complex multilevel – or, in other words, mixed-effects – design. Besides the cross-classified multilevel design, a second issue has to be kept in mind. As described above, the dependent variable *cognitive validity* consists of three possible ratings or values: "0", "1", and "2", and is ordinal in nature. Bearing these two aspects in mind, a cumulative link mixed model, as provided by the R-package Ordinal (Christensen, 2015), is appropriate to model our data.

In the explanatory analyses, an estimation is made of the cumulative probability that the *i*th observation falls in the *j*th cognitive validity rating category. This is formally written in Equation (1), where *i* indexes all observations and *j* the three cognitive validity

Table 2. Cross-table insufficient prompt by technique.

Cognitive validity rating	Sufficient prompt (%)	Insufficient prompt (%)	*Total (%)*
Systematic probing (n = 600)	97.50	2.50	*100.00*
Think-aloud (n = 600)	55.34	44.66	*100.00*

rating possibilities minus 1, whereas the model calculates thresholds between the different rating possibilities. The hierarchical levels in which the cognitive validity ratings are nested, respondents and items, are modelled as random effects and assumed to be normally distributed. This null model will be further reported as Model 0.

$$Logit\,(P(Y_i \leq j)) = \theta_0 - \mu_1(respondent_i) - \mu_2(item_i) \tag{1}$$

In order to know whether the included random effects in the null model are statistically significant, the null model is compared with two new models, wherein each time one random effect has been left out. These models are referred to as Model 0a (respondents out) and Model 0b (items out). In order to compare the different models, we relied on the Akaike information criterion (AIC) and the likelihood ratio test.

In order to facilitate the interpretation of the found statistically significant variance parameters, probabilities are predicted for respondents who tend to answer more cognitively validly (a 90th percentile respondent), in comparison with a respondent who tends not to (a 10th percentile respondent). Analogue predictions are made for items that, to a higher and lower extent, trigger cognitively valid results.

It is suggested that the way items are designed can have an effect on how items are cognitively processed by respondents and, consequently, on their cognitive validity. Items that are characterized with a referent-shift design may ask more cognitive effort of respondents in comparison with consensus design items, and may have a higher chance of being cognitively invalid. Extending the model by introducing this explanatory parameter (β_1) can be written as follows in Equation (2). In the results section, we referred to this model as Model 1.

$$Logit\,(P(Y_i \leq j)) = \theta_0 - \beta_1(item\ design_i) - \mu_1(respondent_i) - \mu_2(item_i) \tag{2}$$

In order to test whether Model 1 fits better compared to Model 0, the same fit statistics as with the different null models were used.

Results

Descriptive results

This part of the Results section focuses on the extent to which SSE results are cognitively valid (RQ1), by examining the different stages in the respondents' answering process (see Table 3).

Interpreting items as intended appears not to be self-evident for respondents. Only 44.49% of the observations are cognitively valid, meaning that respondents are interpreting the item as intended by the instrument developer. Conversely, almost 19% of

Table 3. Descriptive results on the cognitive validity of respondents' answering processes.

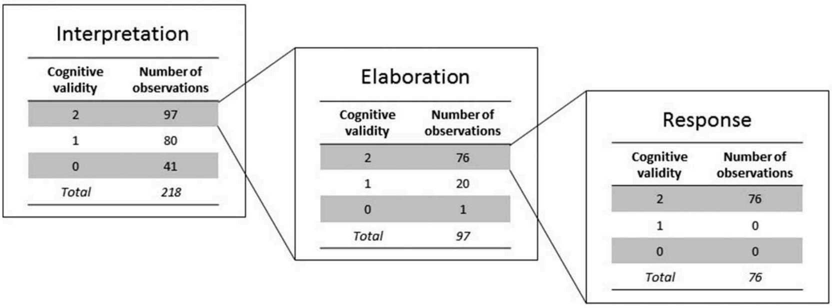
Cognitive validity rating	Interpretation % (n = 218)	Elaboration % (n = 310)	Response % (n = 389)
Cognitively invalid	18.81	14.52	4.63
Partially cognitively valid	36.70	53.87	2.31
Cognitively valid	44.49	31.61	93.06

the observations are cognitively invalid, meaning that none of the respondents' verbalisations on the item interpretations is what the instrument developer was aiming to capture with the item. Almost 37% of the observations were in between. This means that at least one element of the respondent's interpretation is in line with the instrument developers' intention, but that at least one element is not.

Concerning the elaboration stage, the results show a slightly different pattern. Almost 15% of 310 observations are allocated as "cognitively invalid". Remarkably, only 31.61% of the observations are completely cognitively valid. The category "partially cognitively valid" is a rather large one. About 54% of all observations in the elaboration stage are allocated this rating. This means that respondents are retrieving information from their memory that is only partially in line with the instrument developers' intention.

The response stage of the respondents' cognitive answering process seems to be less problematic. To a large extent, the use of the pre-defined answering options is cognitively valid. About 93% of all observations demonstrate a use of the pre-defined answer options which is congruent with the instrument developers' intention and with the respondents' preceding cognitive processes. Only a small minority, up to 4.63%, of all observations in the use of the response options turn out to be cognitively invalid.

Figure 2 shows to what extent observations are considered cognitively valid across all three different stages of item processing. Remarkably, the majority of the observations drop out of the cognitively valid category "2" during the first stage of item interpretation. Only 97 out of 218 observations in the interpretation stage are cognitively valid. When these observations are followed up during the elaboration stage, data show that 76 of these observations remain cognitively valid. Overall, it can be concluded that about one in five respondents is straying off from the initial interpretation of the item while elaborating. After the interpretation and elaboration stages, which are cognitively validly processed by respondents, no problems occur during the last stage of the answering process. Respondents use the pre-defined response options as they are postulated by the instrument developers. As a result, 76 observations out of 218 (34.86%) are cognitively validly processed across the three different cognitive stages.



Interpretation

Cognitive validity	Number of observations
2	97
1	80
0	41
Total	*218*

Elaboration

Cognitive validity	Number of observations
2	76
1	20
0	1
Total	*97*

Response

Cognitive validity	Number of observations
2	76
1	0
0	0
Total	*76*

Figure 2. Cognitive validity across response stages.

Note: Cognitive validity: "2" = cognitively valid; "1" = partially cognitively valid; "0" = cognitively invalid.

Explanatory results

The cognitive validity ratings are conceptually nested into two higher levels: *respondents* and *items*, which raises the question of whether differences in cognitive validity can be explained by these levels (RQ2).

Item interpretation

The model with two random effects, *respondents* and *items*, shows that cognitive validity ratings for item interpretation vary from respondent to respondent ($\sigma^2 = 1.004$) and from item to item ($\sigma^2 = 0.199$). When studying the model comparison statistics (see Table 4), it is found that for the interpretation stage, the statistical significance for one random effect in the model is supported according to the AIC and the loglikelihood ratio test. If the random effect *items* is left out of the model (Model 0b), the AIC increases and the difference in −2Loglikelihood (Δ-2LL) with one degree of freedom (Δ*df*) is statistically significant in comparison with Model 0. These results show that, for the interpretation stage, differences in items predict the extent to which item interpretations are cognitively valid. If *respondents* is left out of the model (Model 0a), no significant change occurs in the model fit.

As stated in RQ3, it is suggested that the design of items can influence respondents' cognitive processes. Therefore, we tested a model that added *item design* (consensus and a referent-shift design) as an explanatory variable (Model 1) and compared it with Model 0. The model fit statistics show that Model 1 suits the data significantly better than Model 0. Nevertheless, the random effect *items* remains significant in Model 1, so item design does not explain all the variance among items.

As the variance between items is significant, it is interesting to demonstrate what effect this variance has on the probability that an item is interpreted in a cognitively valid way. This can be expressed as an estimation of the probability that a random item will fall in one of the cognitive validity rating categories when processed by an average respondent. A simulation, based on Model 1, is made for an item on the 10th percentile (performing badly for cognitive validity) and an item on the 90th percentile (performing well for cognitive validity). Table 5 shows that the 90th percentile item is 28% more likely to receive a completely cognitively valid interpretation, in comparison with the 10th percentile item. Conversely, a 10th percentile item is about 14% more likely to be interpreted in a cognitively invalid way than the 90th percentile item.

Table 4. Model comparison interpretation.

	AIC	loglikelihood	Δ*df*	Δ-2LL	*p*
Model 0 (all random effects in)	445.71	−218.86			
Model 0a (respondents out)	445.39	−219.69	1	1.68	>0.05
Model 0b (items out)	460.46	−227.23	1	16.75	<0.001
Model 1 (item design in)	443.77	−216.89	1	3.94	<0.05

Table 5. Predicted probabilities for cognitive validity based on variance between items.

Interpretation	10th percentile item	90th percentile item
Cognitively invalid	22.82%	08.62%
Partially cognitively valid	46.00%	32.69%
Cognitively valid	31.18%	58.70%

Table 6. Thresholds and effect significance of item design on cognitive validity during interpretation stage.

	Est.	St.Err.	Est./St.Err.	*p*
Thresholds				
"Cognitively invalid" |"Partially cognitively valid"	−2.286	0.412	−5.548	<0.001
"Partially cognitively valid" |"Cognitively valid"	−0.277	0.366	−0.758	>0.05
Regression weights				
β_1 Item design (1 = Referent-shift)	−0.988	0.487	−2.029	<0.05

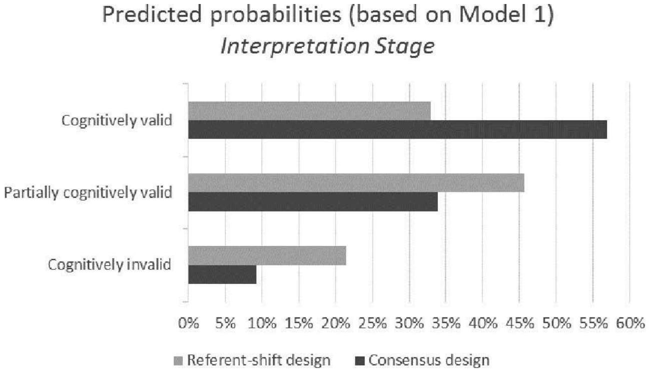


Figure 3. Predicted probabilities for cognitive validity of the interpretation stage by item design.

Model 1 includes *Item design* as a predictive variable, wherein the reference value represents a consensus design where respondents make a statement about themselves ("I …"), while value 1 stands for a referent-shift design, which asks respondents to make a statement about the school as a whole ("In this school …"). Table 6 reports on the estimates in Model 1. The effect of *item design* (–0.988) means that the thresholds are 0.988 logits lower for referent-shift design items in comparison with consensus design items. This implies that referent-shift design items have a smaller chance of being interpreted in a cognitively valid way. This difference between the two item designs is visualised in Figure 3.

During the interpretation stage, consensus design items have only a 9.22% chance of being interpreted cognitively invalidly. For referent-shift items, this chance amounts to 21.46%. Conversely, consensus design items clearly have a much higher probability of being processed in a cognitively valid way. For an average consensus design item, this probability is predicted at 56.88%. For a referent-shift item, this probability drops to 32.94%.

Elaboration

Regarding the elaboration stage of item processing, it can be concluded that a null model including both the random effects *respondents* and *items* fits the data significantly better than the two models each excluding one of both random effects. The AIC of Model 0 is the lowest, and the loglikelihood ratio test indicates that the alternative models fit the data significantly worse when one of the random effects is excluded from

Table 7. Model comparison elaboration.

	AIC	loglikelihood	Δ*df*	Δ-2LL	*p*
Model 0 (all random effects in)	590.14	−291.07			
Model 0a (respondents out)	608.42	−294.27	1	20.28	<0.001
Model 0b (items out)	594.55	−301.21	1	6.41	<0.05
Model 1 (item design in)	591.87	−290.93	1	0.28	>0.05

Table 8. Predicted probabilities for cognitive validity based on variance between items and respondents.

Elaboration	10th percentile item	90th percentile item	10th percentile respondent	90th percentile respondent
Cognitively invalid	19.02%	05.84%	25.10%	04.16%
Partially cognitively valid	63.37%	49.42%	61.87%	42.23%
Cognitively valid	17.61%	44.75%	13.03%	53.61%

the model (see Table 7). Cognitive validity ratings vary from respondent to respondent (σ^2 = 0.635) and from item to item (σ^2 = 0.270). In addition, the model fit statistics show that adding *item design* as an explanatory variable (Model 1) does not lead to a better fit of the model. The difference between consensus and referent-shift items appears to have no impact on the extent of cognitive validity in the elaboration stage (RQ3). Nevertheless, there are still differences between items, which are impeding the cognitive validity of the items' results, but not due to the specific distinction between consensus and referent-shift item designs.

The estimates of Model 0 enable probabilities to be predicted for cognitive validity of item elaborations, taking the variance between items and respondents into account. Table 8 shows that a 10th percentile item (performing badly for cognitive validity), processed by a random respondent, has about an 18% chance of obtaining a "cognitively valid" rating. On the other hand, for a 90th percentile item (performing well for cognitive validity), the predicted probability of a cognitively valid elaboration increases by up to 45%.

The variance between respondents can also be used for predicting probabilities. A 10th percentile respondent (performing badly for cognitive validity), elaborating on a random item, has only a 13.03% chance of doing so in a cognitively valid way. The chance that the respondent does so in a cognitively invalid way amounts to 25.10%. For a 90th percentile respondent (performing well for cognitive validity), this is only 4.16%. The probability for a cognitively valid elaboration on a random item is 53.61% for a 90th percentile respondent. In other words, a 90th percentile respondent has about 40% more chance of processing a random item cognitively validly, in comparison with a 10th percentile respondent.

Response

The next analysis focuses on the response stage, wherein respondents are supposed to select a pre-defined answer option that is congruent with their preceding cognitive tasks. Modelling the random effects *respondents* and *items* (see Table 9) leads to the conclusion that cognitive validity measures vary between respondents (σ^2 = 0.428) and

Table 9. Model comparison response.

	AIC	loglikelihood	Δ*df*	Δ-2LL	*p*
Model 0 (all random effects in)	236.00	−114.00			
Model 0a (respondents out)	234.97	−114.49	1	0.97	>0.05
Model 0b (items out)	235.85	−114.93	1	1.85	>0.05

items (σ^2 = 0.286). Nonetheless, when looking at the model fit comparison statistics, none of the random effects seems to be statistically significant in explaining differences in cognitive validity ratings with regard to the response stage. The use of pre-defined answer options is not impacted by differences between respondents or items. Since *items* is not a significant variance component, no model is estimated with the explanatory variable *item design* (RQ3).

Conclusion and discussion

Answering to SSE questionnaire items requires a great deal of cognitive effort on the part of respondents. This study examines the extent to which respondents perform consecutive cognitive tasks in line with the instrument developers' intention, and the extent to which variation in cognitive validity can be attributed to respondents and/or items.

Interpreting items as they are intended by the instrument developer appears to be a difficult and demanding task for respondents (RQ1a). The interpretation stage in our study is rather poorly executed; only 44.5% of the item interpretations are cognitively valid. Cognitive validity ratings drop further with regard to the elaboration stage. Only one in three item elaborations proves to be cognitively valid (RQ1b). The response stage is less problematic. Of all responses, 93% are cognitively valid, which means that respondents are using the pre-defined answer options as they were intended (RQ1c). Across the entire answering process, only 34.9% of the observations are considered cognitively valid over each of the cognitive stages. This result is in line with findings by Koskey et al. (2010) on students' answers on items concerning mastery goals in their classroom. However, in contrast to this study, we found the elaboration stage to be less cognitively valid. This difference in cognitive validity may be explained by the fact that the examined items in our study may be more abstract or difficult for respondents to retrieve relevant information on from their autobiographical memory (Tourangeau et al., 2000).

Explanatory cross-classified multilevel analyses show that the hierarchical level in which the cognitive validity ratings are nested does explain variance in these ratings, but not for every cognitive stage (RQ2). With regard to the interpretation stage, differences among items do matter in the extent to which results are considered as cognitively valid. This finding is supported by empirical studies, which indicate that several aspects of item formulation, such as the use of many meaningful words in a short space, have a significant influence on how questions are interpreted by respondents (e.g., Belson, 1981; Lenzner, 2012) and consequently impede the cognitive validity. During the elaboration stage, both *respondents* and *items* explain variance in the results' cognitive validity. This supports cognitive processing literature, as at this stage respondents are consulting their autobiographical memory to make statements (Karabenick et al., 2007; Tourangeau et al., 2000), which invites respondents to bring in personal frameworks (e.g., experiences, thoughts, and feelings). Hence, at this stage, differences between respondents can become more manifest than during other stages of the cognitive processes. It is possible that the level of respondents' motivation to engage in and their ability to perform the required cognitive tasks (Krosnick, 1991) can play a role in attaining a cognitively valid answering process. For the response stage, none of the variance components *items* and *respondents* matters significantly. A possible explanation is that the provided answer options were rather straightforward for respondents (Krosnick, Narayan, & Smith, 1996), since the survey consisted of a conventional 4-point Likert scale with a *don't know* option. As this scale is often used in surveys, respondents may already be highly familiar with such a scale. Furthermore, response options, as a part of a whole item, did not differ across the items. This uniformity may additionally explain why no significant differences are found in cognitive validity ratings.

A more elaborate explanatory analysis shows that a referent-shift item design has a negative effect on the probability that an item is interpreted in a cognitively valid way by respondents (RQ3). Although it is argued that referent-shift design items are more appropriate for capturing group-level characteristics (Bliese, 2000; Klein, Conn, Smith, & Sorra, 2001), we demonstrated that they have the drawback that respondents have a smaller chance of interpreting them in a cognitively valid way. Referent-shift items impose additional requirements for respondents at the level about which to make statements, referred to as multilevel thinking. Even when item design is taken into account, the variance parameter *items* remains significant. Given that the distinction between a consensus and referent-shift design is only one way in which items differ, future research may consider examining the effect of, for example, the use of abstract terms or concepts in explaining variation in cognitive validity among items (Lenzner et al., 2010).

In conclusion, the cognitive validity of results of the studied SSE survey is threatened during both the interpretation and the elaboration stage of the answering process. A lack of cognitive validity of SSE results can have consequences for the overall validity of the interpretations deduced from SSE results (Kane, 2006). Valid interpretations of SSE results are not possible when respondents are to a large extent thinking about information which the instrument developer was not aiming for. Moreover, SSE results cannot serve as a basis for appropriate decisions on school policy development when the validity of SSE results is doubtful (Kane, 2013). These findings with regard to cognitive validity are a valuable addition to the insights into the psychometric properties of SSE instruments. However, other techniques and measures to identify the quality of instruments are often reported, such as reliability indicators, information on content validity and face validity (e.g., Antoniou, Myburgh-Louw, & Gronn, 2016). Examining cognitive validity does not imply that other assumptions related to, for instance, content or face validity about data quality should not be checked. These findings are not only relevant in a school context; the way in which SSE questionnaires are built up is similar to how much perception-based survey research is conducted. It is often aimed at making organisational characteristics manifest by means of respondents reporting on themselves or on collective properties (of a team) (Hinkin, 1995), such as organisational commitment (e.g., McGee & Ford, 1987) or transformational leadership in organisations (e.g., Bass & Riggio, 2006). The framework of cognitive validity is expected to elicit similar problems within different contexts.

Findings from this study also have clear implications for SSE practices. When drawing conclusions on SSE results, it may be helpful to discuss the results with the participants. Such a (group) discussion might elicit how respondents construed the items, and may deliver more insight into how results came about and, possibly, foster consensus on how results should be interpreted. This suggestion points at the same time to consequences for the level of evaluation perspectives, school development versus accountability, with regard to SSE (Vanhoof & Van Petegem, 2010). Although a school development (low-stakes) perspective seems to lend itself more easily to providing room for discussion, in comparison with an accountability (high-stakes) perspective, it should be ascertained that, in the latter context, such a discussion should also be embedded in order to make sure that conclusions and uses of SSE results are made in a valid way.

Based on the findings of this study, instrument developers, in the context of SSE and beyond, are advised to cognitively pre-test questionnaires thoroughly. By doing so, questionnaire developers can detect problems in respondents' cognitive answering process and make improvements by clarifying, for example, abstract or complex terms, leading to a lower cognitive burden on respondents (Lenzner et al., 2010). Consequently, more cognitively valid results are fostered.

A number of limitations of the present study need to be acknowledged. Data were collected by means of two techniques of cognitive interviewing: a think-aloud protocol and systematic probing. Despite the training of respondents to think aloud, a great deal of data could not be coded for their cognitive validity. We found that the think-aloud protocol is not ideal for mapping out cognitive processes. Nevertheless, it remains unclear whether this is due to the difficulty respondents experienced with expressing their thoughts (Royston, 1989), or whether they were poorly executing the required cognitive tasks when answering the items. Either way, a think-aloud protocol appears to be problematic, especially when searching for respondents' interpretation of questions. The problems experienced in this study with regard to the think-aloud protocol also justify the allocation of respondents to two groups in order to ensure that ample information is gathered on all items. When investigating the cognitive validity of a questionnaire, it seems to be particularly appropriate to make use of a systematic probing technique. In such a case, there is no need to split the group of respondents into two unless other reasons would justify doing so.

A more profound analysis of the issue of cognitive validity is imperative. Investigating a limited and specific set of items with a larger group of participants might cast more light on what characteristics of items and respondents are crucial in obtaining cognitively valid results. Furthermore, as this study demonstrates that there is a problem with cognitive validity, it still remains unclear what exact problems are encountered by respondents whereby no completely cognitively valid results are obtained. Further analyses on the available cognitive interview data may concentrate on a more in-depth search, for which drivers for cognitive invalid answers can be found. The results thereof can lead to suggestions for the optimisation and development of new instruments.

Despite these limitations, the current study is, to our knowledge, the first to address the issue of cognitive validity within the context of SSE. The findings suggest that the cognitive validity of SSE survey results is threatened during the interpretation and the elaboration stage of the answering process, while the response stage is less problematic. This detected lack of cognitive validity of SSE results has consequences for the overall validity of the interpretations deduced from SSE results.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Jerich Faddar is a PhD candidate at the Department of Training and Education Sciences (Faculty of Social Sciences) of the University of Antwerp and a member of the Edubron research unit. His current work focuses on methodological issues related to school self-evaluations.

Jan Vanhoof is Associate Professor on the staff of the Department of Training and Education Sciences of the University of Antwerp. He is a member of the Edubron research unit. His current research activities focus on school policy and quality assurance in general and on school self-evaluation and the use of data in education in particular.

Sven De Maeyer is Professor at the Department of Training and Education Sciences of the University of Antwerp. He is a member of the Edubron research unit. His research focuses on educational measurement and methodological issues in educational sciences.

ORCID

Jerich Faddar http://orcid.org/0000-0001-5465-8615

References

Alwin, D. F. (1991). Research on survey quality. *Sociological Methods & Research*, *20*, 3–29. doi:10.1177/0049124191020001001

Alwin, D. F. (2010). How good is survey measurement? Assessing the reliability and validity of survey measures. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 405–434). Bingley, UK: Emerald Group.

Antoniou, P., Myburgh-Louw, J., & Gronn, P. (2016). School self-evaluation for school improvement: Examining the measuring properties of the *LEAD* surveys. *Australian Journal of Education*, *60*, 191–210. doi:10.1177/0004944116667310

Bass, B. M., & Riggio, R. E. (2006). *Transformational leadership* (2nd ed.). Mahway, NJ: Lawrence Erlbaum Associates.

Bateson, N. (1984). *Data construction in social surveys*. London, UK: George Allen & Unwin.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*, 287–311. doi:10.1093/poq/nfm006

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

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.), *Multilevel theory, research, and methods in organizations: Foundations, extentions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234–246. doi:10.1037/0021-9010.83.2.234

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). Oxford, UK: Elsevier.

Christensen, R. H. B. (2015). Ordinal – Regression models for ordinal data (R package Version 2015.6-28). Retrieved from https://cran.r-project.org/web/packages/ordinal/ordinal.pdf

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). London, UK: Routledge.

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

Conrad, F., Blair, J., & Tracy, E. (1999, November). *Verbal reports are data! A theoretical approach to cognitive interviews*. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 89–108). Hoboken, NJ: John Wiley & Sons.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev ed.). Cambridge, MA: The MIT press.

Fleiss, J. L. (1981). The measurement of interrater agreement. In *Statistical methods for rates and proportions* (2nd ed., pp. 212–236). New York, NY: John Wiley & Sons.

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

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, *72*, 892–913. doi:10.1093/poq/nfn059

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*, 330–338. doi:10.1016/j.sapharm.2012.04.004

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Hendriks, M. A. (2000). *Kwaliteitszorg voortgezet onderwijs: Instrumenten en organisaties* [Quality care in secondary education: Instruments and organizations]. Utrecht, The Netherlands: VVO/Q5, project kwaliteitszorg voortgezet onderwijs.

Hendriks, M. A., & Bosker, R. (2003). *ZEBO instrument voor zelfevaluatie in het basisonderwijs. Handleiding bij een geautomatiseerd hulpmiddel voor kwaliteitszorg in basischolen* [ZEBO instrumentation for self-evaluation in primary education. Manual to the computerized instrumentation for quality care in primary education]. Enschede, The Netherlands: Twente University Press.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*, 967–988. doi:10.1177/014920639502100509

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

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

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, *86*, 3–16. doi:10.1037/0021-9010.86.1.3

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*, 254–263. doi:10.1016/j.cedpsych.2010.05.004

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* (pp. 3–90). San Francisco, CA: Jossey-Bass.

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

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29–44. doi:10.1002/ev.1033

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

Lenzner, T. (2012). Effects of survey question comprehensibility on response quality. *Field Methods*, *24*, 409–428. doi: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*, 1003–1020. doi:10.1002/acp.1602

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437–448. doi:10.3102/0013189x07311286

MacBeath, J. (1999). *Schools must speak for themselves: The case for school self-evaluation*. London, UK: Routledge.

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

Maslowski, R. (2001). *School culture and school performance: An explorative study into the organizational culture of secondary schools and their effects*. Enschede, The Netherlands: Twente University Press.

Mathieu, J. E., & Chen, G. (2011). The etiology of the multilevel paradigm in management research. *Journal of Management*, *37*, 610–641. doi:10.1177/0149206310364663

McGee, G. W., & Ford, R. C. (1987). Two (or more?) Dimensions of organizational commitment: Reexamination of the affective and continuance commitment scales. *Journal of Applied Psychology*, *72*, 638–641. doi:10.1037/0021-9010.72.4.638

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*, 63–82. doi:10.1080/03323315.2011.535977

Meuret, D., & Morlaix, S. (2003). Conditions of success of a school's self-evaluation: Some lessons of an European experience. *School Effectiveness and School Improvement*, *14*, 53–71. doi:10.1076/sesi.14.1.53.13867

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

Royston, P. N. (1989). Using intensive interviews to evaluate questions. In F. J. Fowler, Jr. (Ed.), *Health survey research methods* (pp. 3–7). Washington, DC: National Center for Health Services Research.

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

Scheerens, J. (2008). *Review and meta-analyses of school and teaching effectiveness*. Berlin, Germany: Bundesministerium für Bildung und Forsching (BMBF).

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

Schildkamp, K., Visscher, A., & Luyten, H. (2009). The effects of the use of a school self-evaluation instrument. *School Effectiveness and School Improvement*, *20*, 69–88. doi:10.1080/09243450802605506

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

Tourangeau, R., & Bradburn, N. M. (2010). The psychology of survey response. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 315–346). Bingley, UK: Emerald Group.

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

Vanhoof, J., Deneire, A., & Van Petegem, P. (2011). *Waar zit beleidsvoerend vermogen in (ver) scholen? Aanknopingspunten voor zelfevaluatie en ontwikkeling* [Where is policymaking capacity hidden in schools? Cruxes for self-evaluation and development]. Mechelen, Belgium: 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*, 20–26. doi:10.1016/j.stueduc.2010.10.001

Van Petegem, P., Cautreels, P., & Deneire, A. (2003). *IZES Basisonderwijs: Instrument voor zelfevaluatie van basisscholen* [IZES primary education: Instrument for self-evaluation in primary schools]. Leuven, Belgium: Acco.

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

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*, 687–700. doi:10.1177/0013164405282467

Appendix 1. Cognitive validity criteria and coding example for a referent-shift example item

Item		In this school everyone has a clear view on the job descriptions of other school staff.
Item response options		*Totally disagree Disagree Agree Totally agree Don't know*
Cognitive validity criteria		
Item interpretation		School staff have a clear view on the activities and responsibilities of their colleagues. One knows whom to approach for specific questions, and who is responsible for a certain task.
Coherent elaboration		
	Context	Respondents refer to the current state of affairs (at the moment of filling in the survey), applied to the whole staff of their school (a pedagogical entity which is meaningful to the respondent), by means of concrete examples.
	Content	There is a clear task allocation. Tasks are assigned to specific positions in the school and these are well known. This clarity concerns both the own function and the positions of others in the school. Furthermore, there are individual function descriptions for every staff member. These do not need to be formally documented, as long as their informal existence is proven from staff members' behaviour.
Congruent response		Responses reflect the extent to which respondents agree with the given statement on the own school. "Don't know" can be used by respondents when they are ill informed or have no knowledge about the topic which is addressed in the item (e.g., a new teacher at school). Situations in which respondents with sufficient knowledge are doubting on an appropriate response option should not result in the answer "Don't know".
Coding for cognitive validity		
		Hypothetical examples of verbalisation of elaboration stage – content
0	Nothing in line with how item was intended	*"Um, I think everyone knows what is expected from teachers by the government. We had a course during our initial teacher training concerning the different roles that teachers have to fill in."*
1	Elements both in line and not in line with how item was intended	*"Yeah, well, our principal has a conversation with each of the teachers at the beginning of the school year and provides us with our individual job description. Afterwards he even distributes those individual job descriptions among the complete staff. So we do know the responsibilities of each of us."*
2	Everything in line with how item was intended	*"I believe that we have indeed a clear idea of whom we can address with specific questions. We know for instance what we can ask to our secretary or, vice versa, our secretary knows who is responsible for the school's open day. We also know the individual job description of all our colleagues."*