SCHOOL SELF-EVALUATION: SELF-PERCEPTION OR SELF-DECEPTION?

STUDIES ON THE VALIDITY OF SCHOOL SELF-EVALUATION RESULTS

Jerich Faddar

SCHOOL SELF-EVALUATION: SELF-PERCEPTION OR SELF-DECEPTION?

STUDIES ON THE VALIDITY OF SCHOOL SELF-EVALUATION RESULTS.

Jerich Faddar

Universiteit Antwerpen

Faculteit Sociale Wetenschappen

Departement Opleidings- en Onderwijswetenschappen

SCHOOL SELF-EVALUATION:

SELF-PERCEPTION OR SELF-DECEPTION?

STUDIES ON THE VALIDITY OF SCHOOL SELF-EVALUATION RESULTS

Dissertation to obtain the degree of doctor in Educational Sciences

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de Onderwijswetenschappen aan de Universiteit Antwerpen te verdedigen door
Jerich FADDAR

Promotoren:

Prof. dr. Sven De Maeyer
Prof. dr. Jan Vanhoof

Antwerpen, 2018

Composition of the doctoral jury

Supervisors

Prof. dr. Sven De Maeyer, University of Antwerp

Prof. dr. Jan Vanhoof, University of Antwerp

Doctoral commission

Prof. dr. Liesje Coertjens, Catholic University of Louvain

Prof. dr. Bieke De Fraine, Catholic University of Leuven

Prof. dr. Joe O’Hara, Dublin City University

Prof. dr. Peter Van Petegem, University of Antwerp (chair)

TABLE OF CONTENT

DANKWOORD

« Roads are made for journeys
not destinations »
- Confucius

Wie me vroeger had verteld dat ik een doctoraat zou schrijven had ik zonder twijfel gek verklaard. Het voelt dan ook bijzonder vreemd aan om op dit moment te kunnen terugkijken op de voorbije jaren, en vast te stellen dat één en ander toch wat anders uitgedraaid is. Bij het inslaan van de doctoraatsweg had ik nog geen concreet idee wat er allemaal op mijn pad zou liggen. Dat ik hier vandaag sta, zou niet mogelijk geweest zijn zonder de directe of indirecte hulp en steun van vele mensen. In dit dankwoord wil ik enkele mensen graag de verdiende aandacht schenken die in de luwte of de branding hebben bijgedragen aan dit werk.

Allereerst wil ik mijn promotoren bedanken. Jan en Sven, ik herinner me ons eerste overleg in het kader van dit project nog goed. Jullie openden met: “goed Jerich, de middelen hebben we, wat wil je nu eigenlijk doen?”. Overdonderd door de vraag – er lag immers toch een projectplan?! – bleek het tekenend te zijn voor de vele gesprekken die daaropvolgende jaren zouden volgen. Wist ik op dat moment veel dat ik op elk overleg met meer vraagtekens ging buitenstappen dan ik was binnen gekomen. Achteraf gezien is dat nochtans de voedingsbodem geweest voor de weg die ik heb afgelegd. Met elk jullie eigen accenten, gaven jullie me de vrijheid om ideeën te laten rijpen en een richting te kiezen waar ik mezelf goed bij zou voelen.

Jan, onze samenwerking is gegroeid vanuit mijn masterproef waar je toen ook de rol van promotor opnam. Sindsdien is mijn waardering voor je kritische, rationele en analytische geest enkel gegroeid. Je hebt me de voorbije jaren weten uit te dagen en doen groeien als onderzoeker. Je bracht me niet alleen onderzoekscompetenties bij, maar je wist me ook altijd te intrigeren met je diplomatische tactieken en fijnbesnaarde verwoordingen.

Sven, ook jij was altijd bereid om te helpen waar dat kon. Tijdens de vele gesprekken die we hadden over mijn proefschrift liet je jouw methodologisch licht over mijn werk schijnen. Bedankt voor al je hulp bij R-stress, de statistische uitdagingen en ideeën. Je bemoedigende woorden en schouderklopjes nu en dan hielpen me om te blijven doorzetten in de afgelopen jaren.

Jan en Sven, jullie hebben een enorme bijdrage geleverd aan wat hier neergeschreven staat. Maar naast dit professionele aspect ben ik blij dat we het ook persoonlijk met elkaar konden vinden. Het was altijd fijn om elkaar te zien en te vertellen hoe het weekend was geweest of welke sportieve mijlpaal we bereikt hadden. Bedankt voor alles.

Peter, jij volgde mijn werk vanop een iets grotere afstand. Tijdens de begeleidingscommissies kon je mij echter telkens op scherp zetten. Je kritische ogen zagen beperkingen die ik weleens voor mezelf durfde uit duwen. Bedankt voor jouw feedback die zonder twijfel mijn inzichten over mijn eigen werk verrijkt heeft.

Joe, many thanks for taking up the role as member of my guidance commission. It even brought me to your beautiful country, Ireland, for an international research stay! That was without any doubt a tremendously inspiring experience. You enlarged my view on education, education systems and how academics can be excellent hosts for their visitors. Taking me to the Gaelic Football game, exposing me to hard-working Irish flatmates, and sending me to the head of Saint Oliver Plunkett are just a few of the many highlights of my stay. Many thanks for all this, and sure we will stay in contact.

Liesje, hoewel je in eerste instantie mijn werk zou volgen vanop kortere afstand, kreeg dit een wat andere wending en werd dat – letterlijk en figuurlijk – een langere afstand. Bedankt voor je feedback en het opnemen van de rol als lid van de doctoraatsjury.

Bieke, voor jou was dit eerder een ‘achterafje’ vermits het proefschrift immers reeds afgerond was. Bedankt voor het aanvaarden van de uitnodiging om deel uit te maken van de doctoraatsjury, en wellicht zien we elkaar in de toekomst nog.

Natuurlijk vermeld ik ook graag mijn collega’s in dit dankwoord. Ik bedank graag al mijn (ex-) collega’s uit de GK10 en de Meerminne (of Venusstraat) voor de spontane en leuke babbels (en dat waren er veel in mijn geval), de bemoedigende woorden en de constructieve feedback. De ijsjespauzes, de koffies, de pintjes of colaatjes na het werk, raftings op een rimpelloze zee of kerstfeestjes maakten en maken de GK10 een enorm fijne en gezellige werkplek.

Er zijn ook de GK10'ers aan het begin van mijn weg. Alexia, bedankt voor dat fijne onthaal aan de prille start en de toffe jaren vol TALIStivity, samen met Maarten P, Gert (ik houd onze lach-traditie in stand), Elien en Roos op het tweede verdiep en de zolder. Mijn bureaugenootjes van de laatste kilometers bedank ik ook graag. Bedankt Boukje, Margot, Steffi en Loth om tijdens de laatste weken en maanden de stress te verduren. Ik beloof plechtig dat ik vanaf nu mijn rommel ga opruimen.

Dan wil ik nog enkele collega's in het bijzonder vernoemen. Zij vervullen hun rol als collega met verve, maar ondertussen zijn ze veel meer dan dat. Tine, bedankt voor je Eerste Hulp Bij R-Ongevallen, je luisterend oor en je aanstekelijk enthousiasme. Bedankt, Marije, voor je niet aflatende bereidheid tot helpen, je meedenken en alle fijne momenten. Leen, gebaseerd op empirische evidentie konden we samen vaststellen dat we gedeelde interesses hebben, en vooral veel oog voor detail. Kristin, we voelen elkaar goed aan en we kunnen eindeloos blijven kletsen. We generen ons al lang niet meer om tegen elkaar leunend te powernappen in het vliegtuig. Kendra, langlaufpartner in crime. De Botrange zal nooit meer hetzelfde zijn, net zoals het gebruik van de selfiestick. Bedankt voor alle weekendupdates en zalige uitjes, zoals de pubcrawling in Temple bar.

Er is ook zo veel meer dan enkel het academische leven. Daarom verdienen ook vrienden en familie een plaats in dit dankwoord. Mijn ploeggenoten van Heren 3 die de voorbije jaren zonder het altijd te weten wat lichte frustraties hebben moeten compenseren tijdens de training. Bedankt aan heel de bende van de Vrienden om te tonen dat het werk maar een bijkomstigheid is in het leven. Hopelijk kunnen we nog lang doorgaan met telkens fijn bij te praten. Sara en Jeroen, het drukke gezinsleven weerhoudt jullie er niet van om me tijdens de afgelopen periode te steunen en een luisterend oor te bieden. Mooi om Jade, Lente en Lore van dichtbij mee te zien opgroeien!

Ook mijn familie wil ik graag bedanken. Ralph en Rebecca, hopelijk kunnen we in de toekomst nog heel wat tripjes en uitjes plannen. Hier dichter bij de heimat in het Antwerpse, of op de parking, in het West-Vlaamse Ieper. Mama en Patrick, we hebben al heel wat watertjes doorzwommen. Dat ik hier vandaag sta, hadden jullie waarschijnlijk zelf ook niet van tevoren geweten. Toch is het mede door jullie hulp, inspanningen, zorgen, steun en toeverlaat dat dit doctoraat er gekomen is. Het academische zegt jullie niet zo heel veel, maar dat hoeft ook niet. Het feit dat jullie er voor me zijn is meer dan genoeg. Tot slot, Frank, dit traject heb ik niet alleen afgelegd. Jij hebt alles van op de eerste rij meegemaakt met alle mogelijke frustraties erbij. Je stak bovendien al eens handjes toe door een Excelleke op te stellen of teksten na te lezen. De laatste maanden nam je nog veel meer van me over om me de kans te geven dit proefschrift af te ronden. Bedankt voor al die niet-vanzelfsprekende vanzelfsprekendheden.

Benieuwd naar welke roadtrips in het verschiet liggen.

INTRODUCTION

1 Quality assurance and the self-evaluating school

In recent years, schools are increasingly expected to monitor their own quality. School self-evaluation, as a mechanism for internal evaluation, is a key strategy for schools to meet this requirement and has become common practice in many education systems (Eurydice, 2015; McNamara, O'Hara, Lisi, & Davidsdottir, 2011; Nelson, Ehren, & Godfrey, 2015; OECD, 2013; Schildkamp, Vanhoof, van Petegem, & Visscher, 2012). School self-evaluation can be defined as a systematic process, in large part initiated by the school itself, whereby 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 & Van Petegem, 2010). Although school development can be the ultimate goal of an SSE, it is also possible to have an accountability perspective within an SSE framework. An accountability perspective on SSE enables schools for instance to compare themselves to other schools, and to examine how a school performs regarding externally predefined quality objectives. A developmental perspective primarily views the SSE process as a means to evaluate internally defined quality objectives and as a start for a dialogue among school team members about their own vision concerning these quality objectives (Nevo, 2001; Nisbet, 1988; Vanhoof & Van Petegem, 2007).

The above mentioned definition of SSE requires value judgements of participants regarding aspects that relate to a school’s functioning and that determine the quality a school delivers. What aspects of a school make it effective in achieving quality objectives has been extensively researched in the past decades within the field of school effectiveness research (e.g., Creemers, 1994; Scheerens, 2008; Teddlie & Reynolds, 2000). Generally, indicators at different levels or stages of the educational process have been discerned, which have been identified as context, input, process and output indicators of the educational process (Hofman, Dijkstra, & Hofman, 2005; Scheerens, 1991, 2008). All these indicators can be part of an SSE. However, some SSEs tend to focus on elements that are situated at a school’s process indicators. Focussing on a school’s processes, it is argued, can lead to a higher impact on the enhancement of school improvement and effectiveness as these indicators can be manipulated more easily (Scheerens, 1991). Process indicators can be situated at classroom level, such as the quality of instruction, or, at school level, for example in the area of school leadership.

It is critical for an SSE to map out the current state of affairs regarding the school processes that are subject to a particular SSE. However, tapping into school processes is a methodological challenge. While it is argued that “Schools must speak for themselves” (MacBeath, 1999), it is obvious that schools as such are not able to directly come up with numbers or narratives to describe their functioning. Instead, an SSE has to rely on informants to provide a description of the current functioning in order to gather information on a school’s processes (MacBeath & McGlynn, 2002). Based on that description, there is also a judgement required from participants to enable the identification of areas for improvement (Vanhoof & Van Petegem, 2010).

Information can be obtained from SSE participants by means of different methods. These methods are given shape in instruments that are supposed to facilitate the SSE process (Hofman et al., 2005; MacBeath & McGlynn, 2002; Vanhoof, 2007). Next to individual interviews or focus groups with eligible participants, SSE instruments can be designed as questionnaires (e.g., Hendriks, Doolaard, & Bosker, 2001; MacBeath & McGlynn, 2002; Vanhoof, Deneire, & Van Petegem, 2011). The questionnaire method enables the collection of information, of either descriptive or explanatory nature, in a limited amount of time from a large number of participants (Cohen, Manion, & Morrison, 2011).

Many different organisations or stakeholders in the field of education have developed SSE questionnaires (e.g., umbrella organisations of educational networks or consultancy organisations). This variety of questionnaire developers has also resulted in a wide array of instruments. A number of studies have questioned the quality of all these SSE questionnaires (Hendriks, 2000; Hofman et al., 2005). Hendriks (2000) argues that many of the available SSE instruments lack a serious underpinning at methodological and psychometric levels, which leads to uncertainty about the validity of their results. The filling in of questionnaires by respondents such as school principals, teachers, school governors or pupils involves a great dependence on self-reporting of respondents, which leads to SSE questionnaires capturing a perceived reality (Cuyvers, 2002). Whereas the value of respondents’ perceptions about the variables under review in an SSE is acknowledged, it is unsure to what extent this perception is distorted. It is possible that there is a systematic distortion as a result of which a respondent’s self-perception about the school is (intentionally or unintentionally) far removed from how a school is actually performing (Alwin, 2010; Groves et al., 2009; Paulhus, 2002).

This could mean that such a self-perception could for instance easily turn into self-deception, where a respondent is unconsciously depicting his/her school more favourably. This is a serious threat to the aim of arriving at a description of a school's functioning from which valid conclusions can be drawn.

If the completing of SSE instruments is lacking a sound underpinning at methodological and psychometric level and, as a consequence, generates distorted results, this is a possible threat to the validity of conclusions drawn from SSE results (Kane, 2013; Meier & O'Toole, 2013). Moreover, policy decisions and actions undertaken at school (or at overarching) levels may be endangered. Especially in an era where there is an increasing interest in and demand for evidence-informed decision-making in schools, this is a worrying finding (OECD, 2007, 2013; Schildkamp, Lai, & Earl, 2013). Up until now, little is known about the validity of SSE results. As a central research goal, this dissertation aims to answer the question of to what extent results from SSE questionnaires are valid? The next paragraphs will elaborate on what perspective of validity is adopted in this dissertation.

2 Perspectives on validity of school self-evaluation results

In this dissertation, validity is not considered a sum of, or split in different types of validity as inherent characteristics of the SSE instrument. Instead, following an argument-based approach to validity, validity of questionnaire results is ensured by building a strong argument, empirically underpinned, about the interpretations and/or uses of the results (Kane, 1992, 2006). This rationale could also be followed in the context of SSE questionnaires. Kane (1992, 2006) argues that one has to specify precisely what the purpose of, in this case, an SSE instrument is, and what interpretations and uses are related to the obtained results in an *interpretive/use argument*. This is done by laying out the network of inferences and assumptions that explain how observed processes lead to conclusions and decisions. An inference is made when, for example, a sample of questions about a certain school characteristic are judged by a respondent, which means that a respondent's perceived reality is reflected in a reported reality by means of a score (see Figure 1). In a subsequent stage, the respondents' scores form the basis for a next stage of inference where, for instance, statements about the functioning of the school as a whole are made. Inferences are supported by a ‘warrant’ (Toulmin, 1958). A warrant lays out which assumptions need to be met to justify the inference.

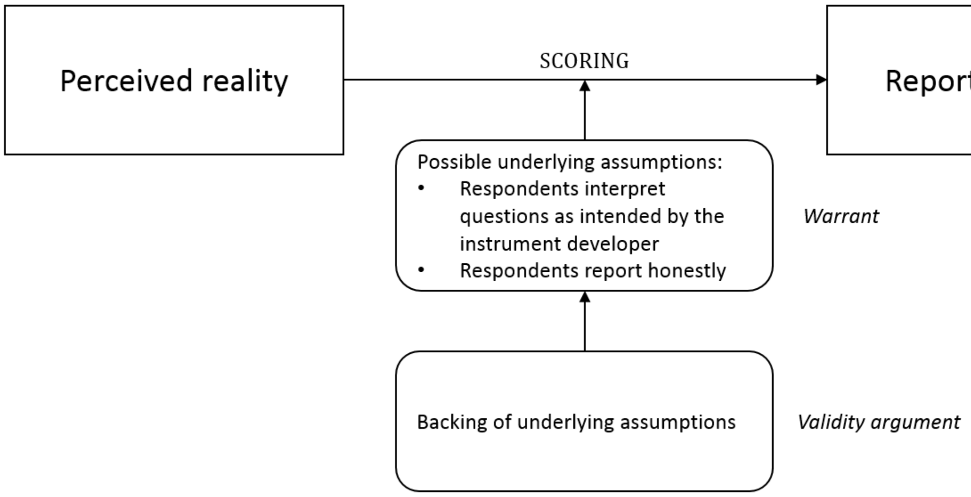

Perceived reality
SCORING
Reported reality
Possible underlying assumptions:
• Respondents interpret questions as intended by the instrument developer
• Respondents report honestly
Warrant
Backing of underlying assumptions
Validity argument

Figure 1 Scoring stage in the argument-based approach to validity in measuring school processes in the context of SSE

The warrants and assumptions need to be evaluated if one wants to make claims about validity (Cronbach, 1988). These evaluations must be provided in a *validity* argument (Kane, 2006, 2013). This implies checks on the clarity of the argument, on whether the proposed interpretations or uses are coherent, and whether the drawn inferences and assumptions are reasonable and plausible. For example, statistical characteristics of SSE questionnaires need to be monitored throughout the development and modifications must be made when required. Another example is to check whether every respondent has interpreted the questions in a similar way. This process provides the backing needed to move to the conclusion that the SSE results obtained are appropriate.

Kane (2006) suggests that in the different stages of inference critical problems can emerge which might undermine the interpretive argument. The stage of scoring for instance, where respondents in the context of SSEs are providing their individual judgement on specific questions, can be vulnerable to (different kinds of) distortion. If procedures are violated during the completion of the SSE questionnaire or if the procedures as such are inappropriately established, the interpretive argument can effectively be overthrown. The current knowledge in survey and methodology literature can help us to understand what sources of distortion might lead to rejecting the interpretative argument in the context of SSE. This literature base identifies different dimensions of error that can occur in the measurement of concepts. The next section discusses how error manifests itself in measurements.

3 Measurement error

Multiple sources of error, which have been identified within the survey methodology literature may also apply in the context of SSEs. Problems can arise with factors beyond the measurement itself, for example, nonresponse error or sampling error. However, in relation to the measurement of processes in SSEs, error can occur as 'measurement error'. This type of error is broken up into two categories: *random error* and *systematic error* (Cote & Buckley, 1987; Groves et al., 2009).

Random error is due to non-systematic influences on the measurement of SSE variables and leads to instability/unreliability in the measurement over several attempts (Groves et al., 2009; Niemi, 1993; Spector, 1994). For example, random error may arise from accidental deterioration of a respondent's attention. Random error levels itself out, and the means across measurements are not affected. Still, variance in variables is increased by random error, and thus weakens the dependencies between them (Niemi, 1993).

Systematic error on the other hand does influence the averages in measurement results. It is systematic as it leads to consistent overestimation or underestimation of the SSE construct being tapped into by the question and also produces errors in the dependencies between variables (Groves et al., 2009; Niemi, 1993; Spector, 1994). For instance, when two SSE variables are mapped out by administering an SSE questionnaire, common method variance can distort the results. Common method variance is attributable to the measurement method rather than to the construct it aims to tap into (D. T. Campbell & Fiske, 1959; Podsakoff, MacKenzie, Jeong-Yeon, & Podsakoff, 2003). Another example is the underreporting of respondents when they are afraid to reveal what is seen by the outside world as bad behaviour (e.g., 'How many times a week are you too late for classes?') (Groves et al., 2009). Systematic error is especially problematic, as averages across measurements are affected and cannot be anticipated by statistical analyses or techniques. Therefore, it can be a critical source of distortion for SSE results.

In the framework of this dissertation, which aims to examine the validity of SSE results, these insights on measurement error are helpful in understanding how validity of SSE results can be impeded. The occurrence of measurement error is caused by not meeting assumptions that underlie a sound measurement. The next section elaborates on what assumptions are in the focus of this dissertation.

4 This dissertation

In response to the central research goal for this dissertation, namely examining the validity of SSE results, it can be argued that, considering an argument-based approach to validity, error can occur at different stages throughout the conduct of an SSE. This might range from the selection of items that are part of the SSE questionnaire, up to the use of the obtained SSE results. There are different assumptions that underlie an appropriate interpretation or use of SSE results. A crucial stage in building a validity argument is when respondents answer items from an instrument (Alwin, 2010; Alwin & Krosnick, 1991; Tourangeau, 2003). This answering process is situated in the scoring stage within the chain of an interpretive and validity argument (Kane, 2006). In the context of SSE, participants are asked to report on items of an SSE questionnaire, generating information on the particular processes which the items are tapping into (see Figure 2Figure 2).

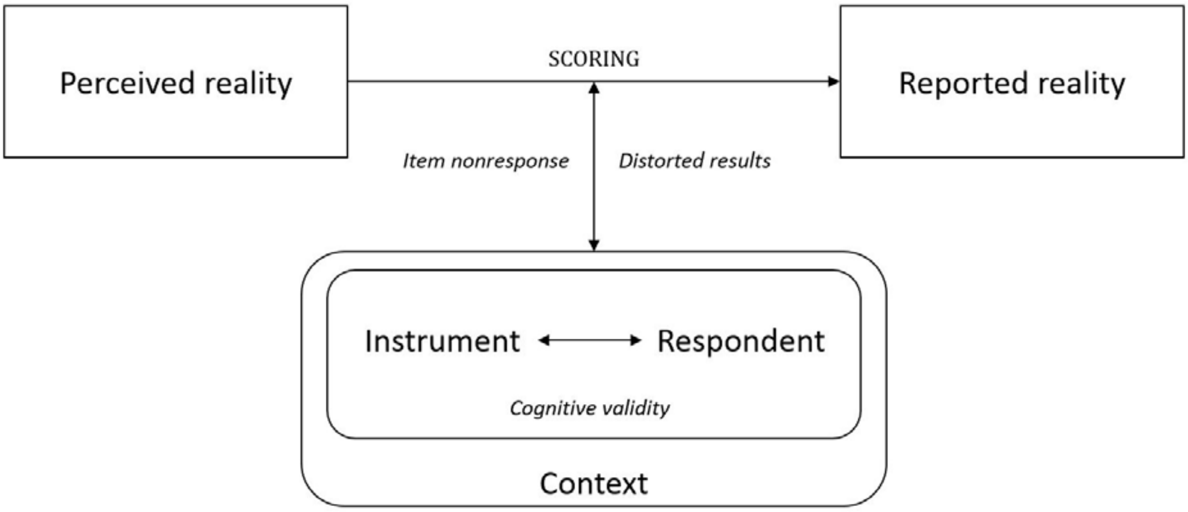


Figure 2 Conceptual model

Respondents are required to perform several tasks when they are asked to complete a questionnaire, of which an appropriate conduct cannot be readily assumed (Karabenick et al., 2007). Although respondents have a vital task in providing their perceptions, general survey method literature points to the insight that a respondent's answer is the result of the interplay between a respondent, an instrument and the context in which the questionnaire is administered (see Figure 2) (Groves et al., 2009; Tourangeau & Bradburn, 2010; Tourangeau, Rips, & Rasinski, 2000). While different assumptions underlie the response to a question and determine the validity of the results that are obtained in the context of SSE, this dissertation examines three assumptions. Firstly, it is assumed that respondents cognitively process SSE questions as intended by the instrument developers, referred to as cognitive validity. Secondly, there is an assumption that respondents' perceptions are not distorted. Finally, it is assumed that respondents are willing to complete each of the items included in the questionnaire. The following sections will elaborate on each of these assumptions which are addressed in multiple studies throughout this dissertation.

4.1 Cognitive validity

In the context of SSE, practitioners often make an appeal to externally developed instruments to map out a current state of affairs about the functioning of the school. The developers of such an instrument aim to make information regarding certain processes explicit. Consequently, respondents are expected to come up with information that is probed for (O'Muircheartaigh, 1999). It is readily assumed that respondents cognitively process items as the instrument developers intended them to be. However, are respondents thinking of the information the instrument developers intended with a particular item? Is a respondent's use of the predefined answering options in line with how the instrument developers intended them? The extent to which respondents cognitively process items in line with the instrument developers' intentions is referred to as cognitive validity (Karabenick et al., 2007; Koskey, Karabenick, Woolley, Bonney, & Dever, 2010). Literature on the cognitive processing of survey items discerns three critical stages in the cognitive process of respondents: interpretation, elaboration and response (Karabenick et al., 2007; Tourangeau & Bradburn, 2010). The interpretation stage comprises several tasks. Respondents need to be able to read and understand the words of the item and to arrive at an interpretation about what is asked in this item (Karabenick et al., 2007; Tourangeau & Bradburn, 2010; Tourangeau et al., 2000). Next, respondents need to elaborate on this interpretation by retrieving relevant information from their memory that relates to the content of this item, and refers to an appropriate reference period and level about which a statement is required (Koskey et al., 2010). Finally, respondents must be able to use the predefined answering options provided by the instrument to reflect their cognitive process. If the instrument developer provides a '*don't know*' option, in case 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. In each of the cognitive stages (interpretation, elaboration and response), respondents can stray from what was intended by the instrument developers. Up until now, little empirical evidence supports the assumption that respondents indeed think of what the instrument developers intended to map out with the items developed.

The cognitive processing of SSE items is highly intertwined with the instrument (Karabenick et al., 2007; Schwarz, 2007). So it must be considered that the design of questionnaire items influences the cognitive process it triggers with respondents. The way in which items are formulated varies across different instruments. Disregarding the specific content of an item, the level to which it refers is of particular interest in the aim of tapping into concepts at a school level (Bliese, 2000; Chen, Mathieu, & Bliese, 2004; Mathieu & Chen, 2011); that is, what the item requires a statement about. Items can require a statement from respondents about themselves, which is a consensus design item (e.g., “I have a clear view on the job description of other school staff”) (Bliese, 2000; Chen et al., 2004). Afterwards, the answers of all respondents can be aggregated onto the school level in case it aims to generate results about the school. Another approach is that items directly require a statement from respondents about the school as a whole, which is called a referent-shift design (e.g., “In this school everyone has a clear view on the job description of other school staff”) (Bliese, 2000; Chen et al., 2004). It could be argued that this difference in design has an effect on the cognitive burden that is placed on respondents, and, consequently has an effect on their answering process (Krosnick, 1991). However, up until now, there is no empirical evidence that in the context of SSE, this difference in item design indeed affects respondents’ answering process. Does this item design (referent-shift vs. consensus design) actually matter in how respondents cognitively process SSE items and in the extent to which results are cognitively valid?

The notion of cognitive validity of SSE results is the central focus of the first two studies in this dissertation. In order to verify whether respondents’ cognitive processes are in line with instrument developers’ intentions, insights in these cognitive processes are indispensable. In order to map out respondents’ cognitive process, cognitive interviews were conducted with 20 primary teachers from four Flemish primary schools while completing an exemplary SSE questionnaire consisting of 20 items (Ericsson & Simon, 1993; Madans, Miller, Maitland, & Willis, 2011; Willis, 2005).

The first study's main research goal is to verify to what extent SSE respondents are cognitively processing items as intended by the instrument developers when answering items (see Figure 3 for an overview of the studies and their main research goals in this dissertation). Based on cognitive validity criteria, which represent the instrument developers' intentions (Woolley, Bowen, & Bowen, 2006), respondents' verbalised cognitive processes are coded for cognitive validity. This is carried out for each of the cognitive stages, leading to 1,200 coding units. In a second stage, cross-classified multilevel analyses identify whether differences in cognitive validity can be attributed to the level of respondents or items. These analyses also explore whether item design predicts the extent of cognitive validity.

Drawing on the results from the first study, the second study aims to examine what problems can be identified during the interpretation, elaboration and response stages of the answering process of SSE items that might explain cognitive invalid results. By means of a content analysis performed on the cognitive interview data, this study aims to examine in depth respondents' reasoning and how these thoughts are not in line with the instrument developers' intentions.

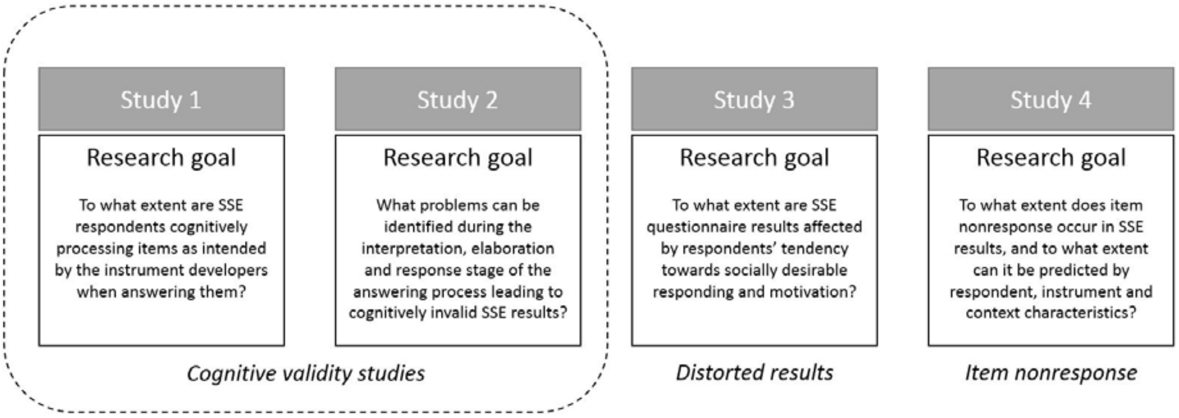

Study 1
Research goal
To what extent are SSE respondents cognitively processing items as intended by the instrument developers when answering them?
Study 2
Research goal
What problems can be identified during the interpretation, elaboration and response stage of the answering process leading to cognitively invalid SSE results?
Cognitive validity studies
Study 3
Research goal
To what extent are SSE questionnaire results affected by respondents' tendency towards socially desirable responding and motivation?
Distorted results
Study 4
Research goal
To what extent does item nonresponse occur in SSE results, and to what extent can it be predicted by respondent, instrument and context characteristics?
Item nonresponse

Figure 3 Overview of studies and their main research goals

4.2 Distorted results

Next to how respondents cognitively process questionnaire items, and whether this is done in line with the instrument developers' intentions, insights from literature on cognitive aspects in the context of questionnaire administration point to the importance of respondents' motivation to read questions and think about relevant information in order to answer them

(Krosnick, 1991). Survey literature has already pointed out that respondent characteristics such as cognitive ability influence respondents' answers to questions (Knäuper, Belli, Hill, & Herzog, 1997; Krosnick, 1991). It could be argued that respondents' motivation to fill in a questionnaire might also influence their response behaviour and the quality of their responses (Bateson, 1984; Cannell, Miller, & Oksenberg, 1981; Heerwegh & Loosveldt, 2009; Krosnick & Presser, 2010). This notion of motivation may also apply in the context of respondents that complete an SSE questionnaire. Generally, motivation is seen as a unidimensional concept that relates to the level of motivation. However, drawing on the self-determination theory (Deci & Ryan, 1985, 2002), not only the quantity of respondents' motivation, but also the quality of their motivation can play a role in their response behaviour. The quality of motivation refers to the nature of respondents' motives. When respondents are autonomously motivated to fill in an SSE questionnaire, they enjoy doing it or at least they see it as a valuable task to achieve their personal goals (e.g., providing high quality education for students). When respondents are to a high extent characterised by controlled motivation to fill in the SSE questionnaire, they engage in the behaviour because they experience an internal or external pressure to do so. If respondents are a-motivated to a high extent, they do not see the value in filling in the SSE questionnaire. Currently, it seems to be readily assumed that respondents' quantity and/or quality of motivation does not matter in arriving at their reported perceptions.

Furthermore, it could be questioned whether in an SSE context, respondents answer genuinely when their perception is sought. Literature in the field of organisational behaviour has, for instance, addressed that respondents tend to respond in a socially desirable way (Donaldson & Grant-Vallone, 2002). Socially desirable responding (SDR) can be defined as a phenomenon where respondents depict themselves overly positively in order to make a positive impression (Paulhus, 1984, 2002). Instead of viewing SDR as a unidimensional construct, research has already demonstrated that SDR can be further broken down into two components (Paulhus, 2002). The first component is impression management, which refers to a respondent that describes him- or herself, deliberately and consciously, in an overly positive way. This behaviour is characterised as a response set since it is a temporary reaction to a particular question or questionnaire. The second component, self-deception, is the phenomenon where a respondent unconsciously and honestly reports an overly positive self-image. This follows the logic of a response style, which causes distortion over time and across different questionnaires. Thomas and Kilmann (1975) argue that SDR is expected to operate in any study with an evaluative overtone, which also makes it a potential cause of distortion in SSE. Furthermore, it has already been found that in the field of education, especially in light of external evaluation, schools can deliberately mislead inspectors by making proactive arrangements in order to be assessed more favourably, which is referred to as window dressing (de Wolf & Janssens, 2007; Perryman, 2009). Evidence on window dressing activities is situated predominantly at school level. Little is known about how individuals' perceptions are distorted in their responses to SSE questionnaires. This raises the question as to whether SDR, which can be seen as a window dressing activity, occurs in the context of SSE, and if so, to what extent results are distorted by this phenomenon. Can users of SSE questionnaire results readily assume that respondents' perceptions are not distorted by their tendency towards SDR?

As its main research goal, the third study in this dissertation examines the extent to which SSE results are affected by respondents' motivation to fill in the SSE questionnaire and by respondents' tendency towards SDR. Regarding respondents' motivation, this study aims to verify whether, next to the quantity, the quality of respondents' motivation to fill in the SSE questionnaire also has an impact on SSE results. Based on the requirement for a high number of participants for this research goal, this study is embedded in an authentic SSE performed in an educational organisation in Flanders. This resulted in an SSE questionnaire completed by 378 teachers. By means of a path analysis, drawing on structural equation modelling, this study gives insight into the extent to which SSE results are distorted by respondent's motivation and their tendency towards SDR.

4.3 Item nonresponse

The search for the distorting effects that impact upon the results of an SSE questionnaire starts from the assumption that respondents are indeed reporting on their perceptions. It is argued in the context of an SSE process that teachers are highly eligible participants since they have day-to-day experience and information to share. However, it can also happen that respondents are unwilling to fill in a complete SSE questionnaire, leading to nonresponse. The issue of nonresponse in administering questionnaires has also been discussed in general survey literature as well (Bosnjak & Tuten, 2001; Groves et al., 2009; Vehovar, Batagelj, Lozar

Manfreda, & Zaletel, 2002). Different types of nonresponse have been discerned in the literature. If a selected respondent does not start to answer the questionnaire, this is referred to as unit nonresponse. When a respondent drops out from the questionnaire at a specific point in the process of completing it because of survey fatigue, for instance, this is called partial nonresponse. Respondents' behaviour can also be characterised by skipping an item or a series of items, after which they restart filling in the questionnaire. There is a failure to obtain information and, consequently, data is missing (de Leeuw, Hox, & Huisman, 2003). This is referred to as item nonresponse. The extent to which item nonresponse threatens the valid use of the questionnaire results depends on the cause of its occurrence. The least problematic is when data are missing completely at random (MCAR). In this case, a respondent might have accidentally overlooked an item. When missingness is related to an observed variable, but not related to the missing response itself, the data are missing at random (MAR). If a respondent does not provide an answer to a question because of the variable under measurement, data are missing not at random (MNAR). In the latter case, missingness is a serious threat to validity and distorts the questionnaire results (R. J. Little & Rubin, 2002; Rubin, 1976). Statistical techniques or strategies to overcome the occurrence of missingness often start from the assumption that missingness is completely at random (de Leeuw et al., 2003; Durrant, 2005). Also, such statistical strategies often require a large dataset, which is mostly not the case in the context of an SSE because schools are not that large by nature. Because of these restrictions in the capacity to anticipate the problem of item nonresponse, it is vital to avoid item nonresponse occurring. However, up until now, the occurrence of item nonresponse has not been studied in the context of SSE.

If SSE respondents are not sharing their perceptions for particular items, this raises the question as to why this is so. General survey literature finds that respondents' behaviour to engage in filling in a questionnaire depends on characteristics of the respondent, the questionnaire and the context (Groves & Couper, 1998; Lozar Manfreda, 2001; Vehovar et al., 2002). Characteristics of respondents are found to affect respondents' intentions to respond to questionnaires. It is argued that answering questions is a demanding task for respondents since they need to carry out several subtasks in order to arrive at a response. Therefore, respondents are required to be motivated to engage in the cognitive process of answering a questionnaire (Heerwegh & Loosveldt, 2009; Krosnick, 1991). The extent to which respondents' quantity and quality of motivation, as adopted in the self-determination theory (Deci & Ryan, 2002), can predict the occurrence of item nonresponse has not yet been examined. Also respondents' tendency towards SDR might not only affect respondents' reported perceptions, as discussed in section 4.2, it might also lead to respondents deciding not to share their perception on an item (Tourangeau & Yan, 2007). In the context of an SSE, respondents can decide to leave an item unanswered as it might reveal a potential weakness in the functioning of their school.

As argued earlier, answering behaviour of respondents is intertwined with the design of a questionnaire. The way items are formulated can increase the cognitive burden placed on respondents. When respondents cannot cope with the complexity of an item, this might lead to the behaviour of skipping an item (de Leeuw et al., 2003). The difficulty can be increased by the topic into which an item aims to tap, but also by the way in which it is designed (Belson, 1981; Krosnick, 1991). The formulation of items that require a statement about the self (consensus design) can be perceived as easier compared to an item that probes for a statement about the school as a whole (referent-shift design) (Chen et al., 2004).

Next to the instrument, the context is also critical for the conduct of an SSE (MacBeath, 1999). As discussed above, there can be a different view on the function of an SSE in a school. A rather developmental or an accountability perspective can characterise the context in which an SSE is conducted (Nevo, 2001; Vanhoof & Van Petegem, 2007). Contexts that are characterised by a strong emphasis on accountability are shown to incite window dressing at a school level, and SSEs can be used in a strategic way in order to make a positive impression (Jansen, 2004; Perryman, 2009; Watling & Arlow, 2002). However, little attention has been paid to how this difference in evaluation perspective might yield an influence at an individual level when respondents are shaping their answers to the items of an SSE questionnaire. Does the assumption hold that the evaluation perspective has no influence on respondents' answering behaviour?

The fourth study included in this dissertation addresses the question to what extent respondents are actually participating in the questionnaire by completing all of the included questions. This study focuses on the phenomenon of item nonresponse, which refers to the extent to which an item or a series of items is unanswered by a respondent, after which he or she restarts completing the SSE questionnaire. The research goal of this study is to identify the extent to which item nonresponse occurs in SSE results, and to what extent can it be predicted by respondent, instrument and context characteristics. This study is based on the collected data in study three, and draws on 378 completed teacher questionnaires. Generalised liner mixed models examine the extent to which item nonresponse can be predicted by respondents' quantity and quality of motivation, their tendency towards SDR, item design and the concept it aims to tap into, and the context (evaluation perspective) in which the SSE is carried out.

5 Methodology

In order to achieve the research goals in each of the studies, this dissertation draws on an array of methods. It can be described as a multi-method design as it involves both qualitative and quantitative strategies (Bryman, 2006; Tashakkori & Teddlie, 2010). In order to address the research goals of the first two studies, respondents' answering process, along with its corresponding cognitive tasks, need to be studied. This requires the collection of in-depth and rich information about respondents' thinking processes. The technique of cognitive interviewing is used to help understand what respondents think when responding to questions, which connects with a qualitative approach of data collection (Ericsson & Simon, 1993; Madans et al., 2011; Willis, 2005). Whereas the first study makes use of a quantitative approach to analyse the cognitive interviewing data, the second study draws on a qualitative analysis method. In order to study the research goals of the third and fourth study in this dissertation, a quantitative approach of data collection is taken. With the objective of studying distorting effects and making predictive analyses, a context is required with a high number of participants making part of one organisation. This enables the exclusion of variance at the organisation level and administering the same questions among the participants.

This dissertation acknowledges that different research methods can fit different research goals. This fits the discourse of a pragmatist research paradigm, in which there is an emphasis on the nature and consequences of human actions in a social context (Schoonenboom, 2017). In line with this pragmatist paradigm, this dissertation approaches its research goals or questions by means of methodologies that are adopted from different research paradigms and strands therein (Morgan, 2014; Schoonenboom, 2017). Corresponding to the research goals of each study, an appropriate design is set up, and a data collection technique chosen.

This implies that choices are made that can link to different strands or paradigms throughout the stages in the conduct of the research (Creswell, 2014; Patton, 2008, 2015; Tashakkori & Teddlie, 1998).

STUDY 1: SCHOOL SELF-EVALUATION INSTRUMENTS AND COGNITIVE VALIDITY. DO ITEMS CAPTURE WHAT THEY INTEND TO?

This chapter is based on:

Faddar, J., Vanhoof, J., & De Maeyer, S. (2017). School self-evaluation instruments and cognitive validity. Do items capture what they intend to? *School Effectiveness and School Improvement*, 1-21. doi: 10.1080/09243453.2017.1360363

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.

1 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 et al., 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 SSEs, 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 (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 et al., 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; Van Petegem, Cautreels, & Deneire, 2003; Vanhoof et al., 2011). 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 et al., 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 et al., 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 multilevel 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.

1.1 Cognitive validity framework

Literature determines three critical steps (see Figure 1) in assessing the cognitive validity of survey responses (Karabenick et al., 2007). Firstly, 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 one stage to another (Collins, 2003; K. E. Ryan, Gannon-Slater, & Culbertson, 2012). The next paragraphs will elaborate more on the details of each of these critical steps.

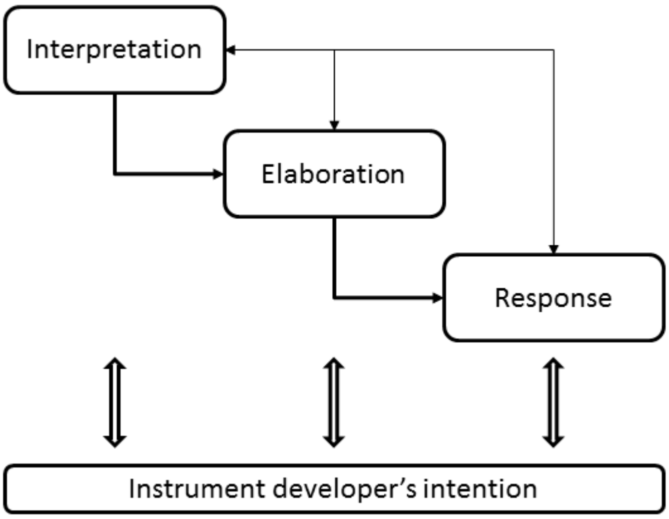


Figure 1 Framework of cognitive validity.

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 et al., 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 organisation level 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.

1.2 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 out 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?

2 Methods

2.1 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 et al., 2011; Collins, 2003; Ericsson & Simon, 1993). To map out the cognitive process of respondents while filling out 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; K. E. 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 out 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 One had to fill out only half of the items while thinking aloud; the other items were administered by means of the systematic probing technique. Group Two also started with the think-aloud protocol, but did so with the items from Group One'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.

2.2 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 '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 four-point scale ranging from 'totally disagree' to 'totally agree', accompanied by a 'don't know' option.

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.*

2.3 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.

2.4 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 et al. (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 the 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.

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

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 the following equation (1), where *i* indexes all observations and *j* the three cognitive validity 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.

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

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 (2). In the results section we referred to this model as Model 1.

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

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.

3 Results

3.1 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 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.

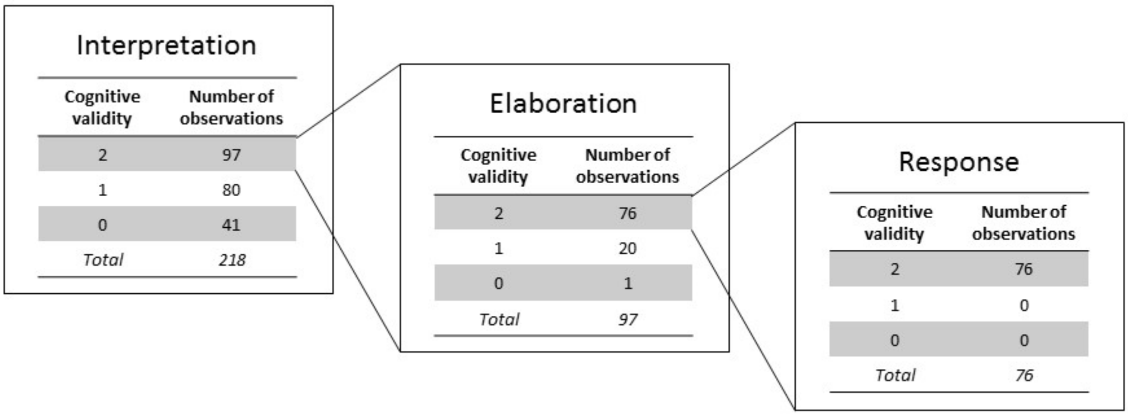
Table 3 Descriptive results on the cognitive validity of respondents' answering processes

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

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 an interpretation and elaboration stage, 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.

3.2 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).

3.2.1 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 (Δdf) 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.

Table 4 Model comparison Interpretation

	AIC	loglikelihood	Δdf	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

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 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 %

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.

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

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 %.

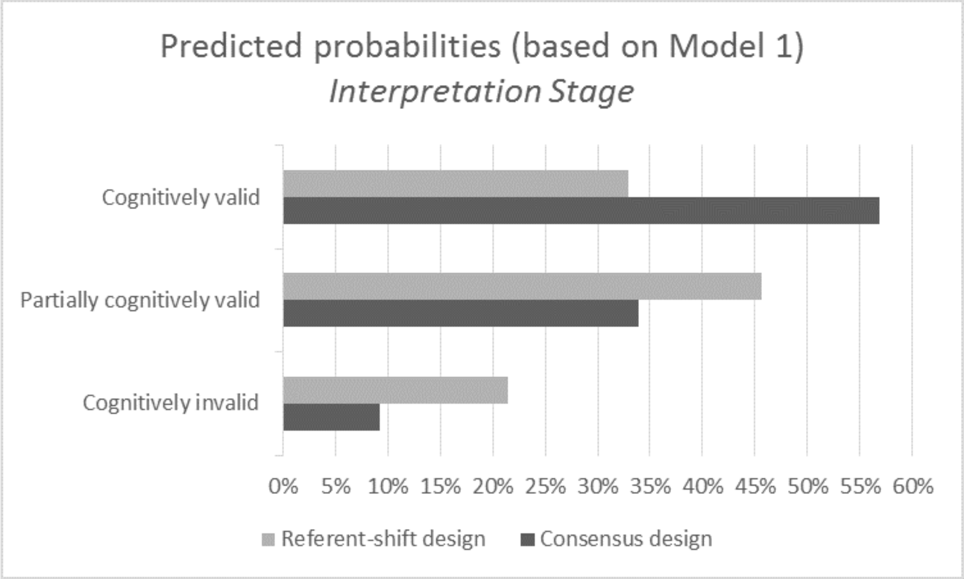


Figure 3 Predicted probabilities for cognitive validity of the interpretation stage by item design.

3.2.2 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 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.

Table 7 Model comparison Elaboration

	AIC	loglikelihood	Δdf	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

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.

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 %

3.2.3 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 ($\sigma^2 = 0.428$) and items ($\sigma^2 = 0.286$). 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).

Table 9 Model comparison Response

	AIC	loglikelihood	Δdf	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

4 Conclusion and discussion

Answering to SSE questionnaire items requires a lot 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 four-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 lot 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.

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 out 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 who 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."*

STUDY 2: INSTRUMENTS FOR SCHOOL SELF-EVALUATION: LOST IN TRANSLATION? A STUDY ON RESPONDENTS' COGNITIVE PROCESSING.

This chapter is based on:

Faddar, J., Vanhoof, J., & De Maeyer, S. (2017). Instruments for school self-evaluation: lost in translation? A study on respondents' cognitive processing. *Educational Assessment, Evaluation and Accountability, 29*(4), 397-420. doi: 10.1007/s11092-017-9270-4

Abstract

School self-evaluation (SSE), as an important leverage for quality assurance, often relies on surveys among staff members to collect information on 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.

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 (MacBeath, 1999; McNamara & O'Hara, 2005; McNamara et al., 2011). SSE can be described as a process, in large part initiated by the school, whereby 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 & Van Petegem, 2010). The procedure of SSE distinguishes between a description of the functioning of the school on the one hand, and a judgment 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-chosen aspects of the school's functioning. School staff have, as it is argued, a good insight in the functioning of the school from their day to day experiences (MacBeath & 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 (MacBeath et al., 2000; Vanhoof, 2007). 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) (Bliese, 2000; Kozlowski & Klein, 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, Vermunt, & Rutte, 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 not only to think 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 an objective reality on the schools‘ functioning (Cohen et al., 2011; Guba & Lincoln, 1994; Patton, 2002). Starting from that point of view, the challenge is to obtain data that reflects this reality. 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 (Hofman et al., 2005; Kane, 2006). 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 (Hofman et al., 2005; Scheerens, 2000; Scheerens & Bosker, 1997).

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 (Karabenick et al., 2007; Schwarz, 2007; Tourangeau et al., 2000). The different stages of this process are shown in Figure 1. While from a theoretical perspective cognitive stages are ordered in the presented sequence, in reality respondents can shift from every stage to another (K. E. Ryan et al., 2012).

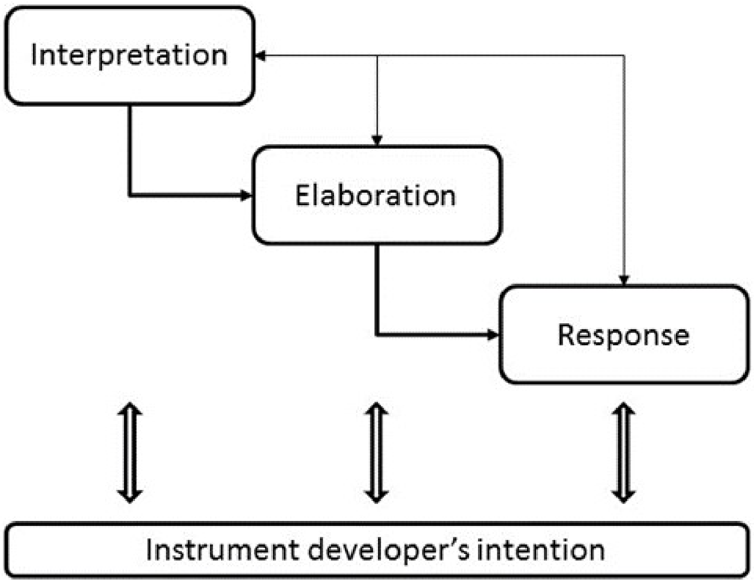


Figure 1 Framework of cognitive validity

First, respondents have to be able to read and interpret the item, involving semantics, syntactics and pragmatics (Lenzner et al., 2010; Tourangeau & 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 map onto multiple underlying representations). Pragmatics can yield problems for example when stylistic elements or other items near to 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 judgment, based on the preceding cognitive stages. This judgment is then formulated in 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, 2014; Krosnick & 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.

Across different contexts the way in which respondents interpret their environment and the task they are asked to perform, depends on the underlying assumptions, values and knowledge they have available on the occasion with which respondents look upon it (e.g., Babik, Singh, Zhao, & Ford, 2015; Mohammed & Ringseis, 2001). The same applies in the context of survey questions wherein respondents are asked for certain information by describing and judging aspects of a school's functioning (Cannell et al., 1981; Tourangeau et al., 2000). From a purely cognitive standpoint, these underlying assumptions, values and knowledge vary among individuals, result in an individual point of reference or mental model (Johnson-Laird, 1983). Asking respondents to make a description or even a judgment on a statement implies an interaction with this point of reference or mental model (Cannell et al., 1981; Tourangeau et al., 2000).

It is often pursued in the context of SSE that different voices are heard. Participants holding different positions in the 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, Day, Arthur Jr, & Bell, 2006; Kozlowski & Ilgen, 2006). This draws on the perspective that these respondents have another background and a different expertise which can generate complementary information on the organisation's functioning (Kyriakides & Campbell, 2004; MacBeath, 1999; MacBeath et al., 2000). 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; Day et al., 2010; OECD, 2014). It is however unknown how these differences affect the cognitive process of respondents in answering an SSE surveys.

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; Krosnick, 1991; Lenzner, 2012). 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 surveys' 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 (Ericsson & Simon, 1993; Presser et al., 2004; K. E. Ryan et al., 2012; Willis, 2005). 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 & 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 & Willis, 2007; Blair & 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 & 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. This resulted in cognitive validity criteria that fed into the analysis as a point of reference. Along the analysis of the interviews clarifying questions were addressed to ensure the instrument developers' intentions were properly reflected in the cognitive validity criteria.

This section outlines the instrument developers' intentions with the designed items for two exemplar items which will be referred to, among other, in the results section. 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.

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 exceeds 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.

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 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 1,200 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 & 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, 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 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 a deductive and inductive coding approach 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 the results 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.

Lost in giving meaning to words

First, at the semantic level, respondents experienced difficulties with terms or concepts that are formulated in 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 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 2.3), 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.

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 categorized in two aspects: content and context.

3.2.1 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.

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)

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)

Elaborating out of scope

A 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)*

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)

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)

3.2.2 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.

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 judgment 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)

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)

3.3 Response stage

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 judgment 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)

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 judgment.

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)*

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)*

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.

3.4 On the respondent's position in the school

The following part of the results section focuses 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.

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.

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)*

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)*

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 conceptualization 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 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 respondents influences their cognitive processes (RQ2).

Results show that respondents have problems with the interpretation of some specific terms. Giving words or concepts an 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 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 (Karabenick et al., 2007; Tourangeau, 2000). Regarding the contextual aspect of an elaboration, it is found that respondents can mistake the appropriate timeframe 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 (Krosnick, 1991; Lenzner, 2012). Furthermore, respondents sometimes lack an answering option that reflects their mental judgement. Findings regarding the use of predefined answer options connects 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 & 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 have also 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. 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.

5 Discussion

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 & 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.

In the chosen method of surveys it is inherently assumed that respondents make the same interpretation of questions asked. This links to the post-positivistic perspective where reality is viewed as an objective truth that is attempted to capture with a measurement instrument (Cohen et al., 2011; Patton, 2015). Simultaneously it is pursued that different perspectives bring in valuable information in the process of SSE. However, the extent to which respondents are expected to bring in different perspectives leading to valuable insights sets off from another research paradigm. The (socio-) constructivistic point of view, where different perspectives are welcomed constructing

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 SSE instruments' quality, this study illustrates the importance of making instrument developers' intentions explicit for every item. The cognitive stages are a suitable framework to that end. Furthermore, instrument developers should conduct a thorough cognitive pre-test 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 & Le Floch, 2004; Madans et al., 2011; K. E. Ryan et al., 2012).

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 a 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 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 (Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006; Saunders, 2000; Williams & Coles, 2007).

The findings of this study also urges a consideration of the use of survey instruments in the process of SSE. 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 facilitating this there should be given priority at policy level in enhancing the (self-)evaluation capacity of participants and schools. This demands efforts at individual participant level as at organisational and leadership level (Preskill & Boyle, 2008). Investing more in the development of tools and resources and an adequate (external) guidance are vital to generate a valid picture of a school's functioning (O'Brien, McNamara, O'Hara, & Brown, 2017). Especially in a context where much emphasis is put on the importance of evidence-based decision-making in 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 (K. E. 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.

Appendix 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.

STUDY 3: SCHOOL SELF-EVALUATION: SELF-PERCEPTION OR SELF-DECEPTION? THE IMPACT OF MOTIVATION AND SOCIALLY DESIRABLE RESPONDING ON SELF-EVALUATION RESULTS.

This chapter is based on:

Faddar, Vanhoof & De Maeyer (revisions submitted) School self-evaluation: self-perception or self-deception? The impact of motivation and socially desirable responding on self-evaluation results.

Abstract

In order to enhance the quality of education, school self-evaluation (SSE) has become a key strategy in many educational systems. During an SSE process schools describe and evaluate their own functioning, often by administering questionnaires among teachers. However, it is unknown to what extent SSE questionnaire results are distorted by respondents' tendencies towards socially desirable responding and their motivation to fill in an SSE questionnaire. This study reports on a path analysis, performed on the results of an authentic SSE with 382 participants. Results indicate that socially desirable responding and motivation have indeed an impact on SSE results. However, the effects are differential and depend on the variable of interest. These findings can have serious implications, and should be taken into account when drawing conclusions and taking (school) policy decisions within the framework of an SSE process.

1 Problem statement

In many educational systems school self-evaluation (SSE) has become a key strategy, next to external evaluation, in efforts to safeguard the quality of education (OECD, 2013). SSE is a mechanism where the school itself takes the initiative to systematically describe its functioning by stakeholders. Drawing on this description an evaluation is made of the school, leading to the consideration of policy decisions and the undertaking of actions (Vanhoof & Van Petegem, 2010). In order to create a description of a school as an organisation there is the need to measure constructs at organisational level, which sets a methodological challenge. Because a school as such cannot speak for itself, SSE often relies on collecting data among teachers or other stakeholders. They are, it is argued, involved in the everyday functioning of the school, and therefore, highly eligible to provide insightful information on their school. In order to capture this information several instruments have already been developed. Often, these are designed as a questionnaire that probes for respondents' perceptions regarding school processes which are found in the literature on school effectiveness (e.g., Hendriks, Doolaard, & Bosker, 2002; MacBeath et al., 2000).

The use of SSE questionnaires is stimulated for several reasons, while literature on survey research, however, already points to methodological concerns with regard to the use of questionnaires as a data collection method. Several factors can distort the answers of respondents on questionnaires, such as the mode of administration (e.g., paper-pencil vs. computer-assisted) or the difficulty of items (Belson, 1981; Krosnick & Presser, 2010). Also, respondent characteristics may be at play, affecting the quality of questionnaire results. There are indications that respondents' tendencies towards socially desirable responding (SDR), a phenomenon where individuals give overly favourable self-descriptions, can impact data quality (Lam & Bengo, 2003; Wayne & Liden, 1995). Moreover, individuals who score highly on SDR scales are "faking good". As a result, individuals' (self-) reports of different concepts in the broad field of SSE research and beyond could be considered distorted. Thomas and Kilmann (1975) argue that this mechanism is expected to operate in any study where ratings are used to assess variables with evaluative overtones. Until now, it is unknown to what extent SSE results and other self-reported measures in the context of SSE are affected by SDR. Next to the issue of SDR, Bateson (1984) points to respondents' willingness to provide an answer as a crucial condition for quality data. When respondents are not motivated to provide an accurate response, they could start relying on response strategies that lead to acceptable, yet inaccurate responses (Krosnick, 1991). In order to be able to make valid conclusions out of SSE questionnaire responses, respondents should be motivated to engage in the cognitive processes that are required to produce an accurate answer to the items. This seems to suggest that not only the quantity or amount of motivation matters, but quality also is an important dimension of motivation. Quality of motivation refers to the underlying attitudes and goals that lead to the action (R. M. Ryan & Deci, 2000). It is expected that SSE participants vary in the quality of their motivation to engage in the SSE process and filling in of the SSE questionnaires. However, it is not known to what extent this quality of motivation can explain respondents' answers on SSE questionnaires. At this moment it is readily assumed that both respondent characteristics (their tendency towards SDR and the quality of motivation to fill in the SSE questionnaire) has no influence on SSE data quality. Nevertheless, valid conclusions on SSE data are of key importance. Especially in an era where there is a strong emphasis on data-based or evidence-based decision making (OECD, 2007; Schildkamp et al., 2013). As gathering data on a school's functioning is part of the SSE procedure, it meets the current tendency towards more attention for data-based decision making in education (C. Campbell & Levin, 2009; Schildkamp et al., 2013), especially because the collected data are used as a basis for school development plans and policy decisions. In light of the implications at school policy level, it is of the utmost importance that the conclusions drawn from SSE data are valid.

This study aims to explore how respondents' motivation and their tendency towards SDR can cause distortions in the different self-reported scores that are obtained in an SSE. As a result, this study will examine how these different variables relate to each other in all their complexity. The manuscript will focus on the extent to which SSE respondents differ in their tendency towards socially desirable responding and their motivation to fill in an SSE questionnaire. Furthermore, the extent to which SSE questionnaire self-report data are affected by respondents' tendency towards socially desirable responding and motivation, is examined.

2 Conceptual framework

The following sections explore and elaborate the key concepts of this study. It sets off with a clear depiction of what school self-evaluation is, and what can be the subject of an SSE process. Next to this, an exploration is made of what is understood by socially desirable responding and the quantity and quality of respondents' motivation.

2.1 School self-evaluation and self-report

School self-evaluation (SSE) is a form of internal evaluation counterbalancing a tendency in many educational systems to rely on external evaluations to guarantee educational quality (OECD, 2013). In this study SSE is defined as '*a systematic process, largely initiated by the school itself, where participants describe and evaluate the functioning of the school for the purposes of making decisions or undertaking initiatives in the context of (aspects of) overall school (policy) development*' (Vanhoof & Van Petegem, 2010, p. 20).

SSEs are requiring value judgements of participants regarding specific indicators that relate to the quality schools deliver, or the functioning of a school as referred to in the above definition. These indicators have often their origin in school effectiveness literature (Scheerens, 1991, 2008; Van Petegem, 1998). Indicators can be situated at different levels and/or stages of the educational process. Generally, the following categories can be discerned: input indicators, process indicators, output indicators and context indicators (Ikemoto & Marsh, 2007; Scheerens, Glas, & Thomas, 2003). An SSE's focus could for instance be narrowed down to hard output indicators such as pupils' results in standardized assessments in reading skills. However, it has already been demonstrated that focussing on school process indicators can lead to more impact on the enhancement of school improvement and school effectiveness as these are more easily manipulated (Scheerens, 1991). Typically, this involves concepts such as the quality of instruction, being a learning organisation or being characterised by distributed leadership (Muijs, Harris, Chapman, Stoll, & Russ, 2004; Reynolds, Sammons, Stoll, Barber, & Hillman, 1996; Scheerens, 1991). In this study there are two central SSE process variables of interest. Although both being process variables they are situated on a different level within the school. Distributed leadership, as an indicator for policymaking capacity in schools (Van Petegem, Devos, Mahieu, Dang Kim, & Warmoes, 2006), is typically situated at the school level and can be described as a form of collective leadership where each team member is empowered to share their expertise to lead collectively (Harris, 2004). Differentiation, however, taps in this study into the classroom level, and focuses on the extent to which teachers adapt their instructions to the particular situation and needs of students (Tomlinson et al., 2003). As these process variables cannot be measured directly, since schools cannot speak for themselves as such, this conceals a methodological challenge. Consequently, SSE relies on the perception of stakeholders, or other well-chosen participants, who can provide insightful information on the topic under review (MacBeath, 1999). Literature points, next to factors at instrumental level, to respondents' characteristics as sources for measurement error . To what extent these characteristics influence the results of SSEs is yet unknown.

2.2 Socially desirable responding

The phenomenon of socially desirable responding is known to influence individuals' behaviour in many different contexts. It has been found that individuals tend to over-report engaging in behaviour which could be described as socially desirable; for example when teachers are asked to self-report on change in their instructional practices during math classes (Lam & Bengo, 2003). But they may also under-report when they are asked about socially undesirable behaviours such as the extent to which they are confronted with discipline difficulties in their classrooms. In essence, individuals vary in the extent to which they depict themselves as overly positive (Paulhus, 2002). As such, when SDR interferes with an accurate response on self-report items, it is considered to be a source of response bias. Holtgraves (2004) identifies three different stages in the cognitive process where SDR can take place. First, SDR can operate during the editing of a response. After having formatted a response, respondents make an evaluation of their response in terms of social desirability. Secondly, the stage of retrieval can simply be bypassed due to SDR. Respondents are basing their answer only on socially desirable implications of the answer, and are not making any attempt to retrieve relevant or accurate information about the item. A last possibility is that respondents are indeed retrieving information, but in a biased way. They selectively retrieve information which places them in a more favourable light. This retrieval is then aimed at confirming one's inquiry and ignoring contradictory information.

The phenomenon of SDR slipping into the answering process of respondents at different stages also raises the question whether different kinds of response behaviour can be discerned. Although SDR has been seen as a unidimensional concept (e.g., Crowne & Marlowe, 1960), research on SDR already demonstrated that a more nuanced view on this phenomenon is advocated (Helmes & Holden, 2003). For a further operationalisation of the concept of SDR, it is important to acknowledge that response bias can be caused by a response *style* and/or a response *set* (Paulhus, 2002). A response style is found to bias individuals' responses over time and, as such, across different questionnaires. A response set, in contrast, occurs only temporarily and is a reaction to a particular question or questionnaire. This distinction is followed in the further conceptualisation of SDR where several authors have suggested a two-component model (Millham & Kellogg, 1980; Paulhus, 1984). Paulhus (1984, 2002) identifies a first component as *impression management*. It is characterized by an individual, deliberately and consciously, describing her- or himself in an overly positive way in response to a certain question or questions, this being a response set. The second component is *self-deception* where a respondent unconsciously and honestly reports an overly positive self-image across questionnaires and time, which follows the logic of response style.

Although there is a debate on the extent to which SDR is a problem in self-report measures, Thomas and Kilmann (1975) argue it is likely that SDR occurs in any context where variables are measured with an evaluative overtone. When SSE is performed either in a developmental or in an accountability oriented context, an evaluation aspect is involved, and implications on the validity of SSE results are to be expected. Furthermore, literature has also demonstrated that the specific traits that are under review can influence the way in which respondents answer them. One could argue that indicators that are closer related to participants' individual responsibilities, such as differentiation, could be perceived as more sensitive questions, and consequently evoke social desirable responding compared to indicators that are more remote from participants individual responsibilities, as distributed leadership (Moorman & Podsakoff, 1992). It could also be argued that both differentiation as well as distributed leadership are prone to SDR. Respondents that are asked about one of their own competences, in this case a skilled one as differentiation, might be triggered to respond in a socially desirable way in order to be seen as a highly competent teacher. The same might be true for distributed leadership, which is a characteristic not commonly found in schools. This might prompt respondents to depict the school more favourably. Also, the evaluative aspect of the SSE might contribute to the occurrence of SDR in the reported SSE scores (Thomas & Kilmann, 1975).

2.3 Respondents' motivation

A key condition for quality data is the willingness of respondents to fill in the SSE questionnaire (Bateson, 1984; Krosnick, 1991). If people do not feel an impulse to act, they are described as unmotivated. When people are activated to an end, they are considered to be motivated. Much research has addressed the concept of motivation as a unitary construct ranging from little motivation to a lot of motivation. However, self-determination theory (SDT) adds, next to the *quantity* of motivation, another dimension: the *quality* of motivation (Deci & Ryan, 2002; Vansteenkiste, Sierens, Soenens, Luyckx, & Lens, 2009). The quality of motivation refers to the reasons why individuals engage in a particular behaviour, and these can vary substantially. According to SDT this variation in the quality of motivation, is due to the extent to which reasons or motives to engage in behaviour are internalised (Deci & Ryan, 2002). This internalisation is a process in which initial external values as a reason to regulate behaviour are becoming part of the self. Drawing on the dimensions of quantity and quality of motivation, which are integrated by SDT, many studies make the distinction between the following types of motivation: a-motivation, autonomous and controlled motivation (Deci & Ryan, 2002; Vansteenkiste, Lens, & Deci, 2006).

Considering the quantity of motivation, a first concept is discerned when an individual lacks motivation. *A-motivation* refers to the lack of motivation to engage in filling in a questionnaire or having no intention to do so. A person can have no trust in achieving a desired outcome, there can be no feeling of competence to execute the task, or it can be perceived as non-relevant (Deci & Ryan, 1985).

Autonomous motivation is characterised by a feeling of freedom, and a person’s reasons to engage in filling in a questionnaire can be described as more or less self-determined. They engage in this activity because of sincere interest, and perceive it as inherently enjoyable and satisfying (R. M. Ryan & Deci, 2000). This type of motivation, with its reasons for engaging in filling in a questionnaire, is situated rather on the higher end of the continuum of internalisation. Often, autonomous motivation is further subdivided into intrinsic motivation, with the highest amount of internalisation, and identification, where individuals engage in certain behaviours as they believe it helps to attain their personal goals.

Individuals, who, by contrast, experience a pressure to fill in a questionnaire, are driven by c*ontrolled motivation*. A subdivision can be made according to the attribution of this pressure. When the experienced pressure is the result of internal feelings of shame or guilt, then it is referred to as introjected regulation. When pressure is external to the self, as in the case of receiving incentives, avoiding punishments or meeting expectations of others, it is referred to as external regulation (Vansteenkiste et al., 2006). Obviously, external regulation is at the lower end of the continuum of internalisation.

3 Method

3.1 Context and participants

In order to examine the aforementioned research questions, the study was embedded in a self-evaluation performed in an educational service organisation in Flanders (Belgium). This organisation provides education and training in several disciplines, ranging from general education over technical education to vocational education. Students can enrol from the age of 16, but the main target students are adults. The SSE dealt with different topics of the organisation's functioning, however, this study will focus on two constructs of interest as typical cases. One variable is a typical organisational construct that is often mentioned in school effectiveness literature and widely debated in the field of school improvement: distributed leadership (e.g., Hallinger & Huber, 2012; Muijs & Harris, 2003). It is a way of thinking about leadership that focuses on the engagement of existing expertise scattered within the school, rather than sticking to formal or hierarchical positions and roles (Harris, 2004). The second variable is about teachers' individual practices within their classrooms, and focuses on differentiation. A differentiated classroom instruction can be defined as a systematic, proactive way of providing instruction tailored to the specific needs and differences between students (Tomlinson et al., 2003).

The teaching staff was familiarised with the notion of self-evaluations as they were asked to evaluate their own teaching on a yearly basis. These self-evaluations are performed from a rather developmental perspective in order to use the results as input for further development of the provided educational quality. Previous self-evaluations were conducted by means of a self-report questionnaire, which means that the teaching staff has experienced in this method. All teaching staff were invited to participate in the self-evaluation, which was administered by means of an online questionnaire. A response rate of 58 % was attained, resulting in 382 completed questionnaires.

3.2 Instruments

For all concepts in this study, items from existing instruments were used and brought together in the SSE questionnaire. However, the instruments were adopted in a new context that urges us to verify their psychometric qualities for this study. To measure socially desirable responding we relied on the Paulhus Deception Scales and selected items that allow us to tap into both concepts of impression management and self-deception (Paulhus, 1998). All 40 items were translated from English into Dutch. As these scales were developed on a body of literature that overarches different contexts and disciplines, an explorative factor analysis (with oblique rotation) was performed to ensure that the two factors were retained in the data. Table 1 includes a sample item of each subscale and reports on the scales' Cronbach's alphas.

Items of the Academic Self-Regulation Questionnaire (SRQ-A) were adopted to tap into the concepts of intrinsic motivation, identified, introjected and external regulation (Vansteenkiste et al., 2009). As these original items were situated in the context of learning behaviour, the items were rephrased so that the behaviour of filling in a questionnaire became central in each item. A-motivation is, however, not integrated into the SRQ-A and therefore we included and rephrased items from the Academic Motivation Scale (AMS) (Vallerand, Blais, Brière, & Pelletier, 1989). As these items were administered in a new context we started with an explorative factor analysis (with oblique rotation) to identify factors based on the data. With regard to the dimension of the quality of motivation, results indicated that only two factors could be discerned in the data. Clearly, two components were retained in the data: autonomous motivation on the one hand, and controlled motivation on the other. Some items were omitted from further analyses because they did not load well on the factors. Possibly, these items did not fit very well the context of answering a questionnaire as the original instrument was constructed to serve in an academic context (e.g., "I fill in this questionnaire because I find it a pleasant activity"). Cronbach's alphas for these scales are satisfying and show no problematic inconsistencies (see Table 1).

The dependent variables in this study, distributed leadership and differentiation, were measured by means of a scale specifically designed for the field of education. For distributed leadership a scale, consisting of six items, was adopted from an instrument to tap into the policymaking capacities of schools (Vanhoof et al., 2011). The extent to which teachers differentiate during their lessons is measured by means of a scale that is borrowed from an existing instrument, which aims to examine whether teachers meet basic teaching competences (Meynen, Struyf, & Adriaensens, 2011). Cronbach's alphas for these scales were satisfying (see Table 1).

Table 1 Instrument's scales, reliabilities and sample items

	N° Items	Cronbach's alpha	Sample items
Socially desirable responding			
Impression management	20	0.79	I sometimes tell lies if I have to. *
Self-deception	20	0.72	I am fully in control of my own fate.
Motivation			
Autonomous motivation	6	0.87	I fill in this questionnaire because I personally find this very valuable.
Controlled motivation	7	0.76	I fill in this questionnaire because others expect me to do so.
A-motivation	4	0.82	Honestly, I don't know. I really feel that I am wasting my time when I'm filling in this questionnaire.
Distributed leadership	7	0.93	On this campus everyone has sufficient possibilities to engage in decision-making procedures.
Differentiation	5	0.74	I take into account the different pace of students.

Note: *= negatively phrased item.

3.3 Analyses

The conceptual framework identified that SDR can have an impact on any kind of self-report, so it must be acknowledged that it could also affect the scores obtained for motivation in our study. Therefore, we decided to identify the effect of SDR on motivation and the SSE variables in a direct and indirect way. This enabled us to control for an SDR effect on respondents' statements about their motivation to fill in the questionnaire on the one hand, and to the relation between motivation and the scores on the SSE variables of interest. In order to accurately estimate the relationships between all these variables, we ran a path analysis by making use of structural equation modelling (SEM). This technique allows us to run complex models with latent variables, making use of measures at item-level and multiple indicators for one latent variable. In this sense, the strength of this technique lies in the fact that it combines a measurement model, using confirmatory factor analysis, and a regression model (Kline, 2015; Ullman & Bentler, 2003). Several indices for model fit were consulted to ensure the quality of our analysis. The Comparative Fit Index (CFI), which makes a comparison between the assumed model and a null model without relationships, indicates an acceptable model fit when higher than 0.90 (Hu & Bentler, 1995; Schreiber, Nora, Stage, Barlow, & King, 2006). The

Root Mean Squared Error of Approximation (RMSEA), which gives an indication of how well the model would fit the population, should not be higher than 0.06 and the Standardised Root Mean Squared Residual (SRMR), which gives an indication of the difference between the predicted and actual matrix; below 0.08 is considered to be acceptable (Hooper, Coughlan, & Mullen, 2008; Hu & Bentler, 1999). Modification indices were consulted in order to optimise the model if the initial model did not fit. Respondents who missed out on an item or did not respond to one of the variables in the questionnaire were retained in the analysis by estimating missing data with Full Information Maximum Likelihood (FIML). This technique performs well in comparison to other techniques for handling missing data (Enders & Bandalos, 2001).

Given the high number of parameters in our model, especially for the scales tapping into SDR, we parcelled the items of impression management and self-deception into four parcels of five items. Each parcel serves as an indicator for the respective latent constructs (T. D. Little, Cunningham, Shahar, & Widaman, 2002; Matsunaga, 2008). The allocation of items to parcels was done randomly and repeated 20 times. For each allocation solution a satisfying level of fit was obtained for the measurement model of the SDR scale.

The strategy of random parcel-allocation for the SDR scale was also adopted in the comprehensive path analysis, leading to 20 SEM models. The fit indices for all estimated models can be found in**Table** Table 2. Only slight differences were found across these different models. One of these 20 models is selected and presented in the results section. It is representative to all other models, as all reported significant relationships were found in every estimated model.

Table 2 Average and range of fit indices of 20 path models using parcelling for social desirability scales

Fit indices	Average	Min	Max
CFI	.912	.904	.920
TLI	.903	.894	.911
RMSEA	.044	.042	.046
SRMR	.073	.070	.076

Relationships in a path model between independent and dependent variables can be of a direct or indirect nature, because of the presence of one or more mediating variables (Alwin & Hauser, 1975). First, the direct relationships will be discussed in the results section. Path analysis, however, also enables us to calculate the indirect effects which can be added up to the direct effects resulting in a total effect parameter. These total effects show the overall effect of the independent variables on the dependent variables, including the effect these have on the mediating variables. The analyses were run using the R-package lavaan (Rosseel, 2012).

4 Results

In the first part of the results section, some descriptive information discusses the respondents' tendency towards socially desirable responding and their motivation to fill in the SSE questionnaire. The second part focuses on the explanatory analyses.

4.1 Descriptive results

With regard to socially desirable responding, respondents are found to have a rather moderate tendency to describe themselves as overly positively. The mean score for self-deception is 3.50 (see Table 3). The standard deviation (*SD* = .58) is found to be rather small, which means that the difference between respondents in our sample is rather small. Respondents score slightly higher for their tendency towards impression management (*M* = 3.70), and the results also show more spread in responses with regard to this concept among the respondents (*SD* = .77).

Respondents' motives to fill in the SSE questionnaire are not highly internalised. Although there are some differences between respondents, they do not see the administration of the SSE questionnaire as an interesting activity, or at least as a valuable activity to achieve their personal goals (*M* = 3.42; *SD* = .77) (see Table 3). To a lesser extent, respondents experience the administration of the SSE questionnaire with a feeling of pressure (*M* = 2.24; *SD* = .76). However, the mean score for a-motivation (*M* = 2.33) is higher than for controlled motivation. This means that respondents tend to be more unmotivated to fill in the questionnaire than they are controlled motivated. For a-motivation, the spread among respondents is highest for all motivation scales (*SD* = .88).

Regarding distributed leadership, respondents are quite critical about it with a mean of 3.42 (see Table 3). However, of all the administered scales, respondents differ most in their opinion about distributed leadership in their school (*SD* = .92). Respondents are slightly more positive about the extent to which they differentiate in their classrooms (*M* = 3.73), and they are more unanimous about this judgement with a standard deviation of .67.

Table 3 Range, mean and standard deviation for scales on socially desirable responding, motivation, distributed leadership and differentiation

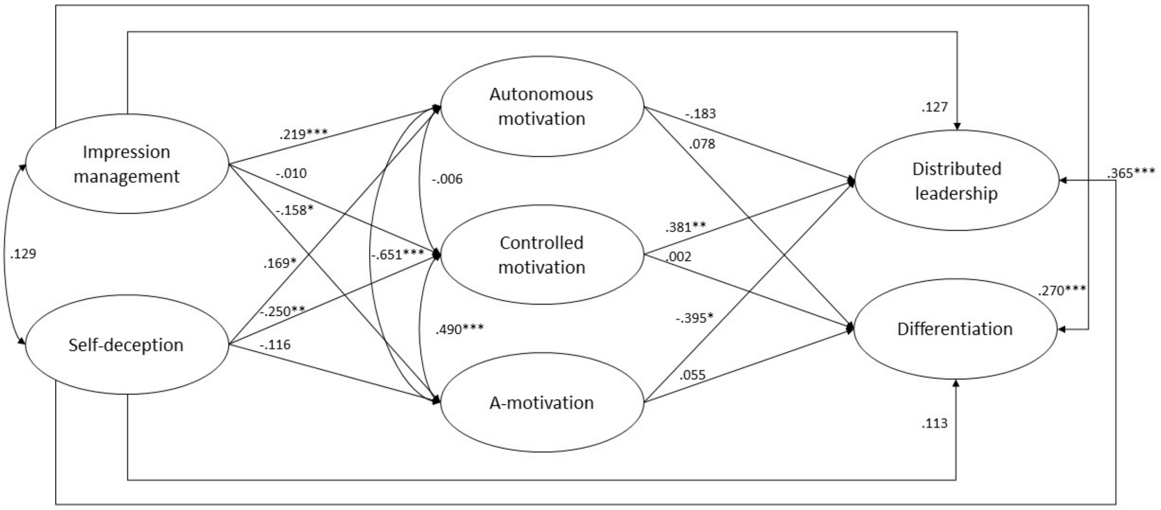
CONCEPT	MIN.	MAX.	MEAN	SD
Socially desirable responding				
Self-deception	1	5	3.50	.58
Impression management	1	5	3.70	.77
Motivation				
Autonomous motivation	1	5	3.42	.79
Controlled motivation	1	5	2.24	.76
A-motivation	1	5	2.33	.88
Distributed leadership	1	5	3.42	.92
Differentiation	1	5	3.73	.67

4.2 Explanatory results

This section reports on a path analysis, with both distributed leadership and differentiation as SSE variables of interest included. The path model presented in FigureFigure 1 has an acceptable fit with the data (CFI = .914; TLI = .905 RMSEA = .043; SRMR = .072).

4.2.1 Direct effects

The model illustrates that SDR has indeed an impact on the SSE variables of interest. Impression management has a significant positive effect on the respondents' self-reported differentiation in the classroom (estimate = .270; *p* < .001). So, respondents who score higher on their tendency to deliberately describe themselves in a more favourable way, provide a more positive picture of the extent to which they differentiate in their classroom. In contrast, impression management has no statistically significant effect on how respondents report on the extent of distributed leadership in their school (estimate = .127; *p* = .062). Impression management has an opposite effect on the variables in comparison to self-deception. Self-deception has a significantly positive effect (estimate = .365; *p* < .001) on respondents' reported distributed leadership in the school. This means that respondents, who have a higher tendency to – unconsciously – give an overly positive self-description, are found to have a more positive view on distributed leadership in their school. However, differentiation cannot be explained by self-deception (estimate = .113; *p* = .137).



* p < .05; ** p < .01; *** p < .001

Figure 1 Relationships between impression management, self-deception, autonomous motivation, controlled motivation, a-motivation and school self-evaluation scores based on the path analysis

The results of the path model also show a differential picture for the relationship between the different SSE variables of interest and the subscales of motivation. Autonomous motivation has no statistically significant effect on how respondents report on the extent of distributed leadership in their school (estimate = -.183; *p* = .145). This means that the extent to which respondents indicate that they see the administration of the SSE questionnaire as a personally valuable or interesting task, has no effect on how they report on distributed leadership in their school. Nor has this motivation dimension an effect on the reported amount of differentiation in the classroom (estimate = .078; *p* = .489). However, the extent to which respondents feel an external or internal pressure to fill in the SSE questionnaire does indeed have an effect. There is a statistically significant effect (estimate = .381; *p* = .002) between controlled motivation and the perception of distributed leadership. The parameter is positive, which means that the more respondents experience an external or internal pressure to fill in the SSE questionnaire, the more positively they report on distributed leadership in their school. With regard to the reported differentiation, no significant effect was found for controlled motivation (estimate = .002; *p* = .986). A-motivation, identifying whether respondents are motivated to fill in the SSE questionnaire anyway, has a negative relationship (estimate = -.395; *p* = .012) with distributed leadership. This would suggest that the more respondents are a-motivated, the more negatively they perceive distributed leadership in their school.

4.2.2 Indirect and total effects

The path analysis shows that there are not only direct effects on the SSE variables of interest. It is clear that SDR also has an indirect effect on the ultimate SSE measurements. As the model takes into account that the self-reported motivation scales could also be distorted by SDR, it is of interest to look at the extent to which they are. Autonomous motivation is affected both by impression management (estimate = .219; *p* < .001) and self-deception (estimate = .169; *p* = .011). The effect is positive which means that the higher respondents score on the SDR scales, the more they state to be motivated by sincere interest or see the questionnaire as a valuable means to achieve personal goals. Furthermore, self-deception is a predictor for the variance in controlled motivation (estimate = -.250; *p* = .002). The found relationship is negative, which means that the higher the respondents' tendency towards self-deception, the lower they score for controlled motivation. Impression management, in contrast, has no significant explanatory effect on controlled motivation (estimate = -.010; *p* = .883). With regard to a-motivation the opposite explanatory effects are found. The variance in a-motivation is not significantly impacted by self-deception (estimate = -.116; *p* = .093). Impression management, however, does indeed influence the a-motivation score in a negative way (estimate = -.158; *p* = .018). Therefore, respondents scoring higher on impression management tend to score lower for a-motivation, which in fact is a reducing effect on the extent to which they perceive filling in the SSE questionnaire as a useless task.

Structural equation modelling enables us to calculate the indirect effects and direct effects into total effect parameters of the adopted impression management and self-deception scale. The indirect effects, via the motivational scales, seem to have no statistical significant effect on the dependent variables (see Table 4). The total effects include both the direct effects, which were discussed above, and the indirect effects, which are due to the path structure in the analysis. Looking at the total effect of impression management on distributed leadership, a remarkable result pops up. Although there is no significant direct effect of impression management on distributed leadership, nor was a significant indirect effect found, the total effect of impression management on distributed leadership turns out to be statistically significant indeed. This means by taking both indirect and direct effects into consideration there is still an effect of impression management on the obtained result for distributed leadership. The total effects of self-deception on distributed leadership, and the total effect of impression management on differentiation could be expected, as the earlier results already point out that there is a strong direct effect between these variables. Still, this indicates that the indirect effects and direct effects are not levelling each other out. The total effect of self-deception on differentiation is, after including direct and indirect effects, not statistically significant.

Table 4 Standardized parameters and p-values for indirect effects and total effects of impression management and self-deception on SSE variables

	Parameter	*p*
Indirect effects		
Impression management → Distributed leadership	0.018	0.547
Self-deception → Distributed leadership	-0.081	0.072
Impression management → Differentiation	0.008	0.546
Self-deception → Differentiation	0.006	0.817
Total effects		
Impression management → Distributed leadership	0.145	0.029
Self-deception → Distributed leadership	0.284	0.000
Impression management → Differentiation	0.278	0.000
Self-deception → Differentiation	0.120	0.083

5 Conclusion and discussion

The objectives of this study are threefold. First, the need to identify how respondents differ regarding their tendency to respond in a socially desirable way, together with the quality and quantity of their motivation to fill in a school self-evaluation (SSE) questionnaire. Second, this study aims to identify to what extent SSE questionnaire data are affected by socially desirable responding (SDR). Third, the study explores to what extent the quantity and quality of respondents' motivation affects SSE questionnaire data.

Results show that there is indeed variation in respondents' tendency towards socially desirable responding. Respondents score more highly for impression management in comparison to self-deception. However, the spread among respondents is also higher for impression management than for self-deception. With regard to their motivation, respondents score rather low for autonomous motivation. Still, although rather on the lower end of the five-point Likert scale, a-motivation obtains a higher average score than controlled motivation.

Furthermore, this study shows that socially desirable responding has both direct and indirect effects on the SSE variables of interest. However, the picture for both dependent variables adopted in our model differs. Whereas there is a direct effect of self-deception on respondents' opinions about distributed leadership, there is none for respondents' self-reported extent of differentiation in the classroom. In contrast, there is a direct effect of impression management on respondents' self-reported extent of differentiation, but not for respondents' opinions about distributed leadership in the school. A significant direct effect of self-deception was found on distributed leadership. Literature describes impression management as a deliberate response behaviour operating for specific questionnaires or questions and as a temporary reaction (Paulhus, 2002). Our study, in the context of SSE, complies with earlier research suggesting that this mechanism indeed depends on the items' subject or the variable under review. The fact that the self-reported extent of differentiation is affected by impression management may be explained by how the respondent relates to the reported behaviour. The differentiation in the classroom is situated more in their own control and teachers may feel more responsible for it, whereas the extent to which their school is characterised by distributed leadership is not solely their own responsibility, nor a description of their exclusively own behaviour. This connects to literature that deals with the question of what could be understood as sensitive, and consequently vulnerable for SDR. Tourangeau et al. (2000) identify that concerns of possible consequences, or the perceived intrusiveness of questions, could trigger respondents' tendency towards SDR. Tapping into differentiation in the classroom might be experienced as more intrusive in comparison to distributed leadership, or respondents may think that they will be held accountable. The direct effect of self-deception on respondents' opinions about distributed leadership means that teachers tend to over-report characteristics at school-level in a genuinely or unconscious way. Possibly, they have an overly positive picture of their school or school-level characteristics because they may believe that they are doing a good job. They might consider all well-intended efforts of their colleagues and the management regarding distributed leadership and the schooling they provide in general, and have a genuinely positive perception about it. Further in-depth research should look into this phenomenon in order to uncover what is at play in this situation.

Indirect effects of impression management and self-deception via the path structure of the model are not significant. However, combining the direct and indirect effects into total effects, identifies that there is also a significant effect of impression management on the reported distributed leadership, although no significant direct and indirect effects were found. This stresses the importance of considering a path model approach as conducted in this study (Alwin & Hauser, 1975).

Motivation of respondents to fill in an SSE questionnaire has indeed, even after correction for the impact of SDR, an impact on an SSE. Results demonstrate that this also depends on the variable of interest. No impact of quantity or quality of motivation was found on the reported extent of differentiation. With regard to reported distributed leadership, motivation has an impact. In terms of quantity of motivation, this study points out that unmotivated respondents (scoring high for a-motivation) are evaluating distributed leadership in their school less positively. In terms of quality of motivation, this study finds that respondents who experience a pressure to fill in an SSE questionnaire, are attributing a higher score for distributed leadership. It remains unclear why the effect of motivation is different for differentiation and distributed leadership. Possibly, a-motivated respondents put less effort in thinking about positive examples of distributed leadership in their organisation, leading to a more negative picture of distributed leadership. Respondents who are reporting a higher extent of controlled motivation may feel an internal or external pressure to think of positive examples of distributed leadership, leading to a more positive score. Furthermore, there might also be a connection with the difficulty of the concept that is subject of the SSE. As distributed leadership is not common to occur among schools, it might be more difficult for respondents to retrieve positive examples or indications thereof. That could make the eventual score more liable to respondents' motivation to fill in the questionnaires. Making statements about differentiation, which is situated more closely to their daily activities, might require less effort from respondents to retrieve examples or indications. Possibly, this explains why motivation to fill in the questionnaire has no impact on the reported score for differentiation. Nonetheless, further research should look into possible explanations for these findings. This study sketches a more nuanced picture than is generally found in the field of self-report methodology, which commonly states that respondents' amount of motivation has an effect on the accuracy of their answers (e.g., Cannell et al., 1981; Kessler, Wittchen, Abelson, Zhao, & Stone, 2000). By demonstrating that discerning the quality next to the quantity of motivation, and identifying that there is a differential effect on the SSE variable of interest, the current study contributes to theory-building in this area of research.

The most important take-away of this study is that it must be acknowledged that data gathered within the process of SSE are not free of distortion. Respondents' self-perceptions or their perceptions of the school are indeed influenced by self-deception, or socially desirable responding in general. Also, motivation has an impact on SSE results. This raises the question to what extent SSE practitioners can rely on such questionnaire data in order to make sound conclusions, or indeed make valid policy decisions. Moreover, the impact of SDR and motivation is not univocal, and depends on the SSE variable of interest. Possibly, the extent to which the variables are under the control and responsibility of the respondents involved, makes it more or less vulnerable for influences of respondents' tendency towards impression management or self-deception. The same applies for respondents' motivation to engage in filling in the SSE questionnaire. These differential findings suggest that turning SSE results into valid interpretations is far from self-evident (Kane, 2013).

This study generates important insights about the conceptualisation of SDR and motivation. The factor analyses (exploratory and confirmatory) conducted in this study support the division of SDR into two subconcepts. Discerning impression management and self-deception is not only theoretically underpinned, but is also supported by the data. Moreover, it can be seen as a necessary approach since effects on the SSE variables are found in a differential fashion, both in a direct and in an indirect way. Conceptualising motivation into subconcepts as autonomous, controlled and a-motivation also proved to be important. Although no significant effects are found on the reported extent of differentiation, distributed leadership is affected by controlled and a-motivation. Autonomous motivation has no significant impact on either of the SSE variables. At the level of measurement, this study contributes to the field by exploring the concepts of SDR and motivation. This study makes a first attempt in translating the Paulhus Deception Scales into Dutch. This instrument makes use of 40 items, which requires a lot of effort from respondents. Further research could focus on psychometric qualities of this instrument and focus on the feasibility of shortening the questionnaire. Concerning motivation, this study was not able to discern a further subdivision of autonomous motivation into intrinsic motivation and identified regulation. Nor was there evidence to subdivide controlled motivation into introjected and external regulation. Possibly, the translation of the instrument into the context of filling in an SSE questionnaire needs further testing and refinement.

This study brings along important implications for researchers and practitioners that want to use SSE questionnaire data in order to inform their decisions, actions and policies. It is vital to avoid distortion in SSE questionnaires as much as possible. For SSE practice it could be advised to motivate in an autonomous way, meaning that respondents engage in filling in the SSE questionnaire out of sincere interest, or that they at least identify it as a means to achieve their personal goals. By doing so, the risk for distorted SSE results is reduced. Autonomous motivation can be stimulated by fostering feelings of autonomy among respondents. This can be enhanced by letting them decide on the focus of the SSE and developing an interest in the SSE by rousing respondents' curiosity in the matter (Hidi & Renninger, 2006). Also, our findings suggest that calling for respondents to be honest in their answering is not sufficient to obtain higher quality data. As self-deception occurs unconsciously, an awareness about their behaviour should be raised among SSE respondents. Triggering and stimulating respondents' critical thinking is an important aspect in the process of SSE. It is a central feature of self-evaluation capacity building (programs) (Labin, 2014). This could be stimulated by creating a safe climate among staff characterised by an openness for constructive critique and feedback (Vanhoof, Van Petegem, Verhoeven, & Buvens, 2009). Respondents can be asked to be hard for themselves when giving their opinion. Furthermore, practitioners are advised to supplement SSE data gathered from questionnaires with other data and data sources (MacBeath & McGlynn, 2002). Individual interview data or information obtained from focus group interviews could provide a deeper and/or broader insight in what could be derived from SSE questionnaires.

An SSE can be performed in varying contexts. When an SSE is performed in a context which is strongly characterised by accountability, respondents might behave in a very different way in comparison to a rather development oriented context. This study took place in a setting where teachers were familiar with the administration of this SSE within the framework of their personal development. An interesting extension to this study would be to examine how respondents behave in contexts that have a different focus (accountability versus development) and what effects occur on the SSE results.

To conclude, it is noteworthy that this study is unique within the field of school self-evaluation. It makes a first attempt to identify how socially desirable responding and the quality and quantity of respondents' motivation to fill in a questionnaire affects the quality of data. This study fits into a trend towards paying more attention to the quality of data that are gathered in the process of school self-evaluation, or where data are gathered to describe schools' own performance or functioning within the framework of quality assurance in general.

STUDY 4: ANSWERING IT OR SKIPPING IT: PREDICTING ITEM NONRESPONSE IN SCHOOL SELF-EVALUATION QUESTIONNAIRES.

This chapter is based on:

Faddar, J., Vanhoof, J., & De Maeyer, S. (submitted) Answering it or skipping it: Predicting item nonresponse in school self-evaluation questionnaires.

Abstract

Teachers are often asked to fill in questionnaires to provide information on the functioning of a school in the process of school self-evaluation (SSE). However, respondents are not always responding to all items. The phenomenon of item nonresponse, where respondents leave one item or a series of items unanswered, after which they continue completing the questionnaire, is not necessarily problematic as long as it occurs randomly. This study examines to what extent item nonresponse depends on respondents’ motivation to fill in the SSE questionnaire, their tendency towards socially desirable responding, the evaluation perspective on the SSE, the item design, and the construct which the item aims to tap into. This study reports on an experimental study with 376 respondents who completed a questionnaire in an authentic SSE setting. Cross classified multilevel modelling identified that respondents’ autonomous motivation predicts item nonresponse. Also, the evaluation perspective on SSE, item design, and the construct the item tends to capture all affect the probability of item nonresponse.

1 Problem statement

Teachers are considered to be highly eligible informants who have a lot of valuable insights into how their schools operate (MacBeath & McGlynn, 2002). This information is often gathered by means of questionnaires. Obviously, this implies that participants are supposed to be able and willing to provide information (Bateson, 1984). When respondents do complete a questionnaire, this is believed to generate valuable and important insights based on their day-to-day experiences in their school. One context in which the administration of questionnaires is used as a means to gather information is school self-evaluation (SSE), which in recent years has become a key strategy for internally evaluating a school's performance (McNamara & O'Hara, 2005; McNamara et al., 2011; OECD, 2013). SSE can be defined as a systematic process in which a school aims to describe and evaluate its functioning, with the purpose of shaping policies or undertaking actions within the framework of school development (Schildkamp, 2007; Vanhoof & Van Petegem, 2010). SSEs tend to gather information about processes taking place in schools, since it has been found that these can have the biggest effect school outcomes (Scheerens, 1991). Such processes can be situated at classroom level such as the quality of instruction (Scheerens, 2008). Next to the classroom level, processes can also be situated at the school level such as distributed leadership (Harris, 2004; Muijs et al., 2004). When schools perform self-evaluations, this is different to traditional self-reports in which individuals evaluate themselves, as an organisational entity as such cannot 'speak' for itself. Therefore, gathering information about a school's processes may be considered a methodological challenge. Schools need to rely on informants if their aim is to create a fair picture of its processes, meaning that the perception of teachers on the topics under review is vital.

However, respondents are not always answering all items in a questionnaire. The fact that 'no information is also information' is often overlooked. When one particular item or a series of items is not answered, after which a respondent restarts completing items in a questionnaire, this is referred to as item nonresponse (de Leeuw, 2002; de Leeuw et al., 2003). Unanswered items as such are not always a problem; it depends on why the items were left unanswered. Seminal literature discerns three different types of missingness: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (R. J. Little & Rubin, 2002). When a missing answer is due to missingness completely at random (MCAR), it means that the answer is lacking by accident (e.g., the respondent overlooked the item). When a value is missing at random (MAR), the missingness is related to an observed variable but not to the (unknown) value of the missing response itself. An example of this is when an elderly respondent does not answer a question because it was beyond recall; this is related to age and not to the item itself. In case of a refusal to provide an answer, for example, because the respondent finds the question too intrusive, the missingness is not at random (MNAR). When occurring not at random, this means that other variables than the ones being measured in the questionnaire are at play in the answering process. In this case, the missingness is not ignorable because it is a serious threat to distortion and bias (Rubin, 1976). Obviously, this is a serious problem for the quality of SSE results. Especially so as statistical techniques or strategies to overcome the problem of missingness most often have the hard assumption that missingness should occur completely at random. Moreover, such analyses require a large dataset, which is mostly not the case in an SSE context because school teams are not that large by nature. This makes it even more important to investigate to what extent item nonresponse occurs and identify determinants of item nonresponse.

The reason why respondents make the decision not to answer an item may be varied, and might depend on the characteristics of the respondent, the questionnaire design and the context (Groves & Couper, 1998; Lozar Manfreda, 2001; Vehovar et al., 2002). Literature already pointed to dispositions of respondents such as attitudes towards the completion of a specific survey, that affects respondents' intention to respond (Heerwegh & Loosveldt, 2009; Hox, de Leeuw, & Vorst, 1995). Interestingly, it has not been studied yet how the quantity and quality of respondents' motivation, as conceptualised in the Self-Determination Theory (Deci & Ryan, 1985, 2002), to respond questions can affect the occurrence of item nonresponse in SSE questionnaires. In addition, respondents may have a tendency to respond in a socially desirable way (Paulhus, 1984, 2002). This means that they want, consciously or subconsciously, to depict their school in a more favourable light. It could be that such a tendency leads to respondents skipping items in an SSE questionnaire.

Next, questionnaire design consists of many features such as the administration mode (e.g., online or on paper), but also the complexity of item wording could influence respondents' response behaviour (Belson, 1981; Fowler, 1992; Lenzner, 2012). In the context of SSE, questions can be formulated in such a way that respondents are required to give a statement about their school as a whole, or about themselves, after which items are aggregated onto the school level. It has not been examined yet what impact such a difference in item formulation can have on respondents' tendency to leave the item unanswered. Furthermore, nonresponse models indicate the survey topic as a predictor for nonresponse (Groves & Couper, 1998; Vehovar et al., 2002). It can be debated as to what extent different topics within one questionnaire could impede respondents' response behaviour. Research in the context of SSE has already demonstrated that different topics can be perceived as more difficult to respond to (Faddar, Vanhoof, & De Maeyer, 2017a).

Another important aspect is the context in which a questionnaire is administered, which can be rather particular for SSEs. Depending on the SSE's evaluation perspective, an SSE can be administered in a context oriented towards a school's own development, or with the aim to meet accountability requirements (Nisbet, 1988; Vanhoof & Van Petegem, 2007). It has not been studied yet whether the orientation of the SSE context might contribute to whether respondents leave items unanswered or not.

Reducing item nonresponse in order to obtain reliable and valid information is of utmost importance. Especially in a context where schools are increasingly required to monitor the quality of delivered education and make informed (policy) decisions by gathering evidence (Schildkamp et al., 2013). Up till now, there have been no studies performed that examine the impact of respondents' motivation, items' formulation and topic, and the SSE evaluation perspective on item nonresponse in the context of SSEs. Therefore, this study aims to explore predictors for why respondents are not sharing information they have about their school. The next research questions are in the focus of this study:

1. To what extent does item nonresponse occur in SSE questionnaires?
2. To what extent can item nonresponse be predicted by respondents' motivation, items' formulation and topic, and the SSE evaluation perspective?

2 Theoretical framework

2.1 Item nonresponse

Respondents can leave items unanswered in many ways. If he or she refuses to participate in the study or to complete any of the questionnaire, this is referred to as unit nonresponse

(Curtin, Presser, & Singer, 2005; Yan & Curtin, 2010). Consequently, there are no data available at all for that respondent. Item-missing data, however, comes from a situation in which although respondents did actually start to complete the questionnaire, data are missing for some items. A first possibility is that from a specific item onwards the respondent dropped out of the questionnaire and stopped completing it. This is referred to as partial nonresponse (de Leeuw, 2002; de Leeuw et al., 2003). The second possibility (and in the focus of this study) is called item nonresponse. This type of item-missing data is defined as the phenomenon where an item or a series of items is not answered, after which the respondent resumes completing the questionnaire (de Leeuw, 2002; de Leeuw et al., 2003). It concerns non-substantive responses that do not give any insight into the information the items are seeking. If the response option 'decline to answer' is provided, this is also looked upon as item nonresponse. As a result, no valuable information on the SSE variable of interest is obtained.

Theoretical insights in the participation of respondents in a survey in general are an interesting perspective for SSE questionnaires. Several models emphasise that respondent, context, and questionnaire design characteristics are elements of importance in predicting (non)response behaviour (Groves & Couper, 1998; Lozar Manfreda, 2001; Vehovar et al., 2002). The current study aims to further examine to what extent respondents' motivation, item formulation and topic, and the SSE evaluation perspective might influence the likelihood of leaving an item unanswered. The following paragraphs will elaborate on each of these elements, which are illustrated in Figure 1.

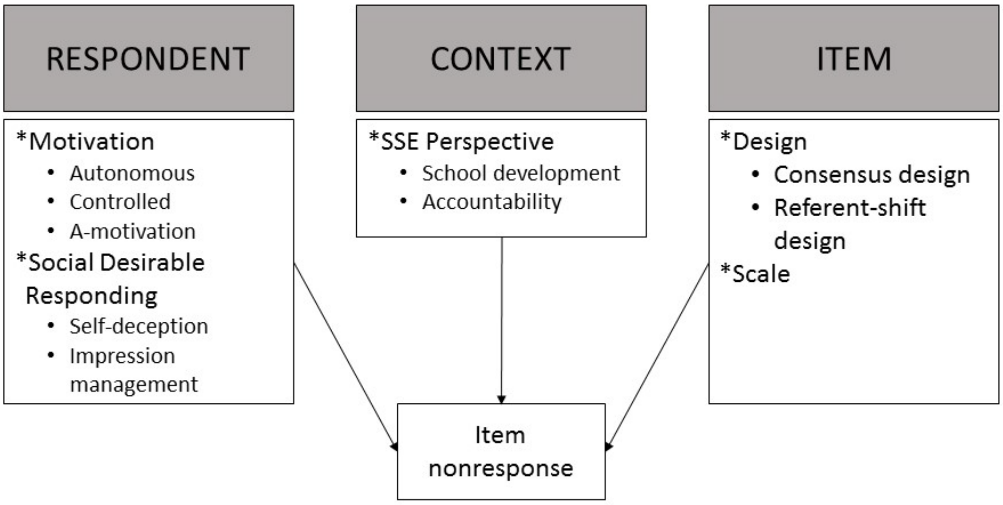


Figure 1 Conceptual framework on characteristics predicting item nonresponse in school self-evaluation

2.2 Respondent's motivation and tendency towards socially desirable responding

An important precondition to obtain information from a respondent is the willingness of a respondent to share her/his insights and to complete the questionnaire (Bateson, 1984; Cannell et al., 1981). Earlier survey research has been approaching the willingness to complete a questionnaire from a quantity perspective (e.g., Krosnick, 1991). Other work have examined the motivational aspect within the framework of the theory of reasoned or planned behaviour (e.g., Heerwegh & Loosveldt, 2009; Hox et al., 1995). This approach maps out a respondent's attitudes and subjective norms, which contribute to a respondent's intention to perform specific behaviour: or in this case, filling in the questionnaire. However, the quality aspect of motivation has, so far, been operationalised rather in a limited way. The current study addresses the willingness of respondents from a dual perspective, adopting both quantity and quality of motivation from the Self-Determination Theory (SDT) (Deci & Ryan, 2002; Vansteenkiste et al., 2009). Respondents are indeed required to reach a certain level of motivation to engage in the behaviour of filling in the questionnaire. When respondents do not want to take part in the questionnaire, they lack the necessary motivation. This, obviously, relates to the amount or *quantity* of motivation. When respondents do take the effort to fill in the questionnaire they have a certain level of motivation, but this gives no indication on why they do so, or the *quality* of motivation. Variation in the dimension of quality of motivation depends on the extent to which motives to engage in certain behaviour are internalised. Generally, three concepts are discerned that adopt both perspectives of quantity and quality of motivation (Deci & Ryan, 2002; R. M. Ryan & Deci, 2000).

The first concept, a-motivation, is characterised by a lack of motivation. Respondents do not see any value in engaging with the questionnaire. They may have no trust in a desired outcome or they may perceive it as a redundant task (Deci & Ryan, 1985). The second concept is autonomous motivation, which is characterised by a feeling of autonomy. Applied in the context of survey response, respondents experience a sense of self-determination in their decision to complete the questionnaire. They engage in the task because they find it sincerely interesting or an enjoyable activity (R. M. Ryan & Deci, 2000). Both of these motives are to a high extent internalised. Furthermore, the concept of autonomous motivation could be broken down into intrinsic motivation, with the highest amount of internalisation, and identification, where respondents complete a questionnaire because they believe that it helps to achieve personal goals. The third concept is controlled motivation where respondents feel a pressure to complete the questionnaire. According to where this pressure comes from, this type of motivation can be further split up. When the pressure to fill in the questionnaire comes from the self, manifesting in feelings of guilt or shame when they would not engage in the task, this is referred to as introjected regulation. When the feeling of pressure is external to the self, for instance a punishment when the task is not completed or a reward when it is, it is referred to as external regulation (Vansteenkiste et al., 2006). In the latter case respondents' motives are rather to a lower extent internalised.

Alongside to motivation, participants' tendency towards social desirable responding (SDR) is found to have an impact on their response behaviour and the quality of obtained data (Krosnick & Presser, 2010; Lam & Bengo, 2003; Wayne & Liden, 1995). The underlying idea is that people have the tendency to want to influence how others see them, and leaving items unanswered could be a strategy for respondents to depict themselves more favourably. They can do that by unconsciously deceiving themselves, which is referred to as self-deception (Paulhus, 1984, 2002). It is a phenomenon that occurs across different questionnaires and is therefore also called a response style. On the other hand, when a respondent tries to deliberately depict him- or herself more favourably, it is referred to in literature as impression management (Paulhus, 1984, 2002). This kind of behaviour can be in response to a particular questionnaire or even particular items and is considered to be a response set. During the process of answering items, it has already been demonstrated that these phenomena can cause bias in the information obtained from participants. Especially in a context which is dominated by an evaluative overtone, it can be expected that this mechanism will occur (Thomas & Kilmann, 1975).

2.3 The evaluation perspective

The implementation SSE can serve two different perspectives on evaluation; a *school development* perspective or an *accountability* perspective (Nisbet, 1988). Which of these perspectives an SSE focusses on is determined by what is evaluated and who sets the criteria for the evaluation. Within the school development perspective, the SSE is characterised by shared vision within the school for the (educational) aims being pursued. In this case, SSE focusses on processes and/or output defined by the school itself. Aims may include mapping out differences between management, staff or pupils to start a form of a dialogue within the school. The initiative to set up this kind of SSE lies within schools.

In an accountability perspective, by contrast, there is strong emphasis on the exercise of control; these usually investigate whether externally formulated objectives, mostly anchored in a legislative framework, have been achieved. There is a strong sense of uniformity enabling comparisons between schools and the evaluation is predominantly initiated by an external impetus.

In the context of SSE, there is not always a strict distinction between these two perspective. In contrary, SSE can serve different purposes that are intertwined with both perspectives. Therefore it seems to be more appropriate to think of both perspectives at each end of a continuum (Vanhoof & Van Petegem, 2007). The distinction between these two perspectives, in combination with the other parameters in the conceptual framework, is of particular interest in this study.

2.4 Item design

The focus of the SSE questionnaire is to a high degree dependent on what the initiator (i.e. the school itself) finds valuable or necessary. Often, SSEs tend to put variables under review that are related to a school's processes. The existing literature demonstrates that the enhancement of process indicators have the highest impact on school outcomes (Scheerens, 2008). Such processes can be situated at school level, of which distributed leadership or reflective capacity are typical examples. It is, however, a methodological challenge to map out such processes taking place at the organisational level. A traditional self-report is impossible since an organisation is not able to speak. Therefore, SSEs need to rely on the school's members to provide information about the school's functioning.

Teachers are considered to be highly eligible informants because of their day-to-day experiences in a school. It is believed that they have good insight in how their school operates (MacBeath & McGlynn, 2002). In order to make statements about the school as a whole, two ways of formulating items are mainstream. A first possibility is that teachers are asked to make a statement about themselves (e.g., I have a clear view on the job description of others in the school), after which the data are aggregated onto the school level. This is referred to as a consensus design (Chan, 1998; Chen et al., 2004). A second possibility is a referent-shift design, which is characterised by respondents making statements straight onto the school level (e.g., In this school everyone has a clear view on the job description of others in the school) (Chan, 1998; Chen et al., 2004).

It has already been demonstrated that different topics or scales and different designs of items generate a differential cognitive complexity for respondents (Belson, 1981; Faddar, Vanhoof, & De Maeyer, 2017b). On its turn, an increased cognitive complexity might influence a respondent’s response behaviour. Respondents can start putting less effort into performing all necessary cognitive tasks to arrive at an optimal response (Krosnick, 1991). Obviously, in a worst case, respondents can abort or not even start the intended answering process.

3 Methods

The following sections provide more insights into the setting in which the study is carried out, the design of the study, and what analyses were conducted.

3.1 Setting and participants

To examine the research questions put forward, this study must meet several requirements in its design. Firstly, it is vital that items be kept unchanged so that enough information per item is gathered. This also requires a context with a high number of participants that are completing the same items. Furthermore, in order to rule out variation at the school level regarding the processes under review, the participants should be making part of one organisation. Meeting these requirements enables to estimate models that identify variables at respondent and item level. Therefore, this study was embedded in an authentic SSE which was performed in a single, large educational organisation in Flanders. This organisation provides education and training in several disciplines, including general, technical and vocational education. Students can enrol from the age of 16, with a main target population of adults. The SSE was initiated by the organisation itself, where in a total of 378 teachers participated in the SSE, resulting in a participation rate of 62%.

The processes under review in the SSE were chosen by the organisation itself and focussed on the organisation’s functioning. In this study, we focus on the nonresponse for 26 items that try to capture two processes at organisational level: ‘distributed leadership’ and ‘reflective capacity’. These items were adopted from an instrument that aims to map out the policymaking capacity of schools (Vanhoof et al., 2011). The 26 items were situated in the middle of the SSE questionnaire, while the complete questionnaire consisted of 131 items.

3.2 An experimental approach

As we want to investigate the impact of the evaluation perspective of the SSE, we have set up an experimental design in which both evaluation perspectives (accountability vs. development) are experimentally manipulated. Respondents were shown an introduction which explained the purpose of the SSE, emphasizing the corresponding perspective. Both introductions can be found in Appendix 1. The introduction was altered for half of the respondents and the allocation of participants to each of the conditions was done randomly.

A second intervention in the study was introduced at the level of the questionnaire itself. Items that were tapping into the selected processes were formulated both in a consensus and a referent-shift. Each participant was asked to answer the two designs of each item. An example item can be found in Table 1.

Table 1 Example item for distributed leadership

	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, everyone has a clear view on the job description of others in the school.*

3.3 Analyses

In order to answer the research questions of this study, analyses should take into account two elements that relate to the data structure: first, the dependent variable is binomial (whether or not item nonresponse occurs for a particular item), and second, that the values for the dependent variable are nested within every respondent and every item. This complex cross-classified multilevel model (see Figure) needs to be considered when making calculations to answer the research questions. Therefore, analyses were performed with generalised linear mixed models by means of the R-package 'lme4' (Bates & Sarkar, 2008).

First, null models were calculated wherein only random effects were included. In order to determine the significance of each of the variance components, each of the variance components was excluded from the null model and the fit of the model with the data was verified. This was done by comparing the AIC and the likelihood ratio test. Based on these null models, we estimate the likelihood for item nonresponse to occur if a random item is answered by a random respondent.

We were also able to generalise the findings of our null model to the population. By taking into account the standard deviation for each variance component, we calculated 95% confidence intervals around the intercept of our null model. These confidence intervals represent the range in which the chance for item nonresponse lies in the population. These findings provide us with the necessary information to answer RQ1.

In a next stage, fixed effects were added to the model to predict the occurrence of item nonresponse. As the dependent variable in our model is binomial, analyses generated the chance of item nonresponse occurrence, expressed as a logit score. Based on these logit scores, probabilities were calculated. These results were used to answer RQ2, that aimed to identify whether each of these variables can predict item nonresponse.

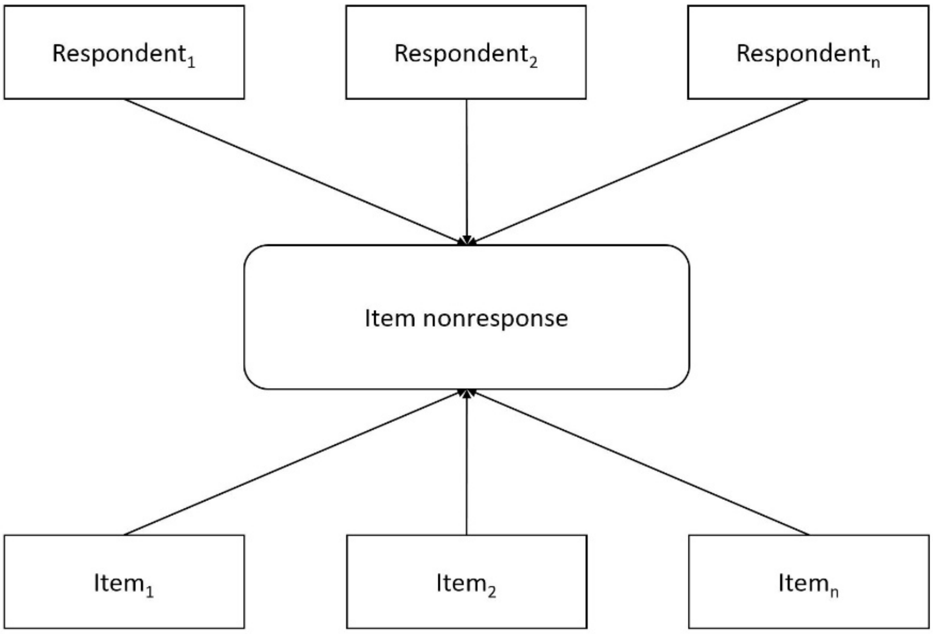


Figure 2 Cross-classified multilevel data structure

4 Results

In the following sections we first map out the occurrence of item nonresponse in the study for the processes of interest (RQ1). Afterwards, the predictive variables at respondent, item and context level are included in the advanced cross-classified multilevel analyses to predict item nonresponse (RQ2).

4.1 Random effects model

A random effects model sheds light on the extent to which variance components are statically significant in predicting item nonresponse. Results show that both variance components indeed predict item nonresponse in the SSE survey (see Table 2). Differences in respondents and items do predict item nonresponse in SSE surveys. This result justifies the addition of explanatory variables related to the level of respondents and items to accurately explain the occurrence of item nonresponse.

Table 2 Identification of variance components by comparing null models

	AIC	Loglikelihood	Δdf	Δ-2LL	p
Model 0	6040.50	-3017.20			
Model 0a (respondents out)	10532.50	-5264.30	1	-4494.00	<0.001
Model 0b (items out)	6521.70	-3258.80	1	-483.16	<0.001

A closer look at the baseline null model, including both random effects, gives information on the chance that item nonresponse will occur. The intercept represents the probability of item nonresponse when a random item is presented to a random participant, and is estimated at 4.35% (see Table 3). This finding points to a rather low probability of item nonresponse.

Taking into account the standard deviation of each variance component, the confidence intervals indicate the range in which the chance for item nonresponse lies for 95% of the population (see Table 3). If 95% of the respondents in the population are asked to complete a random item, the chance for item nonresponse lies between 0% and 98%. If 95% of all possible items are presented to a random respondent, the chance for item nonresponse lies between 7.33% and 22.05%.

Table 3 Null model for item nonresponse

	Est. (logit)	Std.Err.	Prob. (%)
Intercept	-3.09***	.29	4.35

			95% confidence intervals	
Variance components	Variance	Std. Dev.	Min. Prob. (%)	Max. Prob. (%)
Respondents	13.67	3.70	.00	98.47
Items	0.86	0.93	7.33	22.05

4.2 Fixed effects model

This section discusses the fixed effects model (see Table 4), which includes all the predicting variables. The results of this full model show that the central components of the conceptual framework (the respondent, evaluation perspective and questionnaire design) play a role in the occurrence of item nonresponse, yet, not all predicting variables are of statistical significance.

The intercept of this model represents the chance for item nonresponse when a random item is processed by a random respondent, and is scoring zero for all other adopted variables. The latter means that respondents are scoring on average for respondent characteristics, all types of motivation and their tendency towards socially desirable responding. Regarding the evaluation context, this means that the item is administered in the development condition. At the level of item characteristics it involves an item in a consensus design, related to the topic of reflective capacity. In this instance, the intercept results in a probability of .42 (see Table 4), which means that there is a small chance for item nonresponse.

Table 4 Predicting model for item nonresponse

	Predicted Model		
	Est. (logit)	Std.Err.	Prob. (%)
Intercept	-5.47***	.41	.42
Respondent characteristics			
Autonomous motivation (Zscore)	-.69*	.28	.21
Controlled motivation (Zscore)	-.24	.26	.33
A-motivation (Zscore)	.24	.30	.54
Impression management (Zscore)	.10	.24	.46
Self-deception (Zscore)	-.15	.24	.36
Evaluation perspective			
Accountability vs. Development [ref.cat.]	1.06*	.47	1.19
Item characteristics			
Design (Referent-shift design vs. Consensus design [ref.cat.])	1.95***	.11	2.86
Scale (Distributed leadership vs. Reflective capacity [ref.cat.])	.93***	.24	1.05

With regard to respondent characteristics, results show that only autonomous motivation has a statistically significant effect. A respondent who scores higher for autonomous motivation compared to the average is less likely to leave a random item unanswered. More precisely, an increase of one standard deviation leads to a decrease of .69 logits for item nonresponse, resulting in a probability for item nonresponse of .21%. All other characteristics of respondents adopted in our model do not statistically significantly predict the occurrence of item nonresponse. Controlled motivation has no effect on the occurrence of item nonresponse, and although there is a positive relationship between a-motivation and the chance of item nonresponse occurring, the relationship was not found to be statistically significant. Neither impression management or self-deception as subconcepts of SDR were found to be statistically significant.

The evaluation perspective (a development- versus an accountability-oriented context) has a significant direct effect on the likelihood for item nonresponse to occur. Items administered in an accountability context have an increased chance for item nonresponse of 1.06 logits, resulting in a probability of 1.19 %.

With regard to item characteristics, the results show a statistically significant effect for both item design and the scale of which it is part. For design, the model identifies an increase of the chance for item nonresponse to occur with 1.95 logits, leading to a probability of 2.86%. In addition, the scale to which an item belongs is of significance. When all parameters are set to zero, and scale is the only varying factor, the results demonstrate that items on the distributed leadership scale have a higher chance of being left unanswered. The likelihood increases with .93 logits, resulting in a probability of 1.05 %.

5 Illustrations

The predictive model allows to calculate the chance for item nonresponse to occur in some hypothetical cases. The next section illustrates the findings of this study with some specific examples. The calculations mentioned in the particular cases discussed here are also graphically presented in Figure .

5.1 The case of Samir

Samir is filling in his SSE questionnaire with a high autonomous motivation. He is sincerely interested in the questionnaire, or at least he is convinced of the fact that administration of the questionnaire advantageous in achieving his goal of improving his teaching skills. The SSE questionnaire is administered in the school with the idea of obtaining different views of staff on school-specific quality goals. The question Samir tends to answer aims for a statement about his individual behaviour and is about distributed leadership in the school.

> Samir's item: "I have sufficient possibilities to engage in the decision-making processes at our school."

This means that Samir scores one standard deviation higher than average for autonomous motivation. The evaluation perspective with which the SSE questionnaire is administered at his school is developmental. The item is formulated in a consensus design and focusses on distributed leadership.

According to our model, the likelihood that this item will be left unanswered is .21 %.

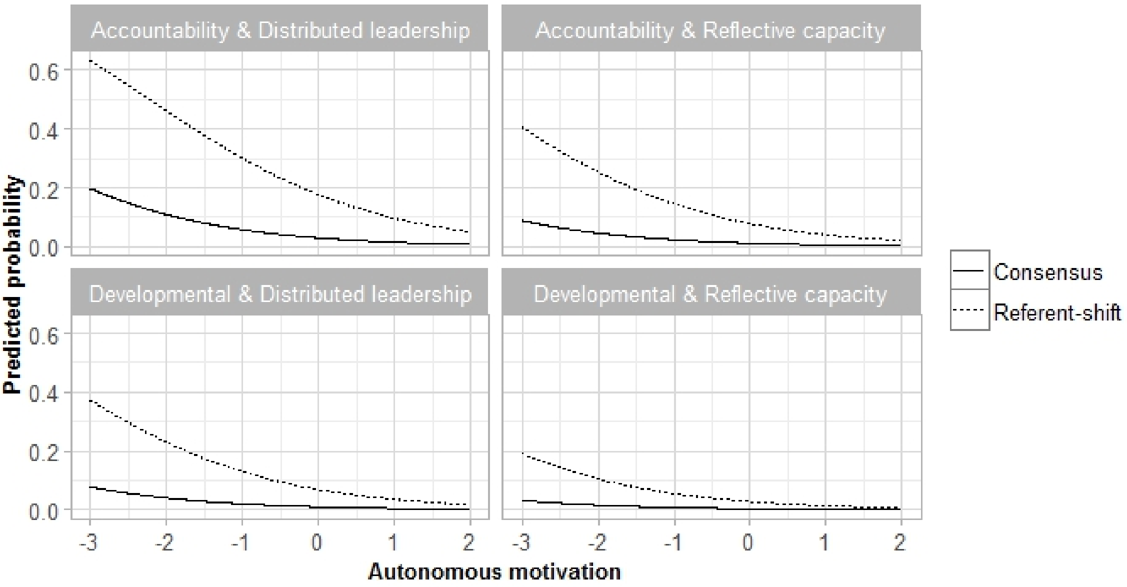


Figure 3 Probabilities for Item nonresponse based on the predictive model

5.2 The case of Erin

Erin does not particularly score high for autonomous motivation. She is not really convinced of filling in the questionnaire as being very interesting, or valuable for her personal goals. The principal of her school has the intention to compare the school's scores with other schools, and wants to track down how the school is performing regarding the quality objectives prescribed by the Ministry of Education. Erin's question is formulated in such a way that it requires a statement about the school as a whole and focuses on the school's reflective capacity.

> Erin's item: "In this school everyone has a critical attitude regarding their own performance."

In terms of our model this means that Erin scores one standard deviation lower than average for autonomous motivation. The SSE questionnaire is administered from an accountability perspective on the SSE process. The item is formulated in a referent-shift design and is related to reflective capacity.

Calculations based on our model shows that there is a 14.52% likelihood that this items will be left unanswered.

5.3 The case of Luke

Luke does not think of the SSE questionnaire of being interesting or valuable at all. Actually, he does not see how it can contribute to his personal goals. In Luke's school the school self-evaluation questionnaire has been introduced to examine how their school is performing compared to other schools in the neighbourhood. The item Luke is asked to answer requires a statement about the school as a whole, and aims to collect information on the extent to which the school is characterised by distributed leadership.

> Luke's item: "In this school everyone has sufficient possibilities to engage in decision-making processes."

When this situation is translated in the modelled variables, Luke is scoring minus two standard deviations for autonomous motivation compared to the average. The item is administered in an accountability perspective on SSE processes. The item design is formulated in a referent-shift design, and relates to the concept of distributed leadership.

In this case, the chance for the item to be left unanswered increases to 63.19%.

6 Conclusion & Discussion

This study addresses the issue of item nonresponse in school self-evaluation (SSE) questionnaires. Item nonresponse is defined in this study as a phenomenon where an item or a series of items are left unanswered, after which the subsequent items are filled in. Too often, it is readily assumed that items are left unanswered randomly. In this study we explored to what extent item nonresponse occurs, and to what extent characteristics of respondents, items and the evaluation perspective in which the SSE questionnaire is administered, or an interplay of these, predict the occurrence of item nonresponse.

The results of this study suggest that item nonresponse can be a big concern in the field of SSE. Although item nonresponse was found to occur in a rather limited extent in this study, the results demonstrate that a combination of variables can turn item nonresponse into a problematic issue. The chance for item nonresponse for a random item processed by a random respondent is 4.35%. It is found that differences among respondents in the chance for skipping items is much higher compared to differences among items for them to be skipped by respondents. Across 95% of all respondents in the population, a random item has a probability between 0.00% and 98.47% of being left unanswered. For 95% of all items in the population, the chance for item nonresponse lies between 7.33% and 22.05%.

The predictive analyses in this study indicate that characteristics of respondents, items and the context in which the SSE is conducted all play a role in the occurrence of item nonresponse. The extent to which respondents are autonomously motivated is significantly related with item nonresponse. The more a respondent is autonomously motivated, the less likely item nonresponse is to occur. Controlled motivation has no impact on the occurrence of item nonresponse. This means that the extent to which respondents are motivated to fill in the SSE questionnaire by internal or external pressure, has no effect on their behaviour in regard to skipping particular items. Looking at the impact of socially desirable responding, both the impression management and self-deception component appear to have no effect on the occurrence of item nonresponse. This finding seems to suggest that respondents, although they may score high for SDR, still want to provide a plausible answer rather than just skip the item.

The context in which an SSE questionnaire is administered was found to be critical in the occurrence of item nonresponse. A context characterised by an emphasis on accountability leads to a higher probability that item nonresponse will occur. Furthermore, the design of the item and the scale in which it originates are significant in predicting item nonresponse. Items that require a statement about the school as a whole are more likely to be skipped. This finding is connected to earlier research on the use of questionnaires in general and in the context of SSE in particular (Faddar et al., 2017a). Referent-shift design items are perceived by respondents as more difficult to answer. Making statements about their colleagues or their school is more difficult: respondents might feel not well-informed on the matter, or if they are, they may have difficulty in selecting a response option that reflects their judgement about the particular item. It is also found that the process indicator under review impacts the likelihood for item nonresponse. Items tapping into the concept of distributed leadership have a higher chance of being skipped, compared to those related to reflective capacity. It is possible that items related to distributed leadership are perceived as more difficult for the respondents to answer. This could be due to the usage of more abstract or unknown words in the items, which may influence the cognitive processing of an item, and influence the likelihood of item nonresponse to occur (Belson, 1981; de Leeuw et al., 2003; Faddar et al., 2017a; Lenzner, 2012).

Results of the current study point to important characteristics of respondents, administered items, and the context in which an SSE is conducted. The use of questionnaires in the context of SSEs can be very valuable, but the completion of items needs to be stimulated in order to avoid distorted results (de Leeuw et al., 2003). Based on our findings, it can be concluded that respondents need to be highly motivated in an autonomous way in order to avoid item nonresponse. Autonomous motivation can be stimulated by fostering feelings of autonomy among respondents. This can be enhanced by letting them decide on the focus of the SSE, enabling them to develop an interest in the SSE by rousing respondents' curiosity in the matter, and feeding back the SSE results (Hidi & Renninger, 2006). This study also suggests carrying out SSEs in a developmental context, in order to obtain as much information as possible from respondents. This may relate to the experience of a safe climate, in which respondents can be open about their judgement of the SSE process under review (MacBeath, 1999; Vanhoof et al., 2009). As soon as SSE becomes mandatory and the stakes of the SSE become higher, different effects may introduce distortion in the data (Swaffield & MacBeath, 2005). Results also demonstrate that it is key to keep SSE items as easy as possible, so that respondents are not cognitively overloaded. An extensive and thorough pretesting of the SSE questionnaire is vital (de Leeuw et al., 2003). These findings regarding item formulation and item topic are particularly valuable for researchers in the field of SSE or for researchers interested in mapping out processes at school or organisational level. Developers of SSE instruments can also benefit from these insights.

The current study extends the current knowledge base on the occurrence of item nonresponse. The inclusion of specific characteristics of respondents, items and the evaluation perspective on the SSE context as predictor variables for item nonresponse had not been examined in this regard prior to this study. Existing survey literature often approaches the motivational aspect of respondents from a quantitative perspective (e.g., Krosnick, 1991). The idea of a quality oriented view on motivation, within the framework of the Theory of Planned Behaviour (Ajzen, 1991), is often operationalised in a limited way (e.g., Heerwegh & Loosveldt, 2009). Adopting the perspective of both quantity and quality of motivation, as conceptualised within the framework of the Self-Determination Theory (Deci & Ryan, 2002), in order to study item nonresponse had not yet been carried out in this context. The results of this study found that a differential effect for both dimensions justify the distinction made in the conceptualisation of motivation in this respect, and has proven to be a valuable addition to the existing body of literature.

The context in which the current study was performed, was valuable to obtain real and authentic SSE results. However, this meant that only a limited number of interventions could be made in the study. Further research could focus on testing more interventions at the item level in an experimental context. For instance, although the tested items were situated in the middle of the questionnaire in this study, it could be tested whether the placement of items in the questionnaire impacts the chance for item nonresponse. Because respondents can get tired of completing a questionnaire, items at the end of a questionnaire may have a higher likelihood of being left unanswered (Herzog & Bachman, 1981). It could be argued that not only the place in a questionnaire, but also what questions preceded a particular item, could play a role in the context of SSE as well (McFarland, 1981).

To the authors' knowledge, this is the first study that addresses the issue of item nonresponse in the context of SSE. Results demonstrate that although item nonresponse occurs in SSE questionnaire results, this is to a limited extent. However, results also show that in a situation with some negative characteristics, the chance for item nonresponse can increase as high as 63.19%, which is problematic and can yield distorted SSE results.

Appendix 1 Introduction to the questionnaire

This Appendix contains the introduction that was displayed to respondents when accessing the online SSE questionnaire.

Accountability-oriented perspective on SSE	Development-oriented perspective on SSE
The **government** expects *School X* to gather information among and about its staff. We will **evaluate** your perceptions in order to be able to prove that we **meet requirements** and **achieve the minimum objectives**. With these data we will be able to **give account of** our functioning **towards the authorities**. This information will also enable us to **measure and make comparisons** with other organisations in order to **steer** our organisation in the right direction.	In the framework of our **own quality assurance** *school X* has chosen to carry out a self-evaluation. It is valuable to fill in a self-evaluation to **improve ourselves**. Your opinion enables the school to start a **maximal dialogue**. From your input, *school X* aims to draw lessons in order to **develop** ourselves in **all transparency**. **Different** opinions lead to different approaches, and this is what we **dedicate ourselves to** for a 100%. The information will be used **internally** in order to create a **common vision on quality of instruction**. The topics addressed are **chosen by and in agreement with** the teachers themselves.

CONCLUSION AND DISCUSSION

This dissertation addresses the methodological and psychometric underpinning of school self-evaluation (SSE) results. More specifically it aims to examine the validity of SSE results. First, this concluding chapter outlines the rationale of the research aim and how this is broken down in four studies. Next, the main findings from this dissertation are synthesised and critically discussed. Finally, this chapter considers the implications of the findings, discusses limitations of the conducted research and includes the main lessons from the research.

1 Rationale and research aims

Schools are increasingly expected to monitor their delivered quality by themselves. SSE, as a mechanism for internal evaluation, is a key strategy for schools to meet this requirement and has become common practice in many education systems (Eurydice, 2015; McNamara et al., 2011; Nelson et al., 2015; OECD, 2013; Schildkamp et al., 2012). In this dissertation, SSE is defined as a systematic process, in large part initiated by the school itself, whereby 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 & Van Petegem, 2010). An SSE can address many different topics, but focussing on a school's processes can, as it is argued, lead to a higher impact on the enhancement of school effectiveness and improvement as these indicators can be manipulated more easily (Scheerens, 1991; Van Petegem, 1998). However, tapping into school processes such as quality of instruction or school leadership is a methodological challenge since schools cannot directly communicate a current state of affairs regarding these indicators. As the aforementioned definition outlines, SSEs often rely on informants' perceptions to generate a picture of a school's functioning (see Figure). To this end, although several other valuable strategies can be used, the administration of questionnaires among staff members can deliver insightful information (e.g., Hendriks et al., 2001; MacBeath & McGlynn, 2002; Vanhoof et al., 2011). It is a method that enables the collection of information in a limited amount of time among a higher number of participants (Cohen et al., 2011). Different instrument developers have created a wide array of questionnaires. However, the quality of these instruments has been questioned and it has already been argued that in some cases there is a lack of methodological and psychometric underpinning (Hendriks, 2000; Hofman et al., 2005). While the value of respondents' perceptions is acknowledged, the extent to which these perceptions are distorted is unclear. A self-perception might (intentionally or unintentionally) be far removed from how a school is actually performing (Alwin, 2010; Paulhus, 2002). A self-perception could, for instance, easily turn into self-deception, where a respondent unconsciously depicts his/her school in a more favourable way. Distorted results are a serious threat to the aim of, in this case, arriving at a description of a school's functioning and from which valid conclusions are supposed to be drawn (Kane, 2013; Meier & O'Toole, 2013). Up until now, little empirical evidence has been examined to verify the validity of SSE results. As a central research aim, this dissertation aims to fill this gap and examines the validity of SSE results.

Considering an argument-based approach to validity (Kane, 2006, 2013), a crucial stage in building a strong interpretive argument is the scoring stage in which respondents answer questions (see Figure) (Alwin, 2010; Alwin & Krosnick, 1991; Tourangeau, 2003). It is readily assumed that respondents cognitively process items as the instrument developers intended them to be, which is referred to as cognitive validity (Karabenick et al., 2007; Koskey et al., 2010). However, no empirical evidence has been collected in the context of SSE that supports this assumption. Furthermore, little is known about the extent to which respondents' perceptions about school processes are distorted by the motivation of respondents to fill in a questionnaire and their tendency towards socially desirable responding (SDR). Finally, it is assumed that respondents only leave an item or a series of items unanswered, after which they restart completing the questionnaire, which is referred to as item nonresponse, on a random basis. This dissertation addresses these assumptions and checks whether they hold in the context of SSE. The examination of these assumptions is broken down in four studies. Each study has a different research goal by altering central concepts.

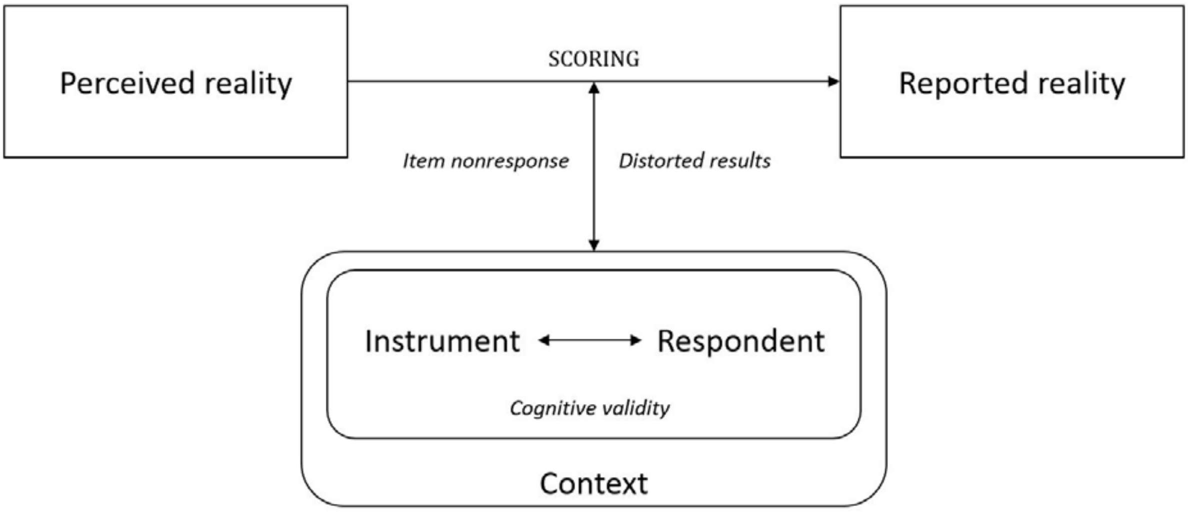


Figure 1 Conceptual model

The first and second study focussed on the respondents' thinking process while answering an SSE questionnaire and examined whether or not respondents think about what the instrument developers intended to map out with the items developed. This cognitive process consists of an interpretation, elaboration and response stage. Furthermore, they examined problems that could be identified during the cognitive stages of respondents' answering process of SSE items.

The third study aimed to identify the extent to which SSE results are affected by respondents' motivation to fill in an SSE and their tendency towards SDR. The study addressed the following questions: are results distorted by the level of respondents' autonomous, controlled or a-motivation? And, does the extent to which respondents are characterised by impression management and self-deception have a distorting effect on SSE results?

The fourth study addressed the phenomenon of item nonresponse in the context of SSEs and questioned whether respondents actually participate in the questionnaire by completing all the questions included. This study focused on the extent to which an item or a series of items were skipped by a respondent, after which he or she restarted completing the SSE questionnaire. The research goal of this study was to identify the extent to which item nonresponse occurs in SSE results, and to what extent it can be predicted by respondent, instrument and context characteristics.

2 Conclusion & discussion

This section makes a synthesis of the main findings drawn from the four studies in this dissertation and discusses these findings critically.

2.1 Cognitive validity of SSE results cannot be readily assumed

By means of the cognitive interviewing technique, respondents' cognitive process was mapped out while answering an SSE questionnaire. Drawing on a body of cognitive interviews with 20 primary teachers and principals from four Flemish primary schools, the first study in this dissertation determined the extent to which SSE results are cognitively valid. Based on cognitive validity criteria for each of the three main cognitive tasks (interpretation, elaboration and response) that are supposed to be carried out by a respondent, the level of cognitive validity was rated for 1200 units of analysis.

SSE RESULTS ARE ONLY COGNITIVELY VALID TO A LIMITED EXTENT

The findings of this study point to a serious problem and show that there is a rather large discrepancy between how instrument developers intended items and how respondents actually cognitively process them. So, cognitive validity of SSE results cannot be readily assumed. Not even half of respondents' interpretations of items were found to be cognitively valid. When looking at the elaboration stage, the extent of cognitive validity dropped even further. In this stage, respondents search their memory for relevant information. Less than a third of the elaborations analysed were found to be cognitively valid. Far more positive was respondents' use of the predefined answering options. More than 90% of the responses were found to be cognitively valid. The study also demonstrated that the interpretation stage is a crucial one. When respondents succeed in making a cognitively valid interpretation, it is likely that the consecutive tasks are also processed in a cognitively valid way.

When predicting the extent of cognitive validity of the interpretation stage, only differences between items mattered, while differences between respondents did not. When predicting the extent of cognitive validity for the elaboration stage, differences among both items and respondents were found to have a statistically significant effect. Differences between respondents only mattered in the elaboration stage, which could be explained by the fact that at this point respondents integrate their personal experiences and mental models in order to answer the item (Karabenick et al., 2007; Schwarz, 2007). However, more research is needed to identify what particular characteristics of respondents matter. Next, the findings point to the vital role of item characteristics for both the interpretation and elaboration stage. This connects with earlier research, which points to the effect of item design on the quality of obtained results (Knäuper et al., 1997; Krosnick & Presser, 2010; Schwarz, 1999). Moreover, a more extended analysis also showed that items which require a statement about the school as a whole, referred to as referent-shift items, are more likely to be elaborated on in an cognitively invalid way compared to consensus design items, which only require a statement about the self. Instrument developers need to take these repercussions regarding cognitive validity into account when developing items.

Based on these findings it is concluded that answering SSE questionnaire items is a quite demanding task for respondents. The assumption that SSE respondents cognitively process items as instrument developers intend them to be does not hold. Respondents only reported on what was aimed for by the instrument developers to a limited extent. In addition, the assumption that the formulation of items (in a referent-shift design or consensus design) does not matter does not hold either. All these findings raise the question as to what problems occur in respondents' cognitive processes that can explain the cognitively invalid results.

DIFFERENT PROBLEMS LEAD TO COGNITIVELY INVALID SSE RESULTS

By means of a content analysis on the body of cognitive interviews from the first study, the second study in this dissertation examined the problems that lead to cognitively invalid answers to SSE items. For each of the cognitive tasks, problems arose in respondents' answering process. Results of this study showed that respondents have difficulties in interpreting some specific terms. In particular, some linguistic aspects were found to generate problems in respondents' process of giving meaning to items and interpreting them. For instance, respondents struggled with the use of rather unfamiliar or abstract words as well as with the sentence structure of an item.

Furthermore, when respondents were elaborating on items (i.e. searching their memory for relevant information), some issues were found. Some respondents experienced difficulties staying on topic; their thoughts strayed and they considered information that was irrelevant.

Others, in contrast, did not cover the broad sense of an item and subsequently mentioned irrelevant information. It was also found that respondents were mistaken about the appropriate timeframe and did not think of a current state of affairs but brought up out-dated information. Furthermore, results showed that respondents were mistaken about whom a statement was required. It happened that respondents only considered themselves when a statement about the school as a whole was required. This endangered the rationale behind the use of referent-shift items.

This study also identified some problems regarding the response stage. Respondents sometimes lacked an answering option that reflected their mental judgement. For instance, when the statement applied for half the school team but not for the other half, they experienced an inability to reflect this in a single answer option. Also, the intention of a predefined 'don't know' option, provided for respondents who had no relevant information, was violated in different ways. Some respondents were found to use this option due to an item's complexity, simply because they did not know what the item was asking about. Further research might focus on what the effect of an answering option 'I don't understand' might be on the yielded responses, and whether this would be used as intended.

ASKING TEACHERS' PERCEPTIONS ABOUT SCHOOL PROCESSES IS A DEMANDING TASK

The aim of asking teachers' perception about school processes is a powerful idea. Within the field of SSE, it is argued that teachers are highly eligible participants because of their day-to-day experiences in the school. They are supposed to have valuable information about the functioning of the school (MacBeath, 1999; MacBeath & McGlynn, 2002). Furthermore, in light of school development, the involvement of staff throughout the SSE process, can also be a driver for change in schools (Fullan, 2007). However, the findings in this dissertation point to a serious drawback of asking respondents' perception by means of an SSE questionnaire. Not even half of the items that were cognitively processed by our SSE respondents were cognitively valid. Even though teachers are highly trained professionals that are supposed to be familiar with an educational vocabulary, it was demonstrated that they experienced difficulties in understanding what is meant by general terms as part of an item such as 'vision' or 'data'. Moreover, asking SSE participants to complete items that aim to map out school processes is far from self-evident. Since teachers are only to a limited extent able to provide perceptions that are asked for in regard to the topic under review, it could be argued whether or not it is indeed a valuable approach to ask for their perception? The findings from the cognitive validity studies in this dissertation seem to suggest that not all teachers are such highly eligible participants to fill in an SSE questionnaire as has been argued. There seem to be teachers that can succeed in providing cognitively valid information, in contrast to others. Based on the findings, it could indeed be questioned to what extent results from SSE questionnaires provide valuable insights to the users of the SSE results. This may be an argument for working with a selection of teachers participating in an SSE to make results more cognitively valid. However, this raises subsequent questions such as which teachers to select and on what basis.

Particularly difficult were the items that probed for respondents' perceptions about the school as a whole. The results in the first two studies of this dissertation pointed to a problem when concepts were measured by means of items that required a statement from respondents about the school as a whole. It can be argued that such referent-shift items have the advantage of better capturing processes in an organisation as they directly address the organisational level (Kirkman, Tesluk, & Rosen, 2001). However, the studies in this dissertation showed that respondents experienced even more problems with such items, as these had a higher probability of being processed in a cognitively invalid way. It was found that SSE respondents experienced difficulties in taking a helicopter view and tended to think about themselves instead of the school as a whole. It seemed more natural for respondents to answer about themselves rather than about the school as a whole, although the item explicitly asked for it. This advocates the use of consensus design items that only require respondents to make statements about themselves.

COGNITIVE VALIDITY AS A THREAT TO THE VALIDITY OF INTERPRETATIONS AND USE OF SSE RESULTS

Considering an argument-based approach to validity, the consecutive stages in building an interpretive argument about the interpretation or use of SSE results are endangered (Kane, 2006, 2013). If respondents are interpreting questionnaire items in many different ways, this complicates the conclusions that are drawn from the results. The question arises as to how users of SSE results are able to arrive at sound conclusions if respondents hardly respond to what was asked for. This is a serious threat to the validity of interpretations and the use of SSE results. In order to align respondents' cognitive processes with the intentions of instrument developers, it might be helpful to install a stage of sense-making in which a common understanding among participants is created about what the items are aiming to capture. This could also anticipate the problems that were identified in the second study. While all respondents in an SSE can have their own view on the topic under review in an SSE, creating such a common understanding would enable respondents to communicate in a common language about what is in focus.

Although these current findings regarding cognitive validity are embedded in an SSE context, the use of questionnaires is widespread in other fields and disciplines with the aim of mapping school processes or, beyond the education context, organisational processes. The way in which items are formulated and questionnaires are set out is similar, for instance, when measuring employees' commitment to the organisation in which they work (Allen & Meyer, 1990; Goffin & Gellatly, 2001). The methodological challenge in measuring latent constructs remain. Moreover, the risk in a research context is that researchers often do not have the opportunity to install a stage of sense-making. In such a case, the validity of the interpretation and use of questionnaire results is even more endangered.

2.2 SSE results are not free from distortion

The results of the third study in this dissertation pointed out that we cannot readily assume that respondents' perceptions are not distorted by their motivation to fill in an SSE questionnaire and their tendency towards socially desirable responding (SDR). This study drew on an authentic SSE conducted in a Flemish educational organisation in which 376 teachers completed an SSE questionnaire. By means of a path model, within a structural equation modelling framework, the effect of respondents' motivation to fill in an SSE questionnaire and their tendency towards SDR on perceptions about distributed leadership and differentiation in the classroom was examined.

BOTH QUANTITY AND QUALITY OF RESPONDENTS' MOTIVATION HAVE AN EFFECT ON SSE RESULTS

Respondents' motivation was found to impact upon SSE results, even after correction for the impact of SDR in respondents' reporting on their motivation. The findings identified a statistically significant negative effect of a-motivation on the perceptions about distributed leadership. This means that the more a respondent did not see the point in filling in an SSE questionnaire, the more negatively he/she reported on distributed leadership. Furthermore, a significant positive effect was found of controlled motivation on the perception regarding distributed leadership. The more a respondent felt an internal or external pressure to fill in an SSE questionnaire, the more positive he/she reported on distributed leadership in their school. Interestingly, the extent to which respondents reported being autonomously motivated was found to have no effect on their reported perception on distributed leadership. With regard to differentiation in the classroom, the reported perceptions were not impacted by respondents' motivation to fill in the SSE questionnaire. Neither a-motivation, nor controlled or autonomous motivation was found to have an effect on respondents' perceptions regarding differentiation.

Research about respondents' motivation to take a survey does not often discern the quality of their motivation (Krosnick, 1991). Although different motives can drive respondents' motivation to engage in completing a questionnaire, its operationalisation in empirical research is rather limited (Cannell et al., 1981; Heerwegh & Loosveldt, 2009; Kessler et al., 2000; Krosnick & Presser, 2010). By adopting the insights from the self-determination theory, the current study broadens the perspective on motivation. For instance, not only external pressures but also pressure that respondents put on themselves to fill in an SSE questionnaire was included in the current study (Deci & Ryan, 1985, 2002). The differential findings for the effect of quantity and quality of motivation justify the approach that was taken in the study.

SSE results are affected by self-deception and impression management

This study found that SDR has both direct and indirect effects on the SSE results. Whereas the results showed a direct effect of self-deception on respondents' perceptions about distributed leadership, none was found for scores on differentiation in the classroom. In contrast, a direct effect of impression management on respondents' perceptions about differentiation was found, but not for respondents' perceptions about distributed leadership. The effect of impression management might be explained by respondents experiencing items about differentiation more intrusively (Tourangeau & Yan, 2007). Differentiation links more directly to the teachers' instructional practices or their individual role in a school compared to distributed leadership. Self-deception, in contrast, only impacts the perception of respondents about distributed leadership. By considering both direct and indirect effects, the study also demonstrated that impression management had a statistically significant effect on the perception of distributed leadership.

While the two component model to operationalise the construct of SDR (Paulhus, 1984, 2002) has proven to yield insightful differences regarding differentiation and distributed leadership, some findings raise new questions. For instance, it remains unclear why self-deception only has an effect on respondents' reported perception about distributed leadership. By probing respondents about their answering process, future research can provide insights on how respondents arrive at their reported perception and how they deceive themselves when it concerns distributed leadership.

REMAINING QUESTIONS ABOUT THE MEASUREMENT OF SCHOOL PROCESSES

As for the reason why autonomous, controlled and a-motivation were found to have a differential effect on the reported perceptions on distributed leadership and differentiation, the third study did not provide many clues. Possibly, there was interplay between the types of motivation, the cognitive effort participants put into the completion of the questionnaire, and the difficulty of the items (Alwin & Krosnick, 1991; Krosnick, 1991). It could be argued that the items regarding differentiation were less difficult for respondents since these tapped into a concept they were more familiar with and which lay nearer to their day-to-day experiences and expertise. The easier an item is for respondents, the less opportunity for introducing distorting effects in SSE results. This would imply a relationship between difficulty and the level of distortion in SSE results. The question can be raised as to what information can be asked from SSE participants. Should SSE respondents only be asked about school processes that lie within their field of expertise and their (individual) responsibilities in order to avoid distorted results?

When considering total effects of SDR, it must be concluded that impression management impacted upon respondents' perceptions of both differentiation and distributed leadership. These findings indicate that respondents indeed gave a more favourable impression about their school. The reason as for why perceptions about distributed leadership are more vulnerable for self-deception compared to respondents' perceptions regarding differentiation remains unclear.

This study only focussed on respondents' perceptions about differentiation and distributed leadership. Further research could examine whether the same effects would be found when other classroom level or school level processes are under measurement. This can reveal whether the level into which items aim to tap might explain the found effects. This might also connect with the cognitive validity studies in which it was found that it is hard for teachers to take a helicopter view and to make judgements about the school level. Given the finding that SSE results are not free from distortion, this raises the question as to how users can make sound interpretations of the SSE results. Moreover, it is a serious threat to a valid interpretation and use of SSE results.

2.3 Item nonresponse is not a random phenomenon

The fourth study in this dissertation focussed on the occurrence of item nonresponse. This is the phenomenon whereby respondents skip an item or a series of items and then restart completing the questionnaire. This experimental study started from an authentic SSE context in which 376 teachers completed an SSE questionnaire on, among other topics, distributed leadership and reflective capacity. About half of the respondents were assigned an introduction to the SSE questionnaire with an accountability orientation. The other half was given a development-oriented introduction. By means of generalised linear mixed models, the probability for item nonresponse occurring was calculated.

ITEM NONRESPONSE OCCURS TO A LIMITED EXTENT, BUT NOT RANDOMLY

The assumption that respondents skip an item or a series of items only on an accidental basis does not hold. While overall, the likelihood for item nonresponse is rather limited; we identified crucial variables that influence the occurrence of item nonresponse in SSE questionnaires. The results showed that the probability for item nonresponse occurring when a random respondent processes a random item was limited up to 4.35%. However, the predictive model identified that different variables can generate a problematic situation in which item nonresponse seriously adds up. It was found that as respondents experienced less autonomous motivation, the probability for item nonresponse rose. Next, this study demonstrated that an accountability perspective on evaluation also lead to a higher probability for item nonresponse. Also, the design in which item were formulated mattered. Items formulated in a referent-shift design had a higher probability for item nonresponse compared to consensus design items. In addition, it was found that items that were tapping into the scale of distributed leadership had a higher probability of being left unanswered compared to those that were tapping into scale of reflective capacity. When combining these findings into a worst-case scenario, the probability for item nonresponse adds up to 63.19%.

ITEM NONRESPONSE DEPENDS ON RESPONDENT, CONTEXT AND ITEM CHARACTERISTICS

The study found that the extent to which respondents were autonomously motivated influenced the likelihood for item nonresponse to occur, which points to the importance of making respondents see the value in completing such an SSE questionnaire. This means that they have to internalise the motives for engaging in sharing their perceptions about the processes under review. Furthermore, it needs to be acknowledged that, based on this study, motivating respondents by putting them under pressure does not have an effect on the likelihood for item nonresponse to occur. Next to motivation as a respondent characteristic, respondents' tendency towards SDR was tested and found to have no effect on the probability for item nonresponse. This seems to suggest that if respondents answer in a socially desirable way, they will still make an effort to give a plausible answer. This connects with literature on satisficing, which is the phenomenon where respondents put the least effort into providing a plausible answer (Krosnick, 1991; Krosnick et al., 1996).

Our study also demonstrated that the context in which an SSE is administered has an effect on the occurrence of item nonresponse. In instances where respondents experienced an emphasis on accountability, this lead to more respondents skipping an item or a series of items. In such an accountability-oriented context, respondents probably did not feel a sufficient openness towards their perceptions. It seemed that there was a reluctance on the part of respondents to report on their perception. Possibly, they wanted to avoid generating a negative picture of the school. In that case, rather than being an individual trait, which was accounted for by the measurement of SDR in this study, leaving items unanswered also connects to the phenomenon of window dressing (de Wolf & Janssens, 2007; Perryman, 2009). Instead of generating a plausible response, respondents probably left the item unanswered to avoid sketching a negative image of the organisation.

Also regarding the design of items, this study yielded some interesting findings. Referent-shift items had a higher probability of generating item nonresponse. It can be concluded that referent-shift items were experienced as more difficult questions that required more cognitive effort from respondents to answer. This finding of a higher cognitive complexity is in line with the cognitive validity studies conducted in this dissertation. Furthermore, it was found that items that tap into distributed leadership were more likely to be skipped compared to those tapping into reflective capacity. Both scales tapped into a school level process; however, it could be argued that respondents more easily relate to the reflective capacity items. Reflection has become a popular issue in education in recent years and teachers probably had a clearer idea about what the items were asking for.

STATISTICAL STRATEGIES CANNOT EASILY OVERCOME ITEM NONRESPONSE

These findings can have serious implications for the use of SSE results, especially when there is interplay of these variables in a worst case scenario. In this case, it is not possible to rely on the common statistical strategies to overcome this shortcoming in the results (de Leeuw et al., 2003; Durrant, 2005). In addition, the context of SSE is often characterised by a small number of staff members, which also impedes the use of statistical methods. A high level of item nonresponse can have the consequence of generating distortion in the results. Anyway, it should be recognised that in case of a high level of item nonresponse SSE results need to be considered with caution. Moreover, it could be argued that results are hardly interpretable in such a case.

3 Implications of this dissertation

Based on the findings of this dissertation, some important implications can be formulated. The first section discusses implications for instrument developers in the context of SSE and beyond. The second section includes implications for the field of SSE practice and policy.

3.1 Implications for instrument developers

The findings of this dissertation have important implications for instrument developers in the field of SSE and beyond. Although the current dissertation was embedded in a context of SSE, constructs at an organisational level are often mapped out by means of questionnaires. For instance, in the field of organisational behaviour, employee’s commitment to the organisation is similarly measured (Allen & Meyer, 1990; Goffin & Gellatly, 2001). This makes the implications of this dissertation also valuable for instrument developers from other contexts such as organisational behaviour. The instrument that was tested in the context of this dissertation was a typical example of many other instruments in the field of SSE or in research contexts in which school processes are mapped out. The instrument was of high quality and had undergone a serious methodological testing in its development (Vanhoof et al., 2011). An expert panel had reviewed the instrument, a pilot study was performed, and refinements were made based on the obtained feedback. Also, an appropriate statistical psychometric test was carried out. Nonetheless, it must be acknowledged that some common assumptions regarding SSE results do not hold.

Tremendous efforts are put into the development of SSE instruments, and psychometric testing is an important approach to demonstrate the validity of the results they generate. Notwithstanding the importance of an statistical approach to (pre)test SSE questionnaires, the findings of this dissertation point to the necessity of (pre)testing SSE instruments even more extensively by checking SSE questionnaires among respondents of the target group. The cognitive interviewing technique, as used in this dissertation, is proven to be an appropriate technique to gain insights in respondents' cognitive process and to identify the extent to which this cognitive construing is in line with the instrument developers' intentions (Beatty & Willis, 2007; Madans et al., 2011; Rothgeb, Willis, & Forsyth, 2001; K. E. Ryan et al., 2012; Willis, 2005).

Pretesting SSE instruments by means of cognitive interviewing reveals information about the complexity of the formulation of an SSE questionnaire. Despite many years of survey research and the well-intended efforts of instrument developers to make high quality instruments, it is found that the wording of items can still confuse or distract respondents from what the item is aiming for. Some terms are too abstract or unfamiliar for teachers in spite of their professional background in education. By identifying problematic areas, instrument developers can anticipate them. For instance, when administering an SSE questionnaire, a definition can be displayed for specific abstract or complex terms (Peytchev, Conrad, Couper, & Tourangeau, 2010). This can give respondents more clues and directions as to what is being aimed for when they report on their perception. Researchers and instrument developers need to take into account that teachers or other SSE participants do not always speak the same language. It is crucial to create a common language in order to facilitate the dialogue that is initiated between different actors during an SSE process.

In a context of SSE, items often aim to tap into concepts at a school level. This can be achieved by formulating items in different designs. Although survey literature has already pointed to respondents being better at providing information about themselves than about others (Alwin, 2007), some questionnaires make use of a so called 'referent-shift' item. Such an item is formulated in a way that it requires respondents to provide a statement about the school as a whole (e.g., "In this school everyone has…") (Bliese, 2000; Chen et al., 2004). The results in this dissertation found that referent-shift items had a higher probability of generating cognitively invalid results, and had a significantly higher probability of being skipped by a respondent, compared to a consensus design item (e.g., “I have…”).

Next to the wording or design of an item, there is also a concern about the topic that is under measurement. Across the different studies in this dissertation it was found that some topics or concepts were more prone to distortion by specific variables compared to other concepts. For instance, respondents’ perceptions regarding distributed leadership were impacted by respondents’ tendency towards self-deception, but this is not the case for their perception about differentiation in the classroom. In study four, for instance, it was found that items related to distributed leadership have a higher probability of being skipped by a respondent compared to those that tapped into reflective capacity. Here, it is hypothesised that the items for reflective capacity are more meaningful and less abstract for a respondent. These findings further emphasise the importance of a thorough cognitive pretesting. Such abstract or complex concepts need more refinement in order to arrive at a sound reporting of respondents’ perceptions.

3.2 Implications for the field of SSE practice

The findings in this dissertation raise the question as to what extent the field of quality assurance in education and SSE practice in particular can rely on available SSE questionnaires that fit the particular context and SSE objectives. It has already been argued that schools have difficulties with or even have a rather naïve attitude regarding the choice of an appropriate instrument to portray the functioning of the school (Clift, Nuttall, & McCormick, 1987; Hofman et al., 2005). Given the methodological and psychometric problems that have been identified in this dissertation regarding SSE results, practitioners should be cautious in drawing conclusions on, and making use of, the SSE questionnaire results. Because of the found variation in respondents’ cognitive processes, it is vital to arrive at a common understanding about how the SSE questionnaire results need to be interpreted before further inferences are made.

It seems that a stage of sense-making is crucial in arriving at valid conclusions about the SSE questionnaire results. In such a stage, a more common understanding should be created about what the different items are tapping into. This would facilitate the process of making conclusions and undertaking actions for school development. Moreover, it could be argued that preferably such a common understanding is created before the administration of the questionnaire starts. This would also anticipate the problems that were found in the second study, whereby respondents strayed, or did not cover the broader meaning of an item. This would also provide schools with the opportunity to make the questionnaire their own as advocated by the field of SSE (van der Bij, Geijsel, & ten Dam, 2016; Vanhoof et al., 2011).

Creating such a sense-making stage prior to the administration of the SSE questionnaire could also foster respondents' motives to engage in the completing of the questionnaire. The findings in this dissertation suggest that these motives should be internalised as much as possible (Deci & Ryan, 2002). The more respondents' motives are internalised, the more they see the value in filling in an SSE questionnaire. Fostering feelings of autonomy among respondents can stimulate autonomous motivation. By rousing respondents' curiosity in the matter under review, and by, for example, making them decide about what is asked in an SSE questionnaire, this can enhance their view on the completing of an SSE questionnaire as a valuable task (Hidi & Renninger, 2006). Furthermore, it was found that putting respondents under pressure to complete a questionnaire has no effect on them to avoid skipping items. Moreover, it can only have a distorting effect on their reported perceptions.

The results in this dissertation also point to the effect of respondents' tendency towards SDR. Both impression management and self-deception were found to have a distorting effect on respondents' reported perceptions. Research in which personality traits are mapped out, shows that respondents are actually aware of the fact that their self-perception is distorted (Bollich, Rogers, & Vazire, 2015). This would suggest that in an SSE context, participants should be stimulated to reflect critically about what is asked for in the SSE questionnaire. Perhaps by making them explicitly aware of their tendency towards SDR, the actual effect of SDR can already be reduced. This cannot be seen separately from the context in which an SSE is set up. An environment in which there is an openness regarding each other's perceptions and opinions can enhance respondents' critical reflection (MacBeath, 1999; Vanhoof, 2007). The fourth study in this dissertation also identified that accountability-oriented context increases the probability that respondents leave items unanswered. Therefore, in order to avoid respondents not sharing their perceptions, a development-oriented context contributes to yielding as much perceptions from participants as possible.

However, a rather simplistic perspective should be avoided when it comes to administering questionnaires within the context of SSE. The results of this dissertation point to a complexity of SSE results, which is not self-evident to deal with. They suggest that users of SSE questionnaire results have a critical role, and need to have a thorough insight as to how SSE results can be interpreted. From a policy level, this is an important point of special interest. The conduct of SSEs requires sufficient resources and competencies to handle SSE questionnaire data. Data literacy literature points towards a lack of appropriate knowledge and skills among school staff to support this process. In addition, literature within the field of data-use in education has also already pointed out that an accurate interpretation of data is hampered by a lack of know-how (Kerr et al., 2006; Williams & Coles, 2007). Support and guidance at the level of data-literacy when analysing evidence is necessary and benefit decision-making for school improvement (Mandinach & Honey, 2008; O'Brien et al., 2017).

4 Limitations and strengths of this dissertation

The definition of cognitive validity adopted in this dissertation, refers to the intentions with which the instrument developers develop the SSE questionnaire items (Karabenick et al., 2007; Koskey et al., 2010). However, this is only one perspective to which respondents' cognitive processes could be reflected. Given the argument-based approach to validity (Kane, 2006, 2013), it could be argued that it is at least equally important that the conclusions drawn from respondents' perceptions are in line with what respondents were actually thinking of when completing the items. This would enable making further inferences up to the level of making valid interpretations and use of the SSE questionnaire results. Based on the results in this dissertation, wherein a large variation was found in how respondents cognitively processed SSE items compared to the instrument developers' intentions, it could be hypothesised that there is also variation in how cognitively valid interpretations are made from respondents' reported perceptions. Nevertheless, it is a strength of this dissertation that the instrument developers' intentions could be included. Obtaining a thorough description of what instrument developers aim to map out, and subsequently developing cognitive validity criteria, is not always possible. By taking the opportunity to adopt the instrument developers' perspectives, this dissertation is the first to address the topic of cognitive validity of SSE results.

A challenge in this dissertation was to capture respondents' cognitive processing of SSE items. This was done by a cognitive interviewing technique. Despite the fact that respondents were trained in thinking aloud, it was found that respondents still experienced difficulties in thinking aloud while answering SSE items. The systematic probing technique, which makes use of short, direct questions regarding the different cognitive stages, yielded more information about respondents' thought processes. It could be concluded that the systematic probing technique is more adequate to gain insight into respondents' cognitive process regarding SSE items.

It can be argued that in order to explain variance in respondents' reported perceptions, we made use of self-report measures. These self-report measures, such as motivation could also be distorted. However, in order to map out these variables, we relied on instruments based on many years of research in this field. Furthermore, where possible, we also accounted for possible distortion in the self-report on motivation in the design of the statistical modelling.

By making use of a path model, we accounted for SDR bias in respondents’ self-reporting for motivation.

In order to make generalised statements about distortion in SSE results, a high number of participants are needed. In this dissertation, the data collection was embedded in an authentic context that only allowed a limited number of interventions in the design of the studies to be made. Future research could include more constructs in order to verify whether the same effects are found for respondents’ perceptions regarding other school and classroom level processes. In addition, the effect the length of the questionnaire has on respondents’ reporting of their perception regarding the topic under measurement could be identified. Given the nature of many schools, such a high number of participants are often not possible, which hinders the conduct of advanced statistical modelling. It is a strength of the third and fourth study in this dissertation that data were collected from an authentic, large sample of respondents. This permitted carrying out advanced statistical analyses based on the high number of participants that were part of one organisation and who completed the same instrument.

This dissertation’s strength also lies in the introduction of methodological insights from other disciplines within the field of SSE. The way in which respondents cognitively process items had not been addressed before in the context of SSE. Although a vast knowledge base is available regarding survey research and question evaluation methods, it is found that certain strategies such as the cognitive interviewing technique still need to find their way in the field of SSE. This dissertation contributes in the integration of insights from different disciplines in the context of SSE.

5 Key findings

- Respondents are only to a limited extent cognitively processing items of school self-evaluation questionnaires in line with the intentions of the instrument developers, referred to as cognitive validity. As a consequence, it cannot be readily assumed that SSE questionnaire results are cognitively valid. *(Study 1)*
- Items of SSE questionnaires contain linguistic difficulties that hamper respondents in interpreting and elaborating on items in line with the instrument developers' intentions. *(Study 2)*
- When an item requires a statement about the school as a whole, respondents tend to answer the item by only considering it in terms of one's self. Often, Respondents fail to think about the school as whole, as was envisaged by the instrument developers. *(Study 2)*
- SSE questionnaire results are not free from distortion. Respondents' reported perceptions about differentiation in the classroom are impacted by the quantity and quality of respondents' motivation to fill in the SSE questionnaire. Respondents' reported perceptions about distributed leadership are not. *(Study 3)*
- Respondents' tendency towards impression management has an effect on the SSE questionnaire results for both differentiation and distributed leadership. Respondents' tendency towards self-deception has only an effect on results for distributed leadership, and not on differentiation. *(Study 3)*
- Respondents' behaviour in skipping an item or a series of items, after which they restart completing the questionnaire (i.e. item nonresponse), occurs to a rather limited extent in an SSE context; however, it cannot be assumed that it occurs on a completely random basis. *(Study 4)*
- The probability for item nonresponse increases when respondents are less autonomously motivated to fill in the questionnaire, the SSE is administered in an accountability context, items are formulated in a referent-shift design (i.e. when a statement about the school instead of themselves is required), and the item aims to tap into the concept of distributed leadership (instead of reflective capacity). *(Study 4)*

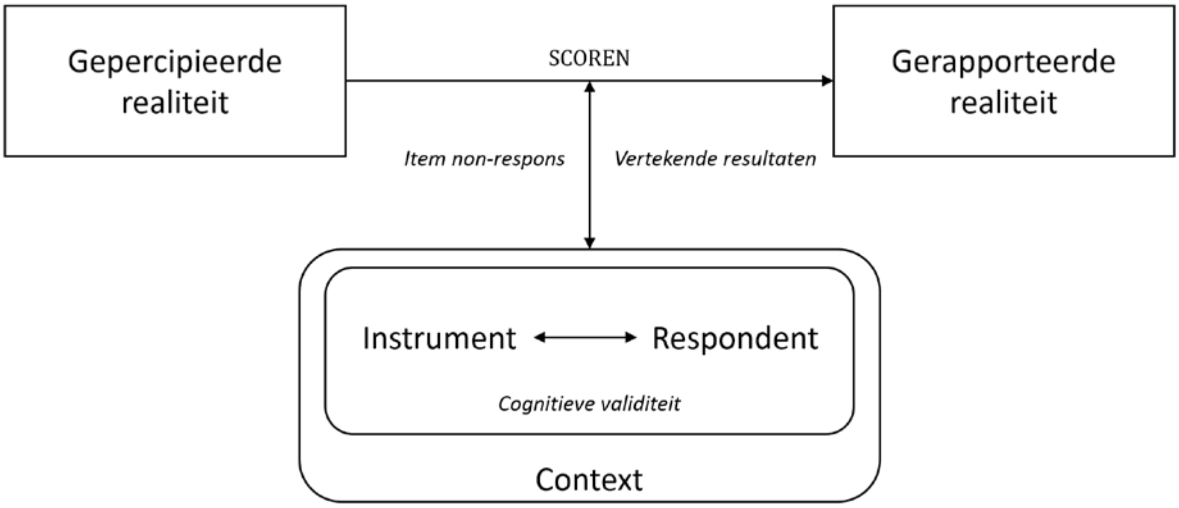
NEDERLANDSTALIGE SAMENVATTING
(Dutch summary)

Dit proefschrift bestudeert de methodologische en psychometrische onderbouwing van schoolzelfevaluatie (SZE) resultaten. Meer specifiek onderzoekt het de validiteit van SZE-resultaten. Deze samenvatting geeft eerst de rationale achter dit onderzoek aan, en beschrijft hoe dit werd opgebroken in vier studies. Vervolgens worden de onderzoeksresultaten beknopt besproken en bediscussieerd.

1 Rationale en onderzoeksdoelen

Van scholen wordt in toenemende mate verwacht dat ze hun eigen kwaliteit monitoren. SZE, als een mechanisme voor interne evaluatie, is een belangrijke strategie voor scholen om tegemoet te komen aan deze verwachting, en is ondertussen een gangbare praktijk in vele onderwijssystemen (Eurydice, 2015; McNamara et al., 2011; OECD, 2013; Schildkamp et al., 2012). In dit proefschrift is SZE gedefinieerd als een systematisch proces, in grote mate door de school zelf geïnitieerd, waarbij welgekozen participanten op een systematische wijze het functioneren van de school beschrijven en beoordelen met het oog op het nemen van beslissingen en initiatieven in het kader van schoolontwikkeling (Vanhoof & Van Petegem, 2010). Een SZE kan verschillende onderwerpen behandelen, maar een focus op processen op schoolniveau kan de impact vergroten in het versterken van de schooleffectiviteit en schoolverbetering vermits deze indicatoren makkelijker beïnvloed kunnen worden (Scheerens, 1991; Van Petegem, 1998). Het zichtbaar of meetbaar maken van processen op schoolniveau is echter een methodologische uitdaging. Een school kan immers niet op een directe manier communiceren over een stand van zaken met betrekking tot die indicatoren. Zoals de voornoemde definitie van SZE omschrijft, wordt er in SZE vaak een beroep gedaan op de percepties van informanten om een beeld te generen van het functioneren van de school (zie Figuur 1). Terwijl verschillende waardevolle strategieën gehanteerd kunnen worden, kan met dit doel voor ogen het afnemen van vragenlijsten inzichtelijke informatie verkregen worden (e.g., Hendriks et al., 2001; MacBeath & McGlynn, 2002; Vanhoof et al., 2011). Deze methode maakt het mogelijk om informatie te verzamelen bij een groot aantal participanten in een beperkte tijdspanne (Cohen et al., 2011). Diverse instrumentontwikkelaars hebben een breed scala aan vragenlijsten uitgewerkt. Echter, de kwaliteit van die instrumenten werd reeds in vraag gesteld en in sommige gevallen wordt aangehaald dat er een gebrek is aan een methodologische en psychometrische onderbouwing (Hendriks, 2000; Hofman et al., 2005). Hoewel de waarde van de perceptie van respondenten wordt onderkend, is het onduidelijk in welke mate deze percepties vertekend zijn. Een zelf-perceptie kan immers (intentioneel of niet-intentioneel) ver verwijderd zijn van hoe een school werkelijk functioneert (Alwin, 2010; Paulhus, 2002). Een zelf-perceptie kan bijvoorbeeld makkelijk ombuigen in zelf-deceptie, waarbij een respondent onbewust zijn of haar school op een positievere manier omschrijft. Vertekende resultaten vormen een ernstige bedreiging voor, in dit geval, het komen tot een beschrijving van het schools functioneren, en waarvan men veronderstelt dat er valide conclusies getrokken worden (Kane, 2013; Meier & O'Toole, 2013). Tot op heden is er weinig empirisch onderzoek uitgevoerd om de validiteit van SZE resultaten. Als centraal onderzoeksdoel beoogt dit proefschrift deze leemte te vullen en de validiteit van SZE resultaten te onderzoeken.

Vertrekkende vanuit een argumentatieve benadering van validiteit (Kane, 2006, 2013), vormt de fase van het scoren, waarbij respondenten vragen beantwoorden, een cruciale stap in het ontwikkelen van een sterk interpretatief argument (zie Figuur 1) (Alwin, 2010; Alwin & Krosnick, 1991; Tourangeau, 2003). Het wordt makkelijk verondersteld dat respondenten items cognitief verwerken zoals de instrument ontwikkelaars ze bedoeld hebben, waarnaar verwezen wordt als cognitieve validiteit (Karabenick et al., 2007; Koskey et al., 2010). In de context van SZE is er echter nog geen empirische evidentie verzameld dat deze assumptie ondersteunt. Bovendien is er weinig geweten over de mate waarin percepties van respondenten vertekend zijn door de motivatie van respondenten om een de vragenlijst in te vullen en hun neiging tot sociaal-wenselijk antwoordgedrag (SWA). Tenslotte wordt het verondersteld dat respondenten een item of een reeks items enkel op een willekeurige basis onbeantwoord laten waarna ze de vragenlijst verder invullen, wat in de literatuur item non-respons genoemd wordt. Dit proefschrift behandelt elk van deze assumpties, en verifieert of deze overeind blijven in de context van SZE. Het testen van deze assumpties is opgebroken in vier studies. Elke studie heeft een ander onderzoeksdoel door centrale concepten af te wisselen.



Figuur 1 Conceptueel model

De eerste twee studies focusten op het denkproces van respondenten bij het beantwoorden van een SZE vragenlijst en onderzochten of respondenten al dan niet denken aan wat de instrument ontwikkelaars bedoeld hadden met de ontwikkelde items. Dit cognitieve proces bestaat uit een interpretatie, elaboratie en respons fase. Daarnaast werd onderzocht welke problemen vastgesteld konden worden in de cognitieve fasen bij het beantwoorden van SZE items door respondenten.

De derde studie beoogde na te gaan in welke mate SZE resultaten beïnvloed zijn door de motivatie van respondenten om een SZE vragenlijst in te vullen, en hun neiging tot SWA. De studie behandelde de volgende onderzoeksvragen: zijn resultaten vertekend door het niveau van respondenten hun autonome, gecontroleerde of a-motivatie. En, heeft de mate waarin respondenten gekenmerkt worden door impressie management en zelf-deceptie een vertekenend effect op SZE resultaten?

De vierde studie behandelde het verschijnsel van item non-respons in de context van SZE's, en stelde in vraag of respondenten werkelijk deelnemen aan de vragenlijst door alle items te beantwoorden. Deze studie focuste op de mate waarin een item of een reeks items werden overgeslagen door een respondent, waarna hij of zij de vragenlijst verder afwerkte. Het onderzoeksdoel van deze studie was om na te gaan of in welke mate item non-respons voorkomt in SZE resultaten, en in welke mate dit voorspeld kan worden door kenmerken van de respondent, het instrument of de context.

2 Belangrijkste onderzoeksbevindingen

Deze sectie maakt een synthese van de belangrijkste onderzoeksbevindingen op basis van de vier studies uit dit proefschrift.

2.1 Cognitieve validiteit kan niet te makkelijk verondersteld worden

Door middel van de cognitieve interview techniek werd het cognitieve proces van respondenten in kaart gebracht bij het beantwoorden van een SZE vragenlijst. Op basis van een corpus van 20 interviews met leerkrachten en directeurs uit vier scholen van het lager onderwijs in Vlaanderen, bepaalde de eerste studie de mate waarin SZE resultaten cognitief valide zijn. Op grond van cognitieve validiteitscriteria voor elk van de drie cognitieve fasen (interpretatie, elaboratie en respons) die respondenten verondersteld worden te doorlopen, werd het niveau van cognitieve validiteit beoordeeld voor 1200 analyse-eenheden.

SZE RESULTATEN ZIJN SLECHTS IN BEPERKTE MATE COGNITIEF VALIDE

De bevindingen uit deze studies wijzen op een ernstig probleem, en tonen dat er een aanzienlijke discrepantie bestaat tussen hoe instrument ontwikkelaars items bedoeld hebben en hoe respondenten ze werkelijk cognitief verwerken. Cognitieve validiteit van SZE resultaten kan niet simpelweg worden verondersteld. Minder dan de helft van de interpretaties van items door respondenten werden als cognitief valide bevonden. Bij het bestuderen van de elaboratie fase, daalde het niveau van cognitieve validiteit nog verder. In deze fase zoeken respondenten in hun geheugen naar relevante informatie. Minder dan een derde van de geanalyseerde elaboraties werden als cognitief valide bevonden. Veel positiever was het gebruik van de vooraf gedefinieerde antwoordopties door respondenten. Meer dan 90 % van de responsen bleken cognitief valide. De studie toonde ook aan dat de interpretatie fase cruciaal is. Wanneer respondenten erin slagen om een cognitief valide interpretatie te maken, is de kans groot dat ook de daaropvolgende taken verwerkt worden op een cognitief valide wijze.

In het voorspellen van de mate van cognitieve validiteit van de interpretatie fase deden enkel verschillen tussen items ertoe, terwijl verschillen tussen respondenten niet significant bleken. Bij het voorspellen van de cognitieve validiteit in de elaboratie fase bleken zowel verschillen tussen items en respondenten statistisch significant effect te hebben. Verschillen tussen respondenten bleken enkel in de elaboratie fase voorspellend te zijn, wat mogelijk verklaard kan worden door het feit dat in die fase respondenten hun persoonlijke ervaringen en mentale modellen integreren bij het beantwoorden van een item (Karabenick et al., 2007; Schwarz, 2007). Meer onderzoek is echter nodig om te achterhalen welke karakteristieken van respondenten hierin een rol spelen. Verder wijst deze bevinding ook op de vitale rol van item kenmerken bij zowel de interpretatie en elaboratie fase. Dit sluit aan bij eerder onderzoek dat wijst op het effect van item design op de kwaliteit van de verkregen resultaten (Knäuper et al., 1997; Krosnick & Presser, 2010; Schwarz, 1999). Een uitgebreidere analyse toonde bovendien aan dat een stelling over de school als geheel, referent-shift items genoemd, een grotere kans hebben om op een cognitief invalide manier verwerkt te worden vergeleken met consensus design items, die enkel een stelling over de eigen persoon vereisen. Instrument ontwikkelaars dienen deze implicaties met betrekking tot cognitieve validiteit in rekening te nemen bij het ontwikkelen van items.

Op basis van deze bevindingen kan geconcludeerd worden dat het beantwoorden van SZE vragenlijst items een behoorlijk veeleisende taak voor respondenten. De assumptie dat SZE respondenten items cognitief verwerken zoals instrument ontwikkelaars ze bedoeld hebben, blijft niet overeind. Respondenten rapporteerden enkel in beperkte mate wat de instrumenten ontwikkelaars voor ogen hadden. Bovendien blijkt de assumptie dat het formuleren van items (in een referent-shift design of consensus design) geen invloed heeft, onterecht. Al deze resultaten doen de vraag rijzen welke problemen in de cognitieve processen van respondenten voorkomen die cognitief invalide resultaten kunnen verklaren.

VERSCHILLENDE PROBLEMEN LEIDEN TOT COGNITIEF INVALIDE SZE RESULTATEN

Door middel van een inhoudsanalyse op de cognitieve interview-data uit de eerste studie, onderzocht de tweede studie in dit proefschrift welke problemen tot cognitief invalide antwoorden op SZE items leiden. Bij elk van de cognitieve taken kwamen problemen in respondenten hun antwoordproces aan het licht. De resultaten uit deze studie toonden aan dat respondenten moeilijkheden ervoeren in het interpreteren van specifieke termen. Vooral enkele linguïstische aspecten werden gevonden die het respondenten moeilijk maakten om betekenis te geven aan items of ze te interpreteren. Respondenten hadden het bijvoorbeeld moeilijk met ongebruikelijke of abstracte woorden of met de zinsstructuur van een item.

Ook bij het elaboreren bij items (waarbij respondenten relevantie informatie uit hun geheugen opzoeken), werden enkele problemen vastgesteld. Sommige respondenten hadden moeite met bij het onderwerp te blijven; hun gedachten dwaalden af en ze namen informatie in overweging dat niet relevant was. Anderen daarentegen namen niet de brede zin van een item in overweging en vermeldden irrelevantie informatie. Er werd ook vastgesteld dat respondenten zich vergisten over een gepast tijdsbestek, niet stilstonden bij een huidige stand van zaken en verouderde informatie aanvoerden. Verder toonden de resultaten aan dat respondenten zich konden vergissen in over wie een stelling gevraagd werd. Het gebeurde dat respondenten enkel zichzelf in beschouwing namen terwijl een stelling over de ganse school vereist was. Dit bracht de idee achter het gebruik van referent-shift items in gevaar.

Deze studie bracht ook enkele problemen aan licht met betrekking tot de respons fase. Respondenten misten soms een antwoordoptie dat hun mentaal oordeel weerspiegelde. Bijvoorbeeld, wanneer een stelling op ging voor de helft van het schoolteam maar niet voor de andere helft, vonden respondenten het onmogelijk om dit te vertalen in een enkele antwoordoptie. Ook het bedoelde gebruik van de ‘weet niet’-optie, voorzien voor respondenten die geen relevantie informatie hadden, werd op verschillende manieren verkeerd gebruikt. Zo gebruikten sommige respondenten deze optie omwille van de complexiteit van een item, simpelweg omdat ze niet wisten wat het item probeerde te vatten. Verder onderzoek zou zich kunnen toespitsen op het effect van een antwoordoptie ‘Ik begrijp dit niet’ zou kunnen zijn op de opgeleverde antwoorden, en of dit zou gebruikt worden zoals bedoeld.

2.2 SZE resultaten zijn niet vrij van vertekening

De resultaten van de derde studie in dit proefschrift wijzen aan dat we niet lichtzinnig kunnen veronderstellen dat de percepties van respondenten vrij zijn van een invloed van hun motivatie om een SZE vragenlijst in te vullen en hun neiging tot sociaal-wenselijk antwoordgedrag (SWA). Deze studie is gebaseerd op een authentieke SZE uitgevoerd in een Vlaamse onderwijsorganisatie waarbij 376 leraren een SZE vragenlijst invulden. Door middel van een padmodel, op basis van een structureel vergelijkingsmodel, werd het effect onderzocht van de motivatie van respondenten om een SZE vragenlijst in te vullen en hun neiging tot SWA op de gerapporteerde percepties over gedeeld leiderschap en differentiatie in de klas.

Zowel kwantiteit als kwaliteit van de motivatie van respondenten beïnvloeden SZE resultaten

De motivatie van respondenten bleek een impact te hebben op SZE resultaten, ook na correctie voor de impact van SWA op de rapportage van respondenten over hun motivatie. De resultaten stelden een statistisch significant negatief verband vast tussen a-motivatie en de percepties over gedeeld leiderschap. Dit betekent dat hoe meer een respondent geen nut ziet in het invullen van de SZE vragenlijst, hoe negatiever hij/zij rapporteert over gedeeld leiderschap. Verder werd een significant verband gevonden tussen gecontroleerde motivatie en de perceptie over gedeeld leiderschap. Des te meer een respondent een interne of externe druk voelt om een SZE vragenlijst in te vullen, des te meer zij/hij positief rapporteert over gedeeld leiderschap in de school. Interessant is de vaststelling dat de mate waarin respondenten autonoom gemotiveerd zijn, geen effect bleek te hebben op de gerapporteerde perceptie over gedeeld leiderschap. Met betrekking tot differentiatie in de klas, werden de gerapporteerde percepties niet beïnvloed door de motivatie van respondenten om de SZE vragenlijst in te vullen. Zowel a-motivatie, gecontroleerde als autonome motivatie bleken geen effect te hebben op de percepties van respondenten omtrent differentiatie.

Onderzoek naar de motivatie van respondenten om een vragenlijst in te vullen maakt vaak geen onderscheid tussen de kwaliteit van die motivatie (Krosnick, 1991). Terwijl verschillende motieven aan de grondslag kunnen liggen van respondenten hun motivatie in het vervolledigen van een vragenlijst, blijft de operationalisering in empirisch onderzoek vrij beperkt (e.g., Cannell et al., 1981; Heerwegh & Loosveldt, 2009; Kessler et al., 2000; Krosnick & Presser, 2010). Door het integreren van de inzichten van de zelf-determinatie theorie, verbreedt de huidige studie de perspectieven over motivatie bij SZE. Bijvoorbeeld, niet enkel externe druk maar ook druk die respondenten die zichzelf opleggen om een SZE vragenlijst in te vullen werd opgenomen in de huidige studie (Deci & Ryan, 1985, 2002). De differentiële bevindingen over het effect van de kwaliteit en kwantiteit van motivatie rechtvaardigen dan ook de gehanteerde aanpak in deze studie.

SZE resultaten zijn beïnvloed door zelf-deceptie en impressie management

Deze studie wees er bovendien op dat SWA zowel directe als indirecte effecten heeft op SZE resultaten. Hoewel de resultaten een direct effect van zelf-deceptie op de percepties van respondenten over gedeeld leiderschap aantoonde, werd er geen effect gevonden op de scores voor differentiatie in de klas. Daarentegen werd een direct effect van impressie management op respondenten hun perceptie over differentiatie gevonden, maar geen voor de percepties over gedeeld leiderschap. Mogelijks ervaren respondenten vragen rond differentiatie als meer gevoelig of indringend, wat het effect van impressie management kan verklaren (Tourangeau & Yan, 2007). Differentiatie relateert immers op een meer directe manier aan de instructiepraktijken van leraren of hun individuele rol binnen een school in vergelijking met gedeeld leiderschap. Zelf-deceptie, in tegenstelling, heeft enkel een impact op de perceptie van respondent rond gedeeld leiderschap. Door zowel directe als indirecte effecten in beschouwing te nemen, toont deze studie aan dat impressie management ook een statistisch significant effect had op gedeeld leiderschap.

Terwijl het twee-componenten model in het operationaliseren van het construct SWA (Paulhus, 1984, 2002) aangetoond heeft interessante verschillen aan het licht te brengen in verband met differentiatie en gedeeld leiderschap, doen sommige bevindingen nieuwe vragen opwerpen. Het blijft, bijvoorbeeld, onduidelijk waarom zelf-deceptie enkel een effect heeft op de door respondenten gerapporteerde percepties rond gedeeld leiderschap. Door het diepgaander bestuderen van respondenten hun antwoordproces, zou toekomstig onderzoek inzicht kunnen verschaffen in hoe respondenten tot hun gerapporteerde perceptie komen en hoe ze zichzelf misleiden inzake gedeeld leiderschap.

2.3 Item non-respons is geen willekeurig fenomeen

De vierde studie in dit proefschrift focuste op het voorkomen van item non-respons. Dit is het fenomeen waarbij respondenten een item of een reeks items onbeantwoord laten en vervolgens de vragenlijst verder invullen. Deze experimentele studie is gebaseerd op een authentieke SZE context waarin 376 een vragenlijst invulden rond, naast andere onderwerpen, gedeeld leiderschap en reflectief vermogen. Ongeveer de helft van de respondenten werden toegewezen aan een introductie tot de vragenlijst met een verantwoordingsgerichte oriëntering. De andere helft werd een ontwikkelings-georiënteerde introductie. Aan de hand van generalised linear mixed models werd de probabiliteit berekend voor het voorkomen van item non-respons.

ITEM NON-RESPONS KOMT IN BEPERKTE MATE VOOR, MAAR NIET WILLEKEURIG

De assumptie dat respondenten een item of een reeks items onbeantwoord laten op toevallige basis blijft niet overeind. Terwijl algemeen genomen de kans op het voorkomen van item non-respons vrij beperkt is, stelden we vast dat enkele cruciale variabelen het voorkomen item non-respons in SZE vragenlijsten beïnvloeden. De resultaten toonden aan dat de probabiliteit voor item non-respons wanneer een willekeurige respondent een willekeurig item verwerkt, beperkt was tot 4.35%. Echter, het voorspellend model stelde vast dat verschillende variabelen een problematische situatie kunnen creëren waarbij item non-respons zwaar kan optellen. Het bleek dat als respondenten minder autonome motivatie ervaarden, de probabiliteit voor item non-respons steeg. Verder toonde deze studie aan dat een verantwoordingsgericht perspectief op evaluatie ook tot een hogere probabiliteit voor item non-respons leidde. Ook het design waarin items werden geformuleerd deed ertoe. Items die geformuleerd werden in een referent-shift design hadden een hogere probabiliteit voor item non-respons, in vergelijking met consensus-design items. Bovendien werd vastgesteld dat items die gedeeld leiderschap in kaart brachten een hogere probabiliteit hadden om onbeantwoord te blijven in vergelijking met items die naar reflectief vermogen peilden. Bij het combineren van deze bevindingen in een worst-case scenario, loopt de probabiliteit voor item non-respons op tot 63.19%.

ITEM NON-RESPONS IS AFHANKELIJK VAN RESPONDENT, CONTEXT EN ITEM KENMERKEN

De studie stelde vast dat de mate waarin respondenten autonoom gemotiveerd waren, beïnvloedde de waarschijnlijkheid waarin item non-respons voorkomt, wat duidt op het belang van respondenten de waarde te doen inzien van een SZE vragenlijst. Dit betekent dat respondenten de motieven moeten internaliseren om hun percepties over de betrokken processen te delen. Verder moet het op basis van deze studie onderkend worden dat het motiveren van respondenten door ze onder druk te zetten geen effect heeft op de mate waarin item non-respons voorkomt. Naast motivatie als respondent kenmerk, werd respondenten hun neiging tot SWA nagegaan, en waarbij geen effect werd vastgesteld op de probabiliteit op item non-respons. Dit zou erop kunnen duiden dat wanneer respondenten op een sociaal-wenselijke wijze antwoorden, ze toch proberen om een plausibel antwoord te genereren. Dit sluit aan bij de literatuur omtrent satisficing, wat het fenomeen is waarbij respondenten met de minst mogelijke inspanning een plausibel antwoord genereren (Krosnick, 1991; Krosnick et al., 1996).

Onze studie toonde ook aan dat de context waarin een SZE wordt afgenomen, een effect heeft op het voorkomen van item non-respons. In situaties waarin respondenten een klemtoon ervaren op verantwoording, leidt dit tot meer item non-respons bij respondenten. In een dergelijke verantwoordingsgerichte context, ervaren respondenten mogelijks onvoldoende openheid ten aanzien van hun percepties. Het leek erop dat er terughoudendheid bij respondenten heerste om hun perceptie te rapporteren. Mogelijks wilde men vermijden om een negatief beeld over de school op te hangen. In dat geval, vormt het onbeantwoord laten van items niet zozeer een individueel kenmerk, waarvoor door middel van het meten van SWA gecontroleerd werd, maar sluit het veeleer aan bij het fenomeen van window dressing (de Wolf & Janssens, 2007; Perryman, 2009). In plaats van het genereren van een plausibel antwoord, lieten respondenten het item mogelijk onbeantwoord om te vermijden de school in een slecht daglicht te stellen.

Ook met betrekking tot het design van items, werden in deze studie interessante vaststellingen gedaan. Referent-shift items hadden een grotere probabiliteit in het genereren van item non-respons. Mogelijks kan geconcludeerd worden dat referent-shift items als moeilijkere items ervaren werden die meer cognitieve inspanning vereisten van respondenten bij het beantwoorden. Deze bevinding van een grotere cognitieve complexiteit is in lijn met de cognitieve validiteitsstudies die in het kader van dit proefschrift werden uitgevoerd. Verder werd vastgesteld dat items die gedeeld leiderschap in kaart brengen een grotere kans hebben om onbeantwoord te laten in vergelijking met die die reflectief vermogen in kaart brengen. Beide schalen peilen naar processen op schoolniveau, echter, het kan gesteld worden dat respondenten zich makkelijker iets kunnen voorstellen bij items die peilen naar reflectief vermogen. Reflectie is immers in de afgelopen jaren een populair concept geworden binnen onderwijs en leraren hadden wellicht een duidelijker beeld van wat de items bevraagden.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179-211. doi: https://doi.org/10.1016/0749-5978(91)90020-T

Allen, N. J., & Meyer, J. P. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology, 63*(1), 1-18. doi: doi:10.1111/j.2044-8325.1990.tb00506.x

Alwin, D. F. (1991). Research on Survey Quality. *Sociological Methods & Research, 20*(1), 3-29. doi: 10.1177/0049124191020001001

Alwin, D. F. (2007). *Margins of error: A study of realiability in survey measurement*. Hoboken, NJ: Wiley.

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* (Second edition ed., pp. 405-434). Bingley, UK: Emerald Group Publishing Limited.

Alwin, D. F., & Hauser, R. M. (1975). The Decomposition of Effects in Path Analysis. *American Sociological Review, 40*(1), 37-47. doi: 10.2307/2094445

Alwin, D. F., & Krosnick, J. A. (1991). The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes. *Sociological Methods & Research, 20*(1), 139-181. doi: 10.1177/0049124191020001005

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*(3), 191-210. doi: 10.1177/0004944116667310

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. doi: 10.1007/s10796-015-9586-x

Bass, B. M., & Riggio, R. E. (2006). *Transformational leadership* (Second Edition ed.). New Jersey: Lawrence Erlbaum Associates.

Bates, D., & Sarkar, D. (2008). *The lme4 Package, 2006*. URL http://cran. r-project. org.

Bateson, N. (1984). *Data Construction in Social Surveys* (Vol. 10). London: George Allen & Unwin.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *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.* Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research (AAPOR), Hollywoord, Florida.

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, Extentions, and New Directions* (pp. 349-381). California; San Francisco: Jossey-Bass.

Bollich, K. L., Rogers, K. H., & Vazire, S. (2015). Knowing More Than We Can Tell:People Are Aware of Their Biased Self-Perceptions. *Personality and Social Psychology Bulletin, 41*(7), 918-929. doi: 10.1177/0146167215583993

Bosnjak, M., & Tuten, T. L. (2001). Classifying Response Behaviors in Web-based Surveys. *Journal of Computer-Mediated Communication, 6*(3), 0-0. doi: 10.1111/j.1083-6101.2001.tb00124.x

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

Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative research, 6*(1), 97-113.

Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability(formerly: Journal of Personnel Evaluation in Education), 21*(1), 47-65. doi: 10.1007/s11092-008-9063-x

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on Interviewing Techniques. *Sociological Methodology, 12*, 389-437. doi: 10.2307/270748

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*(2), 234-246.

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). The Netherlands: Elsevier Ltd.

Christensen, R. H. B. (2015). ordinal - Regression Models for Ordinal Data. R package version 2015.6-28.

Clift, P. S., Nuttall, D. L., & McCormick, R. (Eds.). (1987). *Studies in school self-evaluation*. London: The Falmer Press.

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. doi: 10.1023/a:1023254226592

Conrad, F. G., & Blair, J. (2009). Sources of Error in Cognitive Interviews. *Public Opinion Quarterly, 73*(1), 32-55. doi: 10.1093/poq/nfp013

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

Cote, J. A., & Buckley, M. R. (1987). Estimating Trait, Method, and Error Variance: Generalizing across 70 Construct Validation Studies. *Journal of Marketing Research, 24*(3), 315-318. doi: 10.2307/3151642

Creemers, B. P. M. (1994). 2 - The History, Value and Purpose of School Effectiveness Studies *Advances in School Effectiveness Research and Practice* (pp. 9-23). Amsterdam: Pergamon.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: Sage publications.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly, 69*(1), 87-98. doi: 10.1093/poq/nfi002

Cuyvers, G. (2002). *Kwaliteitsontwikkeling in het onderwijs*. Antwerpen: Garant.

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

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

de Leeuw, E. (2002). International Response Trends: Results of an International Survey. In D. Vaus (Ed.), *Social Surveys* (pp. 121-141). London: Sage.

de Leeuw, E., Hox, J. J., & Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics, 19*(2), 153-176.

de Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of Education, 33*(3), 379-396. doi: 10.1080/03054980701366207

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, E. L., & Ryan, R. M. (2002). *Handbook of self-determination research*. Rochester, NY: University Rochester Press.

DeMaio, T. J., & Landreth, A. (2004). Do Different Cognitive Interview Techniques Produce Different Results? In S. Presser, J. 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, New Jersey, USA: John Wiley & Sons Inc.

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. doi: 10.3102/01623737026001001

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding Self-Report Bias in Organizational Behavior Research. *Journal of Business and Psychology, 17*(2), 245-260. doi: 10.1023/a:1019637632584

Durrant, G. B. (2005). *Imputation methods for handling item-nonresponse in the social sciences: a methodological review*. NCRM Methods Review Papers NCRM/002. Southampton.

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.

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural equation modeling: a multidisciplinary journal, 8*(3), 430-457. doi: 10.1207/S15328007SEM0803_5

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

Eurydice. (2015). Assuring Quality in Education: Policies and Approaches to School Evaluation in Europe. Luxembourgh: European Commission/EACEA.

Faddar, J., Vanhoof, J., & De Maeyer, S. (2017a). Instruments for school self-evaluation: lost in translation? A study on respondents' cognitive processing. *Educational Assessment, Evaluation and Accountability, 29*(4), 397-420. doi: 10.1007/s11092-017-9270-4

Faddar, J., Vanhoof, J., & De Maeyer, S. (2017b). School self-evaluation instruments and cognitive validity. Do items capture what they intend to? *School Effectiveness and School Improvement*, 1-21. doi: 10.1080/09243453.2017.1360363

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

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

Fowler, F. J. (2014). *Survey research methods* (5 ed.): Sage publications.

Fullan, M. (2007). *The new meaning of educational change*. New York: Routledge.

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*(5), 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*(3), 330-338. doi: http://dx.doi.org/10.1016/j.sapharm.2012.04.004

Goffin, R. D., & Gellatly, I. R. (2001). A multi-rater assessment of organizational commitment: are self-report measures biased? *Journal of Organizational Behavior, 22*(4), 437-451. doi: 10.1002/job.94

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, New Jersey: 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.

Hallinger, P., & Huber, S. (2012). School leadership that makes a difference: international perspectives. *School Effectiveness and School Improvement, 23*(4), 359-367. doi: 10.1080/09243453.2012.681508

Harris, A. (2004). Distributed Leadership and School Improvement: Leading or Misleading? *Educational Management Administration & Leadership, 32*(1), 11-24. doi: 10.1177/1741143204039297

Heerwegh, D., & Loosveldt, G. (2009). Explaining the intention to participate in a web survey: a test of the theory of planned behaviour. *International Journal of Social Research Methodology, 12*(3), 181-195. doi: 10.1080/13645570701804235

Helmes, E., & Holden, R. R. (2003). The construct of social desirability: one or two dimensions? *Personality and Individual Differences, 34*(6), 1015-1023. doi: http://dx.doi.org/10.1016/S0191-8869(02)00086-7

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

Hendriks, M., & Bosker, R. J. (2003). *ZEBO: instrument voor zelfevaluatie in het basisonderwijs. Handleiding bij een geautomatiseerd hulpmiddel voor kwaliteitszorg in basischolen.* Enschede: Twente University Press.

Hendriks, M., Doolaard, S., & Bosker, R. J. (2001). School self-evaluation in the Netherlands: Development of the ZEBO-instrumentation. *Prospects, 31*(4), 503-518.

Hendriks, M., Doolaard, S., & Bosker, R. J. (2002). Using school effectiveness as a knowledge base for self-evaluation in Dutch schools: the ZEBO-project. . In A. Visscher & R. Coe (Eds.), *School improvement through performance feedback* (pp. 114-142). Lisse: Swets & Zeitlinger.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly, 45*(4), 549-559.

Hidi, S., & Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist, 41*(2), 111-127. doi: 10.1207/s15326985ep4102_4

Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. *Journal of Management, 21*(5), 967-988. doi: 10.1177/014920639502100509

Hofman, R. H., Dijkstra, N. J., & Hofman, W. H. A. (2005). School Self-evaluation instruments: An assessment Framework. *International Journal of Leadership in Education, 8*(3), 253-272. doi: 10.1080/13603120500088802

Holtgraves, T. (2004). Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding. *Personality and Social Psychology Bulletin, 30*(2), 161-172. doi: 10.1177/0146167203259930

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60.

Hox, J. J., de Leeuw, E., & Vorst, H. (1995). Survey Participation as Reasoned Action; a Behavioral Paradigm for Survey Nonresponse? *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 48*(1), 52-67. doi: 10.1177/075910639504800109

Hu, L. t., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modelling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA, USA: Sage Publications.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55.

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the data-driven mantra: different conceptions of data-driven decision making. In P. A. Moss (Ed.), *Evidence and decision making*. USA: Wiley-Blackwell.

Jansen, J. D. (2004). Autonomy and accountability in the regulation of the teaching profession: a South African case study. *Research Papers in Education, 19*(1), 51-66. doi: 10.1080/0267152032000176972

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

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin, 112*(3), 527.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: 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., . . . Kelly, K. L. (2007). Cognitive Processing of Self-Report Items in Educational Research: Do They Think What We Mean? *Educational Psychologist, 42*(3), 139-151. doi: 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 fromThree Urban Districts. *American Journal of Education, 112*(4), 496-520. doi: 10.1086/505057

Kessler, R. C., Wittchen, H.-U., Abelson, J., Zhao, S., & Stone, A. (Eds.). (2000). *Methodological issues in assessing psychiatric disorders with self-reports*. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.

Kirkman, B. L., Tesluk, P. E., & Rosen, B. (2001). Assessing the incremental validity of team consensus ratings over aggregation of individual-level data in predicting team effectiveness *Personnel Psychology, 54*(3), 645-667.

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*(1), 3.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York: Guilford publications.

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. doi: http://dx.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. doi: 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* (pp. 3-90). California; San Francisco: Jossey-Bass.

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

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, UK: 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. doi: http://dx.doi.org/10.1016/S0191-491X(04)90002-8

Labin, S. N. (2014). Developing Common Measures in Evaluation Capacity Building: An Iterative Science and Practice Process. *American Journal of Evaluation, 35*(1), 107-115. doi: 10.1177/1098214013499965

Lam, T. C. M., & Bengo, P. (2003). A Comparison of Three Retrospective Self-reporting Methods of Measuring Change in Instructional Practice. *American Journal of Evaluation, 24*(1), 65-80. doi: 10.1177/109821400302400106

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *biometrics, 33*(1), 159-174. doi: 10.2307/2529310

Lenzner, T. (2012). Effects of Survey Question Comprehensibility on Response Quality. *Field Methods, 24*(4), 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*(7), 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*(8), 437-448. doi: 10.3102/0013189x07311286

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural equation modeling: a multidisciplinary journal, 9*(2), 151-173. doi: 10.1207/S15328007SEM0902_1

Lozar Manfreda, K. (2001). *Web Survey Errors*. Unpublished PhD dissertation. University of Ljubljana. Ljubljana, Slovenia.

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

MacBeath, J., & McGlynn, A. (2002). *Self-evaluation: What's in it for Schools?* : 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, New Jersey, USA: John Wiley & Sons.

Mandinach, E. B., & Honey, M. (Eds.). (2008). *Data-driven school improvement: Linking data and learning*. New York, USA: Teachers College Press.

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

Mathieu, J. E., & Chen, G. (2011). The etiology of the multilevel paradigm in management research. *Journal of Management, 37*(2), 610-641.

Matsunaga, M. (2008). Item Parceling in Structural Equation Modeling: A Primer. *Communication Methods and Measures, 2*(4), 260-293. doi: 10.1080/19312450802458935

McFarland, S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly, 45*(2), 208-215.

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*(4), 638.

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. doi: http://dx.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. doi: 10.1080/03323315.2011.535977

Meier, K. J., & O'Toole, L. J. (2013). Subjective Organizational Performance and Measurement Error: Common Source Bias and Spurious Relationships. *Journal of Public Administration Research and Theory, 23*(2), 429-456. doi: 10.1093/jopart/mus057

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*(1), 53-71. doi: 10.1076/sesi.14.1.53.13867

Meynen, K., Struyf, E., & Adriaensens, S. (2011). Is the beginning teacher ready for the job? The validation of an instrument to measure the basic skills of the beginning teacher in secondary education. *Pedagogische Studien, 88*(4), 266-282.

Millham, J., & Kellogg, R. W. (1980). Need for social approval: Impression management or self-deception? *Journal of Research in Personality, 14*(4), 445-457. doi: http://dx.doi.org/10.1016/0092-6566(80)90003-3

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. doi: http://dx.doi.org/10.1006/obhd.2000.2943

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65*(2), 131-149. doi: 10.1111/j.2044-8325.1992.tb00490.x

Morgan, D. L. (2014). Pragmatism as a Paradigm for Social Research. *Qualitative Inquiry, 20*(8), 1045-1053. doi: 10.1177/1077800413513733

Muijs, D., & Harris, A. (2003). Teacher Leadership—Improvement through Empowerment? *Educational Management & Administration, 31*(4), 437-448. doi: 10.1177/0263211X030314007

Muijs, D., Harris, A., Chapman, C., Stoll, L., & Russ, J. (2004). Improving Schools in Socioeconomically Disadvantaged Areas – A Review of Research Evidence. *School Effectiveness and School Improvement, 15*(2), 149-175. doi: 10.1076/sesi.15.2.149.30433

Nelson, R., Ehren, M., & Godfrey, D. (2015). *Literature Review on Internal Evaluation*. Institude of Education, University College London. London.

Nevo, D. (2001). School evaluation: internal or external? *Studies in Educational Evaluation, 27*(2), 95-106. doi: http://dx.doi.org/10.1016/S0191-491X(01)00016-5

Niemi, I. (1993). Systematic Error in Behavioural Measurement: Comparing Results from Interview and Time Budget Studies. *Social Indicators Research, 30*(2/3), 229-244. doi: 10.2307/27522726

Nisbet, J. (1988). Rapporteur's report. . In C. o. E. S. C. f. R. i. Education (Ed.), *The evaluation of educational programmes: methods, uses and benefits* (pp. 1-9). Amsterdam: Swets & Zeitlinger.

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. doi: doi:10.1177/1356389016684248

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.

OECD. (2007). *Evidence in education: linking research and policy*. Paris, France: OECD.

OECD. (2013). Syngergies for Better Learning: An International Perspective on Evaluation and Assessment *Reviews of Evaluation and Assessment in Education*. Paris.

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

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

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage publications.

Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods. Integrating Theory and Practice* (Fourth Edition ed.). Thousand Oaks, CA: SAGE Publications.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of personality and social psychology, 46*(3), 598.

Paulhus, D. L. (1998). *Paulhus Deception Scales (PDS): The balanced inventory of desirable responding-7*. Toronto, Ontario, Canada: Multi-Health Systems.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ, USA: Erlbaum.

Perryman, J. (2009). Inspection and the fabrication of professional and performative processes. *Journal of education policy, 24*(5), 611-631. doi: 10.1080/02680930903125129

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.

Podsakoff, P. M., MacKenzie, S. B., Jeong-Yeon, L., & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology, 88*(5), 879.

Preskill, H., & Boyle, S. (2008). A Multidisciplinary Model of Evaluation Capacity Building. *American Journal of Evaluation, 29*(4), 443-459. doi: 10.1177/1098214008324182

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for Testing and Evaluating Survey Questions. *The Public Opinion Quarterly, 68*(1), 109-130. doi: 10.2307/3521540

Reynolds, D., Sammons, P., Stoll, L., Barber, M., & Hillman, J. (1996). School Effectiveness and School Improvement in the United Kingdom. *School Effectiveness and School Improvement, 7*(2), 133-158. doi: 10.1080/0924345960070203

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *2012, 48*(2), 36. doi: 10.18637/jss.v048.i02

Rothgeb, J., Willis, G., & Forsyth, B. (2001). *Questionnaire pretesting methods: Do different techniques and different organizations produce similar results?* Paper presented at the Proceedings of the Annual Meeting of the American Statistical Association.

Royston, P. (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.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

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. doi: 10.1177/1098214012441499

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology, 25*(1), 54-67. doi: http://dx.doi.org/10.1006/ceps.1999.1020

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

Scheerens, J. (1991). Process indicators of school functioning: A selection based on the research literature on school effectiveness. *Studies in Educational Evaluation, 17*(2–3), 371-403. doi: http://dx.doi.org/10.1016/S0191-491X(05)80091-4

Scheerens, J. (2000). *Improving school effectiveness*. Paris: UNESCO International Institute for Educational Planning.

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

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

Scheerens, J., Glas, C. A. W., & Thomas, S. (2003). *Educational evaluation, assessment, and monitoring. A systemic approach*. London: Taylor & Francis.

Schildkamp, K. (2007). *The utilisation of a self-evaluation instrument for primary education*: University of Twente.

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., Vanhoof, J., van Petegem, P., & Visscher, A. (2012). The use of school self-evaluation results in the Netherlands and Flanders. *British Educational Research Journal, 38*(1), 125-152. doi: 10.1080/01411926.2010.528556

Schildkamp, K., Visscher, A., & Luyten, H. (2009). The effects of the use of a school self-evaluation instrument. *School Effectiveness and School Improvement, 20*(1), 69-88. doi: 10.1080/09243450802605506

Schoonenboom, J. (2017). A Performative Paradigm for Mixed Methods Research. *Journal of Mixed Methods Research*, 1558689817722889. doi: 10.1177/1558689817722889

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research, 99*(6), 323-338. doi: 10.3200/JOER.99.6.323-338

Schwarz, N. (1999). Self-reports. How the questions shape the answers. *American Psychologist, 54*(2), 93-105.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*(2), 277-287. doi: 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): 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.

Spector, P. E. (1994). Using Self-Report Questionnaires in OB Research: A Comment on the Use of a Controversial Method. *Journal of Organizational Behavior, 15*(5), 385-392. doi: 10.2307/2488210

Swaffield, S., & MacBeath, J. (2005). School self-evaluation and the role of a critical friend. *Cambridge Journal of Education, 35*(2), 239-252.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (2010). *Sage handbook of mixed methods in social & behavioral research* (Second Ed ed.). Thousand Oaks, CA: Sage.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.

Thomas, K. W., & Kilmann, R. H. (1975). The Social Desirability Variable in Organizational Research: An Alternative Explanation for Reported Findings. *The Academy of Management Journal, 18*(4), 741-752. doi: 10.2307/255376

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., . . . Reynolds, T. (2003). Differentiating Instruction in Response to Student Readiness, Interest, and Learning Profile in Academically Diverse Classrooms: A Review of Literature. *Journal for the Education of the Gifted, 27*(2-3), 119-145. doi: 10.1177/016235320302700203

Toulmin, S. (1958). *The uses of argument*. Cambridge, MA: Harvard University Press.

Tourangeau, R. (2000). Remembering what happended: 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). New Jersey, USA: Lawrence Erlbaum Associates.

Tourangeau, R. (2003). Cognitive Aspects of Survey Measurement and Mismeasurement. *International Journal of Public Opinion Research, 15*(1), 3-7. doi: 10.1093/ijpor/15.1.3

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., pp. 315-346). Bingley, UK: Emerald Group Publishing Limited.

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

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin, 133*(5), 859.

Ullman, J. B., & Bentler, P. M. (2003). Structural Equation Modeling. In I. B. Weiner (Ed.), *Handbook of Psychology*. NJ, USA: John Wiley & Sons, Inc.

Vallerand, R. J., Blais, M. R., Brière, N. M., & Pelletier, L. G. (1989). Construction et validation de l'échelle de motivation en éducation (EME) [Construction and validation of the motivation scale in education]. *Canadian Journal of Behavioural Science, 21*(3), 323.

van der Bij, T., Geijsel, F. P., & ten Dam, G. T. M. (2016). Improving the quality of education through self-evaluation in Dutch secondary schools. *Studies in Educational Evaluation, 49*, 42-50. doi: https://doi.org/10.1016/j.stueduc.2016.04.001

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. doi: 10.1177/1094428107309322

Van Petegem, P. (1998). *Vormgeven aan schoolbeleid. Effectiviteitsonderzoek als inspiratiebron voor de zelfevaluatie van scholen*. Leuven: Acco.

Van Petegem, P., Cautreels, P., & Deneire, A. (2003). *IZES Basisonderwijs. Instrument voor zelfevaluatie van basisscholen*. Leuven: Uitgeverij Acco.

Van Petegem, P., Devos, G., Mahieu, P., Dang Kim, T., & Warmoes, V. (2006). *Hoe sterk is mijn school? Het beleidsvoerend vermogen van Vlaamse scholen*. Mechelen: Wolters-Plantyn.

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

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: Plantyn.

Vanhoof, J., & Van Petegem, P. (2007). Matching internal and external evaluation in an era of accountability and school development: Lessons from a Flemish perspective. *Studies in Educational Evaluation, 33*(2), 101-119. doi: http://dx.doi.org/10.1016/j.stueduc.2007.04.001

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. doi: http://dx.doi.org/10.1016/j.stueduc.2010.10.001

Vanhoof, J., Van Petegem, P., Verhoeven, J. C., & Buvens, I. (2009). Linking the Policymaking Capacities of Schools and the Quality of School Self-evaluations: The View of School Leaders. *Educational Management Administration & Leadership, 37*(5), 667-686. doi: 10.1177/1741143209339653

Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic Versus Extrinsic Goal Contents in Self-Determination Theory: Another Look at the Quality of Academic Motivation. *Educational Psychologist, 41*(1), 19-31. doi: 10.1207/s15326985ep4101_4

Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., & Lens, W. (2009). Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of Educational Psychology, 101*(3), 671.

Vehovar, V., Batagelj, Z., Lozar Manfreda, K., & Zaletel, M. (2002). Nonresponse in Web Surveys. In R. M. Groves, D. Dillman, J. L. Eltinge & R. J. Little (Eds.), *Survey nonresponse* (pp. 229-242). New York, USA: John Wiley & Sons Inc.

Watling, R., & Arlow, M. (2002). Wishful Thinking: Lessons from the Internal and External Evaluations of an Innovatory Education Project in Northern Ireland. *Evaluation & Research in Education, 16*(3), 166-181. doi: 10.1080/09500790208667016

Wayne, S. J., & Liden, R. C. (1995). Effects of Impression Management on Performance Ratings: A Longitudinal Study. *The Academy of Management Journal, 38*(1), 232-260. doi: 10.2307/256734

Williams, D., & Coles, L. (2007). Teachers' approaches to finding and using research evidence: an information literacy perspective. *Educational Research, 49*(2), 185-206. doi: 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. doi: doi:10.1177/0013164405282467

Yan, T., & Curtin, R. (2010). The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research, 22*(4), 535-551. doi: 10.1093/ijpor/edq037

edu bron

Edubron
University of Antwerp