PEDAGOGISCHE STUDIËN
TIJDSCHRIFT VOOR ONDERWIJSKUNDE EN OPVOEDKUNDE

Principals' and Teachers' Comprehension of School Performance Feedback Reports. Exploring Misconceptions from a User Validity Perspective

E. Goffin, R. Janssen, J. Vanhoof

Summary School performance feedback can be a tool for school improvement. However, when educational professionals do not comprehend the data they are provided with, they will not arrive at valid inferences and correct diagnoses. We interviewed 23 Flemish primary school teachers and principals, asking them to explain authentic feedback from a national assessment. Framework analysis of think-aloud data reveals that participants' comprehension of typical concepts is clouded by a range of misconceptions. We observed that that visual, verbal and mathematical building blocks in the report can become stumbling blocks. Moreover, misconceptions can be attributed to a certain extent to disconnects between feedback providers' and feedback users' frames of reference. These findings have important implications for data providers, considering they have a responsibility to cater to the interpretability of the data they provide.

Keywords data-based decision making, school performance feedback, score reporting, sensemaking, user validity

PEDAGOGISCHE STUDIËN
https://doi.org/10.59302/ps.v100i1.13991
2023 (100) 67-97

Artikelgeschiedenis
Ontvangen: 8 juli 2022
Ontvangen in gereviseerde vorm: 2 februari 2023
Geaccepteerd: 22 maart 2023
Online: 28 april 2023

Contactpersoon
Evelyn Goffin,
evelyn.goffin@uantwerpen.be

Financiering onderzoek
This study is part of a research project on school performance feedback use within the Policy Research Center for Test Development and Assessments (STEP), funded by the Flemish Ministry of Education and Training (Flanders, Belgium).

Belangen
De auteurs hebben geen belangen te vermelden.

1 Introduction

Policymakers, researchers and test developers provide schools with high quality achievement data, expecting those data to become drivers for school improvement (Hellrung & Hartig, 2013; van der Kleij & Eggen, 2013; Visscher & Coe, 2003). The assumption is that teachers and principals will use school performance feedback (SPF), for instance from a standardized assessment, as a mirror to identify strengths and weaknesses, and take action accordingly. In practice, however, distribution of test scores and assessment feedback may bring about no effects at all (Hopster-den Otter, Wools, Eggen, & Veldkamp, 2017; Vanhoof, Verhaeghe, Verhaeghe, Valcke, & Van Petegem, 2011; Verhaeghe, Schildkamp, Luyten, & Valcke, 2015) or result in unintended effects (Spillane, 2012; Visscher & Coe, 2003). Misuse, underuse and unintended uses of SPF sometimes stem from recipients' issues with accurately comprehending the data provided. In the present study, we address a fundamental complication that compromises (the effectiveness of) SPF use: the nature of educational professionals' misconceptions when processing typical SPF reports.

Contemporary models emphasize that validity is a property of human interpretation rather than a property of an inanimate test or a score report (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Kane, 2013; O'Leary, Hattie, & Griffin, 2017). Unfortunately, educational professionals often lack the necessary skills and knowledge to effectively interpret data (Hellrung & Hartig, 2013; Hopster-den Otter et al., 2017), as they struggle with comprehending statistical measures and/or visualizations of those measures. In order to determine how SPF can be optimally tailored to educational professionals' data literacy, more insight is needed into actual user interpretations of pupil achievement data (O'Leary et al., 2017; Shivraj & Ketterlin-Geller, 2019; van der Kleij, Eggen, & Engelen, 2014). SPF reports and dashboards are "the primary interface between test developers and […] educational stakeholders" (Gotch & Roduta Roberts, 2018, p. 46) and the way they present information is instrumental in determining whether SPF users will be capable of arriving at valid interpretations.

A central issue is that educational professionals do not simply use data i.e. receive a message and implement adjustments accordingly – data users make sense of data (Earl & Fullan, 2003; Schildkamp, 2019). Interpretive sensemaking processes are at the core of contemporary theories of action on data use (Schildkamp, 2019), but they are complex and rooted in sensemakers' personal lenses, prior knowledge, and social and organizational contexts (Goffin, Janssen, & Vanhoof, 2022). Sensemaking entails asking oneself what the data mean, what the data mean for one's class or school, and what to do next. One of the first stages in this process is (individually) picking up cues from raw data: reading the reports and figuring out what the graphs and numbers mean. Comprehension and initial interpretations are crucial as they guide diagnosis and further stages of educational decision-making.

https://doi.org/10.59302/ps.v100i1.13991

Using a qualitative approach, we examine how teachers and principals construct an understanding of elements presented in authentic SPF reports. Our first research question is descriptive: Do educational professionals comprehend concepts that are central to SPF? (RQ1). This question is rooted in an information-processing paradigm where providers are senders and users are receivers (Ryan, 2006). Our second research question is inspired by a semiotic paradigm and shifts from a mere sender-receiver outlook to a perspective in which SPF reports are seen as communicative tools between providers and recipients (Gotch & Roduta Roberts, 2018; Roduta Roberts, Gotch, & Lester, 2018). How can we explain educational professionals' misconceptions when interpreting SPF? (RQ2). We explore how SPF users interact with graphical, mathematical and linguistic cues in the reports, and how this interaction relates to their (mis)understanding of the data.

2 Theoretical framework

2.1 School performance feedback (SPF)

SPF systems provide schools with formal data about student outcomes or other aspects of school functioning (Hellrung & Hartig, 2013; Visscher & Coe, 2003). Examples range from designated self-evaluation tools, over pupil monitoring systems, to (inter)national assessment programs and central examinations (Verhaeghe et al., 2015). Typically, standardized tests are used, and analyses are based on Item Response Theory (IRT). Performance indicators are fed back on an absolute level (i.e. criterion-referenced, e.g. How do students perform for a particular subject domain?), a relative level for benchmarking (i.e. norm-referenced, e.g. How does group/school-level performance compare to that of a reference group/school?) and/or an ipsative level (i.e. self-referenced, e.g. by giving data about trends over time).

SPF reports characteristically contain numerical, graphical and textual elements. Typical numerical measures include ability scores that express achievement on a certain scale, often including performance levels or score ranges delineated by cut scores. Graphical displays in SPF can take on many forms and levels of complexity. Particular attention in this regard has been given to optimal ways of visualizing measurement error, a concept found to be particularly elusive to SPF recipients (Hopster-den Otter, Muilenburg, Wools, Veldkamp, & Eggen, 2019; Means, Chen, DeBarger, & Padilla, 2011; Zapata-Rivera, Zwick, & Vezzu, 2016). Furthermore, reporting instances vary in the extent to which they provide interpretive guides and other ancillary materials to guide recipients' sensemaking of the data.

2.2 (Ensuring) the validity of SPF

The Standards for Educational and Psychological Testing regard validity and validation as a shared responsibility of feedback providers and feedback users (American Educational Research Association et al., 2014). Feedback providers tread the tightrope of making sure that measurements are technically sound and statistically sophisticated, without compromising reports' interpretability and ease of use. Feedback users, on their part, are expected to possess the capacity to accurately interpret data and effectively use inferences based on those data for decision making. The latter is often referred to as 'data literacy', an umbrella term understood to comprise a rich spectrum of knowledge and skills (Beck & Nunnaley, 2021; Mandinach & Gummer, 2016).

Several authors advocate to place the greater responsibility with feedback providers, stating that it is up to developers to ensure the comprehensibility of SPF (Hattie, 2009) (see 2.2.1). This entails a sensitivity to the fact that there is great individual variability in terms of SPF users' data literacy (Visscher & Coe, 2003; Zapata-Rivera & Katz, 2014; Zenisky, Hambleton, & Sireci, 2009). We will embed data literacy in a broader sensemaking perspective here (see 2.2.2).

2.2.1 Comprehensibility of SPF

Interpretive issues threaten the user validity (a term coined by MacIver, Anderson, Costa, and Evers, 2014) of score reports. However, the literature paints a disconcerting picture with regard to the overall interpretability of score reports (Gotch & French, 2013; Hellrung & Hartig, 2013; O'Leary et al., 2017). On a conceptual level, educational professionals demonstrate a lack of understanding of the constraints of assessment systems (Shivraj & Ketterlin-Geller, 2019) and both criterion- and norm-referenced information in SPF are found to present interpretive challenges (Hellrung & Hartig, 2013). Even basic statistical concepts such as means and percentages have been found to pose problems (Hambleton & Slater, 1997). Educational professionals are also found to struggle with procedural tasks, i.e. extracting information from displays such as charts, graphs and tables in order to subsequently formulate diagnoses and decisions (Gotch & French, 2013; Hambleton & Slater, 1997; Vanhoof et al., 2011; Zenisky et al., 2009). This is particularly the case when no explicit clarification or contextual information is provided (Hellrung & Hartig, 2013) or when additional clarification is in itself too extensive or complex (Hambleton & Slater, 1997).

Research exploring disconnects between SPF provider intentions and user interpretations suggests that choice of words and choice of visual presentations matter in score report design. For instance, the amount of specialized and statistical vocabulary to use is a critical consideration (Shivraj & Ketterlin-Geller, 2019) as narrative elements can be too lengthy, too succinct, or otherwise confusing (Hambleton & Slater, 1997). Jargon can be unfamiliar and sometimes

https://doi.org/10.59302/ps.v100i1.13991

intimidating to SPF users, but at the same time vocabulary can also establish tone and authority (Fjørtoft & Lai, 2021; Roduta Roberts et al., 2018). In some cases, supportive information and tutorials can provide guidance (Zapata-Rivera et al., 2016). However, when sophisticated statistical concepts are employed, such as measurement error, score intervals, reliability and confidence levels, or value-added effects, additional explanations do not appear to suffice to augment comprehension (Gotch & French, 2013; Hopster-den Otter et al., 2019; Zapata-Rivera et al., 2016).

An added challenge is that concepts are often presented using unfamiliar visualizations. Good practices in terms of visual presentation that have been identified are to avoid overly complex or unclear tables and figures, and to favor chart forms that are familiar to users (Hambleton & Slater, 1997; Zapata-Rivera, Vezzu, & VanWinkle, 2013). Other general recommendations are to avoid density and clutter (Goodman & Hambleton, 2004) and to take care that the general lay-out, and the use of colors and symbols are unambiguous (van der Kleij & Eggen, 2013). Furthermore, initial framing is a point of attention: ideally the user's eye is caught by the most important elements first, filling in the details later (Hattie, 2009). Reporting information in different forms (i.e. narrative, numeric, and graphic) shows promise (Goodman & Hambleton, 2004; Visscher & Coe, 2003). However, although presenting a wealth of data can be considered a plus, it can also become overwhelming (Hambleton & Slater, 1997).

2.2.2 Sensemaking of SPF

In line with an argument-based approach to validity (Kane, 2013), a sensemaking perspective in data use research underlines that raw data ('numbers on a page') do not mean anything until a sensemaker has constructed meaning. Sensemaking describes how people make meaning of something new and/or unexpected by figuring out how it fits in with what they already know and assume (Klein, Phillips, Rall, & Peluso, 2007; Maitlis & Christianson, 2014; Weick, 1995). This entails noticing and bracketing certain elements (Coburn & Turner, 2011; Maitlis & Christianson, 2014; Starbuck & Milliken, 1988; Weick, 1995) and weighing them up to personally and/or organizationally held knowledge and beliefs (Klein et al., 2007; Spillane, 2012). If 'conceptions' are the nodes of knowledge that make up the frames people use to make sense of (new) information, 'misconceptions' can be interpreted as the *incorrect* assumptions and convictions that seep into these frames and lead to (systematic and persistent) errors (Prinz, Golke, & Wittwer, 2021; Smith, diSessa, & Roschelle, 1994).

Because sensemaking is a search for coherence, people tend to focus on elements that they perceive as important and relevant, and attempt to frame new information into familiar models and schemata (Klein et al., 2007; Starbuck & Milliken, 1988; Weick, 1995). In terms of SPF use, (un)familiarity with concepts and representations can stem from the amount of experience one actually has with processing SPF, but also to one's work role, training or general statistical knowledge (van der Kleij et al., 2014; Zapata-Rivera et al., 2013). Score report interpretation can also be colored by the way a user relates the new information to their own (assessment) context (Means et al., 2011), by users' motives to consult SPF (Roduta Roberts et al., 2018) and by past uses (Meyer-Beining, 2020). Prior research found that users disregard elements which elude or confuse them, because they do not find them to be sufficiently meaningful (Hambleton & Slater, 1997; Hellrung & Hartig, 2013; van der Kleij & Eggen, 2013).

As documents-in-interaction (Meyer-Beining, 2020) SPF reports mediate meaning between parties, here: SPF providers and users. The present study zooms in on SPF users' initial analyses of raw data: figuring out what the 'numbers on the page' mean. A sensemaking perspective allows us to regard SPF reports as sensemaking resources that have interpretive flexibility over individual data users (Cho & Wayman, 2014). Looking at SPF reports as material-semiotic artefacts (Fjørtoft & Lai, 2021) proves a framework to acknowledge that properties of the data (their source, the specific verbal and visual cues in the reports, or even data *being* numerical or narrative) can trigger certain frames in SPF users (Farley-Ripple, Jennings, & Jennings, 2021; Fjørtoft & Lai, 2021). Moreover, it provides a framework to both academically understand and practically ensure the (user) validity of SPF.

3 Research context and case

This research was carried out in Flanders (Belgium). Periodically, government-commissioned national assessments (NA) are organized to monitor the extent to which attainment targets are achieved on system level, typically for one particular curricular domain at a time. For each NA, a representative sample of schools is selected for participation, which is low-stakes as individual schools' results carry no consequences and are never made public.

Participating schools receive a confidential SPF report. Reports are distributed in PDF format via email to the school, and have a set structure. They start with general information about the setup of the NA program. An interpretive guide explains how system-, school- and class-level results were calculated, what the different components of graphical representations refer to, and what is meant with central concepts such as statistical significance. General guidelines are provided for using the results, including where to turn to for support: users can contact the research team when they have questions about the NA and about specific elements in the SPF report, and are explicitly encouraged to call upon pedagogical counsellors in order to interpret the SPF in light of their schools' own goals, strengths and weaknesses.

Personalized school results in the SPF are broken down into results per test, i.e. per cluster of attainment targets. This feedback is both criterion-referenced (What proportion of pupils reach the attainment targets?) and norm-referenced (How did the school perform compared to the general population and to schools with a similar student population?). First, a brief overview is given of the number of participating students. Second, a table shows the distribution of ability scores, as well as the number of students reaching the attainment targets, and the mean ability score. This table includes school- and class-level results and juxtaposes them to the national results from the reference group. An example of this table is included as Figure A1 in the Appendix, accompanied by a short annotation explaining the setup and the different elements. Third, two caterpillar plots position the school within the sample. One plot compares the school's actual score to the national average and to the statistically expected score based on pupil characteristics. The other plot expresses value-added effects, i.e. differences between schools' actual and expected scores. Annotated examples are included in the Appendix, see Figure A2 and Figure A3. Please note that the figures and annotations in the Appendix provide background information needed to fully appreciate the setup of the data collection and the findings as presented in the following sections.

4 Methodology

https://doi.org/10.59302/ps.v100i1.13991

4.1 Instrument

We examined teachers' and principals' analysis of authentic SPF reports by conducting semi-structured interviews with a think-aloud procedure, because this methodology is considered particularly fit to examine actual user interpretations (Espin, Wayman, Deno, McMaster, & de Rooij, 2017; Goodman & Hambleton, 2004). Moreover, this approach resonates with the discursive nature of sensemaking (Maitlis & Christianson, 2014) and with a semiotic perspective aimed at investigating the meaning that people attribute to signs (Patton, 2015).

In order to ensure a sufficient degree of standardization, the largest part of the interview focused on schools' results on one focal test from an SPF report users were recently presented with. In the think-aloud section, participants were asked to explain the table (see Figure A1 in the Appendix) and caterpillar plots (see Figure A2 and Figure A3 in the Appendix) in their own time and "as if speaking to a colleague". The interviewer noted which components (see Table A1 and Table A2 in the Appendix) were addressed, and probed them where necessary and feasible. As the data collection served a broader purpose beyond the scope of the present study, the full interview protocol also included a range of questions to illuminate other aspects of educational professionals' sensemaking of authentic SPF, such as their appraisal of the results and the factors to which they attribute school performance.

4.2 Participants and data collection

The target population consisted of Flemish primary schools that participated in the 2019 NA of People and Society (formerly a subdomain of the world studies curriculum) in the sixth grade (*N*=99). Spatial use, Traffic and Mobility was selected as the focal test. To avoid school self-selection, i.e., to prevent that only schools performing exceptionally well or poor would volunteer or agree to participate, we pursued a design with sufficient variance in both criterion- and norm-referenced school results (purposive sampling, Patton, 2015). In order to allow for targeted recruitment, all schools that had taken the focal test (*N*=57) were categorized into a crossed design consisting of four profiles based on two dimensions: the percentage of pupils that had reached the attainment targets (i.e. criterion-referenced: "high" versus "low", with 70% of pupils as a cutoff) and school performance compared to similar schools based on statistical expectations for the student population (i.e. norm-referenced, higher or lower). Prospective schools were approached approximately one week after having received the SPF. Interviews were planned over the course of the following four weeks at times best suited to participants' schedules.

As SPF aims to inform both school policy and instructional practice, and since NA are conducted at the end of specific grades, we sought the cooperation of principals as well as sixth-grade teachers. In total, we needed to contact 26 schools in order to be able to recruit sufficient participants. Reasons to actively decline participation, included lack of time and reluctance to participate because the invitee(s) were new at their school or in their function. Ultimately, 1 joint interview and 21 one-on-one interviews were held with 23 participants (11 teachers and 12 principals) from 13 schools. As shown in Table 1, participants' ages ranged from 26 to 60 years old (mean age: 42) and their experience in education ranged from 5 to 40 years (mean experience: 18 years). The majority held a bachelor's degree and had not received any (extensive or specific) training in statistics.

All interviews were organized and conducted by the first author, who identified herself to participants as an employee of the NA research center. Prior to the interviews, participants were informed about the general goals of the study. The invitation letter stated that the interviews were aimed at exploring the "readability" of feedback reports, and the way educational professionals give meaning to results from standardized tests such as the national assessment in practice. Participants were also advised of the ethical clearance obtained, and were told they did not need to prepare in advance. Interviews were conducted online with an average duration of 48 on topic minutes. Video and audio recordings were transcribed verbatim.

Table 1

Participants

School	Participant	Role	Age	Degree	Years of experience in education	Stat Train [a,d]	Stat Prof [b,d]	Inf Use [c,d]
01	Valerie	principal	36	MA	13	Yes	No	Yes
	Sandra	teacher	37	BA	6	Yes	No	na
02	Rebecca	teacher	53	BA	5	No	No	No
03	Paula	principal	36	BA	15	No	No	Yes
04	Frank	principal	52	BA	32	No	No	No
	Natalie	teacher	36	BA	15	No	No	No
05	Jenny [e]	principal	50	BA	28	No	No	Yes
	Melanie [e]	principal	33	MA	10	Yes	na	Yes
	Laura	teacher	39	BA	18	Yes	No	No
06	Heidi	teacher	26	BA	6	Yes	Yes	No
07	Gina	principal	54	BA	34	No	No	No
	Erika	teacher	36	BA	15	Yes	No	No
08	Isaac	principal	39	BA	16	No	No	na
09	Ken	principal	55	BA	32	na	No	N
	Oscar	teacher	29	BA	9	Yes	No	Yes
10	Denise	principal	43	BA	21	No	Yes	Yes
	Quentin	teacher	30	BA	7	No	No	No
11	William	principal	42	BA	21	No	No	Yes
	Tony	teacher	51	BA	26	No	na	Yes
12	Brenda	principal	55	MA	13	Yes	No	na
	Catherine	teacher	39	BA	18	na	No	No
13	Andrea	principal	60	BA	40	Yes	No	Yes
	Xavier	teacher	31	BA	10	Yes	Yes	Yes

Notes.

[a] Stat Train: “I was taught statistics during my training in higher education”.

[b] Stat Prof: “I professionalized in statistics in the course of my career”.

[c] InfUse: “I professionalized in information use in the course of my career (for example: a refresher course in data literacy)”.

[d] Collected via drop-off. Yes = “completely agree” or “somewhat agree”; No = “completely disagree” or “somewhat disagree”; na = “neither agree nor disagree” or “this is not applicable / I don’t know”.

[e] Joint interview.

Table 2

Interpretive scheme for assessing conceptual comprehension

Conceptual dimension	Interpretation
	(How) does the participant express/explain …
ESA – Expression of student achievement	… that this SFB is about students achieving the AT? … ability scores (and how these came about)? … the cutoff i.e. what/where the difference is between reaching and not reaching the AT? … schools' actual scores?
BSP – Benchmarks of s chool performance	… that the school is being compared to the national sample / reference group? … the school's expected score? … the difference between the school's actual score and expected score? … value-added? … statistical significance and its relevance?

Note. AT = attainment targets.

4.3 Data analysis

Transcriptions were analyzed with NVivo. The analysis for the present study focused primarily on the think-aloud section, but also incorporated other parts of the interview, for instance, where participants made inferences about their results or talked about their main take-ways from the report. Framework analysis (Gale et al., 2013) allowed us to search for patterns suggested by the theoretical framework, while also taking into account the idiosyncratic nature of individual participants' sensemaking.

A first step involved isolating participants' utterances about the structural components of the SPF and critically assessing their accuracy. An overview of the components that were elicited (during the interviews) and coded (during analysis) is included in Table A1 and Table A2 in the Appendix, including salient examples of misconceptions we detected. Note that our focus is on the nature of these misconceptions, rather than their prevalence. Particularly in a small, qualitative sample such as ours, a misconception that is uttered once is as informative as one that prevails more broadly.

In a second step, based on a thorough reading of the transcriptions, we interpreted how participants expressed their overall understanding of SPF concepts in reference to the aforementioned report components. The scheme presented in Table 2 served as a guide to assess whether and to what extent these concepts were (sufficiently) comprehended. On the level of individual participants, this comprehension-related information was linked (where meaningful) with the component-related codes.

5 Findings

In section 5.1, we describe whether or not participants succeed in conceptually comprehending the SPF (cf. RQ1), and explore whether (mis)comprehension relates to participants' interaction with report elements (cf. RQ2). In section 5.2, we reflect on misconceptions and the SPF's overall interpretability (cf. RQ2) by taking on broader sensemaking perspective.

5.1 Participants' conceptual comprehension of SPF and the role of SPF elements

5.1.1 Expression of student achievement

The great majority of the participants understand that the SPF pertains to the extent that Flemish attainment targets were reached by pupils in their school, and that the columns in the table (see Figure A1) refer to levels of increasing ability (labeled by many as "categories" or "zones"). Likewise, the divide between 4 and 5 as a cutoff between students that have or have not reached the attainment targets is generally interpreted adequately. While a few participants state they are predominantly interested in 'the bigger picture', a large number of participants critically reflect on table's distribution of low achievers, top scorers, and a middle bracket around the cutoff.

In order to fully grasp what the ability scores refer to, participants need to have read the interpretive guide. One participant states she deliberately disregarded the narrative explanatory sections altogether because she proclaims to be more visually inclined.

> "But I am someone – and that is personal of course – who is better at understanding things when I can see them, rather than when I am reading words. […] So, well, I just make up my own thing from this." (Laura, School 05, teacher)

Even when the concept of ability scores is understood (by reading the interpretive guide), participants do not necessarily possess the vocabulary to reiterate. Some participants explicitly address their lack of confidence in putting it into words. Other participants are not able to articulate at all what ability scores signify or how they came about, or voice clear and striking misconceptions, for example, that the numbers (0-9) refer to specific test items, or to the number of attainment targets that were reached.

> "So, actually, when you look at the Flemish average… If the ability score is 5.9 there, that means that they reach about 60% of the attainment targets?" (Brenda, School 12, principal)

Overall, teachers' and principals' understanding of mean ability scores in the table is strongly linked to the way they understand the construct of ability scores itself. For instance, participants who interpret it as categorical information (insufficient, satisfactory, good etc.) have trouble explaining a mean ability score. Among participants who do accurately interpret (mean) ability scores, the levels of sophistication of analyses diverge. While most will compare the school's mean ability score correctly to that of the reference group, one participant also uses the mean ability score of the reference group as an interpretive benchmark when reflecting on the distribution of ability scores in their own school.

Participants' understanding of the cutoff in the table is aided by the visualization, i.e. the vertical line between 4 and 5, and by explicit verbal cues that state "these pupils have (NOT) reached the attainment targets". However, in order to describe what it means to reach or surpass the cutoff, several participants try to fit SPF concepts into a familiar vocabulary from day-to-day (assessment) practice. The cutoff is for instance incorrectly referred to as "the average", and surpassing the cutoff is described as "passing the test" (a formulation that is justifiable though a little unclear) or "scoring more than half" on the test (which is incorrect).

In the table, many participants focus on the percentage of pupils reaching the attainment targets. However, these percentages are also associated with a myriad of misconceptions. For instance, some participants incorrectly mistake them for the number of attainment targets that have been reached. Additionally, some participants inaccurately label the percentages as "score" or "final grades", making inaccurate statements about how their school "scored X% on the test". Moreover, misconceptions are sometimes extrapolated to the distribution of ability scores. A few participants erroneously claim that the columns describe how many pupils were in "the 10% category, the 20% category and so on". Thus, a percentage is a numerical format that clearly triggers a specific frame of meaning in participants.

5.1.2 Benchmarks of school performance

Many, though not all, participants compare aspects of their school's (or classes') performance to that of the reference group. In the table, the majority of the participants can distinguish between the reference group, the school-level, and the class-level rows (see Figure A1). When interpreting the actual and expected score plot (see Figure A2), the great majority of participants voice clearly that the red dot labeled S indicates their own school's actual score. Participants focus on "their red dot" to make comparative assumptions and inferences by positioning it to other plot elements. With regard to the value-added plot (see Figure A3), a number of participants state that they did not really use it to interpret their result, and/or that they did not manage to make sense of the concept.

The ranking of schools along the X-axis in the caterpillar plot(s) is only addressed in just over half of the interviews, often only implicitly. Nevertheless, those participants tend to understand that the dots represent schools, and that those on the far left resp. the far right have scored the lowest resp. the highest. In order to discuss their school's relative position, participants refer most to the horizontal zero line on the Y-axis (e.g. “we are well above the line”). The majority of participants that explicitly discuss the horizontal line in the caterpillar plot(s) describe it correctly as depicting “the (Flemish) average”, a literal phrase that is present in the plot's auxiliary text.

When prompted, many of the participants who refer to the blue dot as “expected score”, can also express that this is the school's position that would have been expected when taking a range of background characteristics into account. They appear to take their cue from the auxiliary text below the plot. Correct and specific terms like “SES-population” are often used to further elaborate, as this is not an unfamiliar concept to Flemish educational professionals.

Depending on their own school's visual positions in the plots, some participants mistakenly consider their blue dot and the horizontal line to refer to the same thing. One teacher puts the zero line on a par with the cutoff as presented in the table. Without really grasping what is discussed in the caterpillar plot, and finding their school's actual score (just) above the horizontal line and (just) above their expected score, they state they are content with finding their school “above the average”.

“If you are far below the average, you know: ‘oh, that is a problem, we will need to really work on that’. But honestly, anything above is, for me personally, ‘fine’.” (Quentin, School 10, teacher)

A majority of participants disregard confidence intervals when interpreting their schools' position, because they cannot make sense of them at all, and/or because they regard them as non-essential information that could only serve to nuance their interpretation.

“And those vertical lines, well I guess they reflect other things as well but I just read past that. I think.” (Natalie, School 04, teacher)

A small number of participants correctly reiterate from the interpretive guide that confidence intervals express something about the reliability of the NA measurement and that their length depends on the number of participating students. However, a few participants misconstrue the confidence interval as the “range between the strongest and the weakest pupil”. Only a few participants are vocal about the fact that most schools, in the end, do not deviate significantly from the Flemish average.

Finally, in the table, a few participants mistake the system-level information in the top row for school-level results, or indicate that they would expect their colleagues to get confused, because this row is marked in color which draws attention. In the same vein, one participant points out that the use of color in the tables and plots is confusing as the table's top row is highlighted in blue and the expected score dot in the upper caterpillar plot is blue as well.

5.2 Disconnects between SPF providers' and SPF users' frames of reference

Large-scale assessments such as the Flemish NA and the resulting SPF are situated within a specific frame of reference. Our data demonstrate that this frame can conflict with those that teachers and principals employ in daily practice, and inevitably invoke when they make sense of data such as SPF.

5.2.1 (Un)familiar indicators

A sound comprehension of SPF starts with grasping what has been measured. The Flemish attainment targets, as formulated by the educational government, are not always top-of-mind in educational professionals' day-to-day frame of reference. In practice, they work with methods and materials in which the attainment targets have been translated into more concrete terms and objectives. However, particularly when discussing the table, participants do tend to explicitly use the word "attainment targets" or similar terms such as "objectives" or "(minimum) goals" that are commonly used in the Flemish context.

Nevertheless, a number of participants state that, while they are aware of the subject matter the SPF pertains to, they do not exactly know which attainment targets were tested, and would need to look at the documentation in order to refresh their memory. Some participants describe the objectives that were measured predominantly in terms of practical skills, in reference to the concept of *ability* in "ability scores" and/or reminiscing about a practical performance assessment that was also part of the NA.

5.2.2 (Lack of) normative interpretations

There are no explicit normative prescriptions that state which percentage of pupils reaching the attainment targets is considered satisfactory. However, the reference group results are labeled by some as "the standard" or "the expectation", while these elements in fact (neutrally) depict the average achievement on system level. This suggests friction in terms of normative connotations. A school can compare its performance to that of the population, but this does not mean that the average attainment is the criterion to strive towards. Similarly, some participants interpret the idea of an “expected” score as a score to strive for rather than a theoretical construct.

5.2.3 Clashing psychometric perspectives

The measurements presented in the SPF are IRT-based. A student’s position on the measurement scale is not a sum score, as might be the case in classical test theory (CTT) and in typical classroom practice. This disconnect manifests itself in the observations that most participants cannot explain how ability scores were calculated, and that participants inappropriately apply their familiar vocabularies to measurements that do not share the same theoretical foundations. For instance, some participants pick up the recognizable term “the average(s)” and extensively apply it as a label to nearly all different elements in the SPF, such as the cutoff on the measurement scale. It needs to be noted, however, that the SPF providers themselves use the term “average” to refer to multiple constructs (schools’ actual and expected scores as well as the national average from reference group), which may have contributed to confusion.

A related complication is that the IRT-oriented test design of the NA is targeted at group-level and generalized conclusions, and does not allow to make valid statements about individual pupils, individual attainment targets, or even properties of individual test items in terms of detailed error analyses. This is perceived by some as a significant roadblock to being able to interpret the SPF. Typical classroom assessment has a different focus and tends to focus on item-level (error) analysis.

https://doi.org/10.59302/ps.v100i1.13991

5.2.4 (Mis)alignment between the SPF’s statistical complexity and users’ statistical literacy

A number of participants suggest that (particularly) teachers will have trouble in grasping the complexity and level of abstraction of the SPF. Overall, certain central aspects of the SPF are perceived by some as abstract extras that add a layer of complexity unnecessary to form an understanding of the most important messages in the SPF. Consequently, users are not motivated to look at or into them in depth.

> “I can imagine that if you are a layman in statistics, that you just don’t read that part. That you skip it, thinking: ‘is this really essential for me to know?’.” (Melanie, School 05, principal)

This pertains particularly to statistical and psychometric information that requires (some) expertise and/or at least a thorough reading of the interpretive guide. Salient examples are the confidence intervals expressing statistical significance in the caterpillar plots, and the value-added plot in its entirety. Overall, a number of participants state that they feel better able to extract essential information from the tables, with the caterpillar plots having a distinct aura of being harder to digest.

> "I looked at the result first, yes. That was the main thing for me, the extent to which we reached the attainment targets. I have to say that I had to do a double-take on the… uhm… Well, they are in front of me here. … The statistics! I really had to take a real hard look at how this all fits together." (Ken, School 09, principal)

Although we identified a number of misconceptions, most (though not all) participants claim to be confident that they are able to construe at least a basic understanding of the SPF reports. Whereas the extensive interpretive guide was perceived as lengthy and daunting upon first glance, most users need and appreciate the explanations provided in this guide. They generally appreciate the clarity of descriptions and the annotated examples, and the possibility to look up information when struggling to interpret their schools' results. Overall, participants state that the vocabulary used in the SPF is not overly complex. The visual representations in the SPF, and particularly the unfamiliar caterpillar plots, are generally perceived as fairly intricate, but manageable provided there is sufficient processing time.

5.2.5 Diverse preferences and information needs over users

Although we can identify trends, the data illustrate that there is no such thing as "*the* SPF user" and confirm that users make sense of SPF from their own personal perspective.

As illustrated (see 5.2.4), a number of participants focus on the table and regard the caterpillar plots as a nice-to-know extra. One participant explains this by relating that their focus is on "achieving as much as possible with their pupils" and not so much on looking at how the school compares to others or to averages. However, another user regards the confidence intervals as a crucial element and states this was the very first concept they attempted to address. Moreover, the concept of value-added was precisely the element that they were most interested in.

Overall, principals seem somewhat more interested than teachers in benchmarks, i.e. comparing their school's performance to that of other schools. In schools that participated with multiple classes, nearly all participants indicate that they will also compare classes' results. However, teachers tend to particularly zoom in on the results of their own class in the first place.

Finally, notwithstanding that most participants are more invested in the table than in the caterpillar plots, a couple of participants explicitly remark that they would have preferred a graph such as a bar chart to display the distribution of ability scores, adding that other known data providers "also do it like that".

6 Conclusion and discussion

6.1 Conclusion

In this study, we recorded how teachers and principals explain authentic, personalized SPF results from a national assessment in their own words. A first question we sought to explore was whether educational professionals are capable of comprehending concepts that are central to SPF (RQ1). Our findings suggest a nuanced answer. Participants did generally succeed in grasping main messages conveyed in the SPF in terms of expressing student achievement and benchmarking school performance. However, both across participants and within participants, there is a continuum between elementary understanding and being able to handle and/or reiterate more sophisticated conceptualizations. Moreover, we identified a number of concrete misconceptions.

In some cases, misconceptions conceivably invalidate all further interpretation of the results. An example is confusion pertaining to the percentages in the table. When these are misconstrued, further inferences stand no ground. Other examples include participants' difficulties in distinguishing between system-level and school-level results, which inhibit correct benchmarking of school performance. In other cases, one could argue that proverbial pebble stones on the road merely blur a certain aspect of (more advanced) comprehension. For example, without a deep conceptual understanding of measurement scales, ability scores are still accurately interpretable as levels of student achievement. Another (and admittedly more controversial) example would be participants' difficulties with grasping what confidence intervals mean. From an SPF provider's point of view, measurement error and statistical reliability are crucial aspects to interpret psychometric measurements. However, most SPF users feel they succeed in forming an image of their own school's position without using this information. The question remains whether this self-constructed image can (always) be regarded as valid.

In sum, our findings confirm interpretive issues identified in prior research and demonstrate that users' analyses of SPF are not at all straightforward. However, they also suggest that necessary stepping stones are present. SPF providers could reflect on conceptual scaffolding: which elements does a recipient need to construe correct messages in an adequate fashion?

In addition to the descriptive research aim of this study we looked at the way SPF providers represent concepts central to SPF and the way SPF users interact with those representations, in order to find out what contributes to misconceptions (RQ2). We connected with prior research studying said gaps or disconnects by zooming in on users' interpretations of elements in the score reports from an information-processing and semiotic perspective.

To communicate SPF-specific concepts and personalized school results, SPF

https://doi.org/10.59302/ps.v100i1.13991

providers use linguistic, visual and mathematical building blocks. Our findings confirm that these can become stumbling blocks. For one, words matter. Educational professionals use a different vocabulary than SPF providers to talk about achievement, and give their own semantic interpretation to terms and concepts that seem familiar such as ability, average, expectation or significance. This can lead to terminological conflation and sensed discrepancies. Visual presentation matters as well. Even on a very basic level, for instance, use of color merits conscious consideration in SPF report design. Colored highlights direct attention, yet can cause confusion as well. Furthermore, the mathematical and statistical representations SPF providers employ, are not necessarily known or familiar to SPF users – with the caterpillar plots as one of the most striking examples. Overall, even the mere fact that a representation is rooted in statistics, triggers certain frames of meaning in data users (cf. Fjørtoft & Lai, 2021).

Our findings suggest that, in order to aid users' interpretations, SPF providers should build in sufficient demarcation. In the reports' vocabulary, for instance, describing (minimally) different concepts with (overly) similar terms, is a recipe for confusion. The provision of both verbal and visual cues is sensible, but presentations of similar information in different ways should be mutually reinforcing, not obscuring. Rather than trying to fit as much information as possible into one frame, scaffolding of information is advisable (Zapata-Rivera & Katz, 2014).

We also interpreted disconnects in SPF users' take-aways from a broader sensemaking perspective, taking into account that making sense of SPF starts with noticing certain elements (Coburn & Turner, 2011) and involves favoring what matters and what is familiar (Starbuck & Milliken, 1988). We found for instance that some teachers tend to zoom in on their classes, that people are inclined to jump the gun when presented with formats they are used to seeing such as percentages, and that statistical information is sometimes regarded as the bridge too far. These findings demonstrate that even data in raw form cannot be considered neutral, because even at the most fundamental stages of sensemaking there is a sensemaker who constructs meaning from what they see. As further interpretation builds from these nuclear, analytical stages of sensemaking, that are recognition-primed to a certain extent (Klein et al., 2007), it risks becoming monolithic in its inaccuracy.

An overarching observation is that SPF users start within their own frames of reference when interpreting SPF data. These frames differ from those of SPF providers, which to an great extent explains misalignment between providers' intentions and users' interpretations. Moreover, it illuminates the fact that there is no such person as *the* SPF user. Among educational professionals, competences, needs, preferences and expectations diverge. Overall, SPF providers should keep in mind that the language spoken in typical SPF reports is essentially foreign to teachers and school leaders. In order to find alignment, providers should examine what range of frames educational professionals possess, critically assess which frames are necessary to accurately interpret SPF, and gauge whether the frames they build into the SPF (e.g. through an interpretive guide) are sufficiently clear and useful to a recipient. Put differently: preparation entails looking at your data through users' eyes, exploring their frames of references by making them explicit.

6.2 Discussion

Effectively using data for decision-making and for formative purposes in terms of school development and instructional practice, starts with reading and analyzing those data. The sensemaking perspective we took on, postulates that meaning is created instead of given, which has important implications in terms of user validity of SPF. SPF providers may distribute results based on rigorous analysis and envision specific interpretations and uses, but the reports themselves "are where the 'rubber hits the road' in the validity argument for a test" (Zapata-Rivera & Katz, 2014, p. 442). Test developers and SPF providers need to be aware of (potential) roadblocks and disconnects in order to align SPF reports to SPF users' literacy (Hellrung & Hartig, 2013; Hopster-den Otter et al., 2017) and to make sure everyone is 'speaking the same language'. After all, in order to ensure ease of use and to promote valid interpretations, data providers have a responsibility to cater to the interpretability of the data they provide (American Educational Research Association et al., 2014; O'Leary et al., 2017; van der Kleij et al., 2014). The idea of handing out unequivocal meaning on a silver platter is an illusion. In order to find alignment, it is important to not merely define SPF users by their assessment literacy or their statistical literacy (Zapata-Rivera & Katz, 2014). Moreover, as Hattie (2009, p.10) puts it, perhaps we need to reevaluate our sense of directionality: "[…] it is argued that there is no need for "assessment literacy" as teachers need not be required to learn the language of psychometricians. Instead test report developers need to learn the language of teachers, which is teaching and learning.".

This perspective also offers insights into the hazards and opportunities of SPF use in practice. For instance, a negative scenario might be where one team member acts as designated interpreter and introduce static on the line when inaccurately translating SPF results to the rest of the team. However, a positive scenario might include collective sensemaking endeavors that stimulate team members to make their interpretive frames of reference explicit, contributing to the overall richness of interpretation.

Of course, the present study is not without its limitations. The SPF data from our research case were in the form of a static report, which provided us with a stable source of standardization over interview participants. The question is how our findings hold up or need to be interpreted in relation to dynamic forms of score reporting such as data dashboards. The personalization opportunities that such dashboards offer, conceivably put forward even greater challenges in terms of interpretive flexibility over users (Cho & Wayman, 2014; Farley-Ripple et al., 2021). Furthermore, although we discussed authentic SPF data with their actual recipients, the interviews did not constitute an authentic sensemaking setting. Participants were asked to voice individual interpretations in the presence of an interviewer, and we may not assume that participants would construe the same utterances and ideas unprompted, in daily practice. Moreover, as instructed, participants did not specifically prepare for the interview. The course of the interviews showed that certain questions caught several participants off guard, which suggests that they had not yet performed the interpretive exercise on their own.

In order to open the black box of real-life sensemaking of SPF without these distractions, micro-process studies would be particularly suited (Little, 2012; Schildkamp, 2019). Additionally, it would be interesting to embed such studies in a cognitive task analysis or CTA (Clark, Feldon, van Merriënboer, & Yates, 2008). In the present study, much like in CTA, we pre-identified threshold concepts, made use of document analysis and allowed participants to freely voice their train of thought. However, the setup of our study was essentially phenomenographic in nature, as we sought to describe variation in conceptions (Marton, 1981). A systematic CTA-endeavor aimed at identifying typical patterns of reasoning would be useful as a next step, in order to inform further research on specific data sources aimed at educational professionals, and substantiate worked examples of conceptual scaffolding (as suggested in section 6.1).

In any case, as we argued, in order to promote effective data-based decision making, it is necessary to further investigate data use in practice (Coburn & Turner, 2011; Spillane, 2012). Sensemaking is an act of processing reality, therefore we need to take a closer look at how it takes shape in reality. If we want to arm and equip educational professionals with evidence to inform their policy and practice with, we must avoid losing it all to translation.

https://doi.org/10.59302/ps.v100i1.13991

References

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

Beck, J. S., & Nunnaley, D. (2021). A continuum of data literacy for teaching. *Studies in Educational Evaluation, 69*, 100871.

Cho, V., & Wayman, J. C. (2014). Districts' efforts for data use and computer data systems: The role of sensemaking in system use and implementation. *Teachers College Record,* *116*(2), 1–44.

Clark, R. E., Feldon, D. F., van Merriënboer, J. J. G., & Yates, K. A. (2008). Cognitive Task Analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 577–593). Routledge.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective, 9*(4), 173–206.

Earl, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education, 33*(3), 383–394.

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & de Rooij, M. (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research and Practice, 32*(1), 8–21.

Farley-Ripple, E. N., Jennings, A., & Jennings, A. B. (2021). Tools of the trade: A look at educators' use of assessment systems. *School Effectiveness and School Improvement, 32*(1), 96–117.

Fjørtoft, H., & Lai, M. K. (2021). Affordances of narrative and numerical data: A social-semiotic approach to data use. *Studies in Educational Evaluation, 69,* 100846.

Goffin, E., Janssen, R., & Vanhoof, J. (2022). Teachers' and school leaders' sensemaking of formal achievement data: A conceptual review. *Review of Education, 10*(1), e3334.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides : Review of current practices and suggestions for future research. *Applied Measurement in Education, 7347*(July), 1–75.

Gotch, C. M., & French, B. F. (2013). Elementary teachers' knowledge and self-efficacy for measurement concepts. *The Teacher Educator, 48*(1), 46–57.

Gotch, C. M., & Roduta Roberts, M. (2018). A review of recent research on individual-level score reports. *Educational Measurement: Issues and Practice, 37*(3), 46–54.

Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/reports/TECH430.pdf

Hattie, J. A. C. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*, 1–15. Retrieved from http://community.dur.ac.uk/p.b.tymms/oerj/publications/4.pdf

Hellrung, K., & Hartig, J. (2013). Understanding and using feedback – A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review, 9,* 174–190.

Hopster-den Otter, D., Muilenburg, S. N., Wools, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2019). Comparing the influence of various measurement error presentations in test score reports on educational decision-making. *Assessment in Education: Principles, Policy & Practice, 26*(2), 123–142.

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2017). Formative use of test results: A user's perspective. *Studies in Educational Evaluation, 52*, 12–23.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of*

Educational Measurement, 50(1), 1–73.

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In R. R. Hoffman (Ed.), *Expertise Out of Context - Proceedings of the Sixth International Conference on Naturalistic Decision Making* (pp. 113–155). New York, NY: Lawrence Erlbaum Associates.

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education, 118*(2), 143–166.

MacIver, R., Anderson, N., Costa, A.-C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment, 22*(2), 149–164.

Maitlis, S., & Christianson, M. (2014). Sensemaking in organizations: Taking stock and moving forward. *The Academy of Management Annals, 8*(1), 57–125.

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education, 60*, 366–376.

Marton, F. (1981). Phenomenography - Describing conceptions of the world around us. *Instructional Science, 10*(2), 177-200.

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' ability to use data to inform instruction: Challenges and supports*. Washington, DC: Office of Planning, Evaluation and Policy Development, US Department of Education.

Meyer-Beining, J. (2020). "Of course we have criteria". Assessment criteria as material semiotic means in face-to-face assessment interaction. *Learning, Culture and Social Interaction, 24*, 100368.

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice, 36*(2), 16–23.

Patton, M. Q. (2015). *Qualitative Research and Evaluation Methods* (4th ed.). Thousand Oaks, CA: Sage Publications.

Prinz, A., Golke, S., & Wittwer, J. (2021). Counteracting detrimental effects of misconceptions on learning and metacomprehension accuracy: The utility of refutation texts and think sheets. *Instructional Science, 49*(2), 165–195.

Roduta Roberts, M., Gotch, C. M., & Lester, J. N. (2018). Examining score report language in accountability testing. *Frontiers in Education, 3*(June), 1–17.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In *Handbook of Test Development* (pp. 677–710). Routledge.

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research, 61*(3), 257–273.

Shivraj, P., & Ketterlin-Geller, L. R. (2019). Interpreting reports from universal screeners: roadblocks, solutions, and implications for designing score reports. *Frontiers in Education, 4*(108).

Smith, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3*(2), 115–163.

Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education, 118*(2), 113–141.

Starbuck, W. H., & Milliken, F. J. (1988). Executives' perceptual filters: What they notice and how they make sense. In D. C. Hambrick (Ed.), *The executive effect: Concepts and methods for studying top managers* (pp. 35–65). Greenwich, CT: JAI Press.

van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the computer program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation, 39*(3), 144–152.

van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation, 43*, 24–39.

Vanhoof, J., Verhaeghe, G., Verhaeghe, J. P., Valcke, M., & Van Petegem, P. (2011). The influence of competences and support on school performance feedback use. *Educational Studies, 37*(2), 141–154.

Verhaeghe, G., Schildkamp, K., Luyten, H., & Valcke, M. (2015). Diversity in school performance feedback systems. *School Effectiveness and School Improvement, 26*(4), 612–638.

Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement, 14*(3), 321–349.

Weick, K. E. (1995). *Sensemaking in Organizations.* Thousand Oaks, CA: Sage Publications.

Zapata-Rivera, D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy and Practice, 21*(4), 442–463.

Zapata-Rivera, D., Vezzu, M., & VanWinkle, W. (2013). *Exploring teachers' understanding of graphical representations of group performance* (No. RM-13-04). Princeton, NJ.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment, 21*(3), 215–229.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359–375.

Auteurs

Evelyn Goffin is a doctoral researcher in educational sciences at the University of Antwerp and KU Leuven. Her research focuses on the use of school performance feedback for school improvement.

Rianne Janssen is a full professor at the Faculty of Psychology and Educational Sciences at KU Leuven. Her primary research interests include psychometrics and educational measurement.

Jan Vanhoof is a full professor at the Faculty of Social Sciences at the University of Antwerp. His primary research interests include school policy and quality assurance in general and school self-evaluation and data use in education in particular.

Corresponding author: Evelyn Goffin, University of Antwerp, Faculty of Social Sciences, Department of Training and Education Sciences, Sint-Jacobstraat 2, 2000 Antwerp, Belgium. Email: evelyn.goffin@uantwerpen.be

Samenvatting

Begrip en Gebruikersvaliditeit van Schoolfeedback. Onderzoek naar Misconcepties bij Schoolleiders en Leraren.

Schoolfeedback kan een instrument zijn voor schoolverbetering. Echter, wanneer onderwijsprofessionals de data die ze ontvangen niet begrijpen, zullen zij ook niet tot valide conclusies en correcte diagnoses komen. Wij interviewden 23 Vlaamse leerkrachten en directeurs uit het basisonderwijs en vroegen hen om authentieke schoolfeedback uit een peilingsonderzoek te bespreken. Een framework-analyse legt misconcepties bloot die het begrip van typische schoolfeedbackconcepten troebleren. We stellen vast dat de visuele, verbale en wiskundige bouwstenen in het rapport struikelblokken kunnen vormen. Bovendien kunnen misvattingen tot op zekere hoogte worden toegeschreven aan verschillen tussen de referentiekaders van feedbackverstrekkers en feedbackgebruikers. Deze bevindingen hebben belangrijke implicaties voor schoolfeedbackaanbieders, aangezien zij de verantwoordelijkheid hebben om de interpreteerbaarheid van de data die zij ter beschikking stellen te bewaken.

Kernwoorden: geïnformeerde besluitvorming, schoolfeedback, scorerapportering, betekenisgeving, gebruikersvaliditeit

https://doi.org/10.59302/ps.v100i1.13991

Appendix: SPF report elements with annotation

Preliminary note

The figures in this Appendix have been lifted from an authentic SPF report, and were translated from Dutch for the purpose of this paper. The school ID has been fictionalized. The annotations are based on the type of information that is provided more extensively and tailored to the target group in the reports' interpretive guide. Complete examples (in Dutch) of similar SPF reports from the NA's parallel tests, are available online at https://paralleltoetsen.be/voorbeelden.

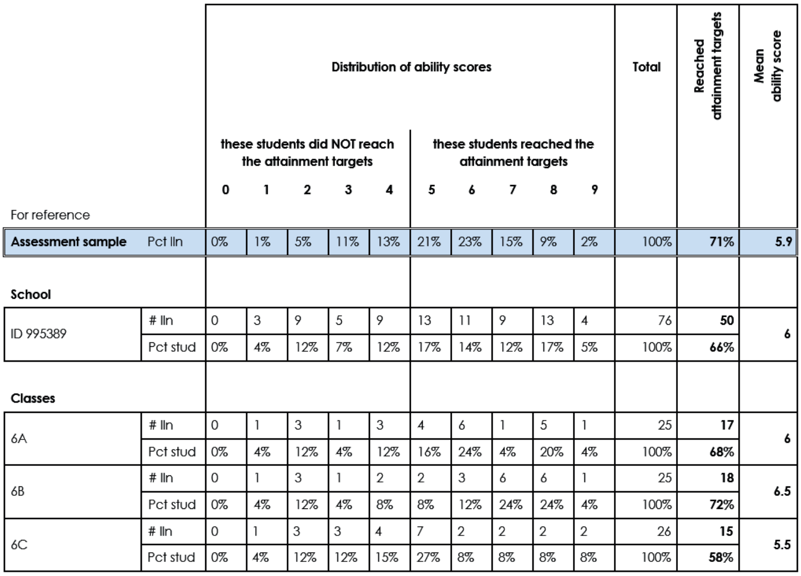
Note that no evaluative judgement is provided in the SPF reports, not for the individual school nor on system level. Users are directed to supplementary material in which the general results of the NA are interpreted and discussed. The emphasis there lies on whether, on system-level, sufficient Flemish pupils reach the attainment targets. Analyses are also presented about background characteristics of schools, classes and pupils that correlate with higher and lower performance levels.

Table expressing student achievement in terms of reaching the attainment targets

Figure A1

Table: Reaching the attainment targets

The table pictured in Figure A1 gives information about the extent to which the



		Distribution of ability scores										Total	Reached attainment targets	Mean ability score
		these students did NOT reach the attainment targets					these students reached the attainment targets							
For reference		0	1	2	3	4	5	6	7	8	9			
Assessment sample	Pct lln	0%	1%	5%	11%	13%	21%	23%	15%	9%	2%	100%	**71%**	**5.9**
School														
ID 995389	# lln	0	3	9	5	9	13	11	9	13	4	76	**50**	**6**
	Pct stud	0%	4%	12%	7%	12%	17%	14%	12%	17%	5%	100%	**66%**	
Classes														
6A	# lln	0	1	3	1	3	4	6	1	5	1	25	**17**	**6**
	Pct stud	0%	4%	12%	4%	12%	16%	24%	4%	20%	4%	100%	**68%**	
6B	# lln	0	1	3	1	2	2	3	6	6	1	25	**18**	**6.5**
	Pct stud	0%	4%	12%	4%	8%	8%	12%	24%	24%	4%	100%	**72%**	
6C	# lln	0	1	3	3	4	7	2	2	2	2	26	**15**	**5.5**
	Pct stud	0%	4%	12%	12%	15%	27%	8%	8%	8%	8%	100%	**58%**	

tested attainment targets have been reached. Information about the extent to which attainment targets have been reached is expressed by way of ability scores (0-9). The cutoff is a psychometric construct derived from the measurement scale: an ability score of 5 and higher corresponds to reaching the attainment targets. The results from the full sample of schools that participated in the assessment, i.e. the reference group, are given for contextualization.

The rows of the table refer to: the reference group on top (marked in a blue color), the school-level, and the class-level. These rows are marked with verbal labels. On school and class level, results are presented in both absolute and relative numbers. Note that in practice, many Flemish primary schools only have one sixth grade class, causing the school and class level rows to show the same numbers.

The columns of the table refer to: the distribution of ability scores (0-9) with an indication of the cutoff between 4 and 5, the total number of participating pupils, the proportion of pupils that have reached the attainment targets, and the mean ability score. All columns have verbal labels. Groups separated by the cutoff are explicitly marked “these pupils have NOT reached the attainment targets” and “these pupils have reached the attainment targets”.

This table is stand-alone i.e. there is no accompanying text that summarizes the main points. However, other chapters of the SPF report explain which attainment targets were tested, reiterate what the setup was of the NA, give basic information about how ability scores were calculated with IRT, and explain how the cutoff needs to be interpreted. An interpretive guide includes a fictionalized example of this table, indicating what the different structural elements of the table refer to.

In Table A1, we list a number of examples of ‘unclarities’ pertaining to the table in the SPF report, that emerged as particularly salient during the interviews. Note that this overview is not intended to be exhaustive. Furthermore, while it indicates a varying range over different components, it does not contain information about the (relative) prevalence of misinterpretations.

Table A1

Examples of problematic aspects and misinterpretations pertaining to the table

Component	Dimension [a]	Examples
Column-level		
Distribution of ability scores / Number of pupils reaching the AT (absolute & relative)	ESA	Numeric labels on top (ability scores) interpreted as referring to specific test items Numeric labels on top (ability scores) and/or relative numbers (percentages) interpreted as test scores Numeric labels on top (ability scores) interpreted as the number of AT (not) reached Highest ability score (9) interpreted as the norm for reaching the AT Idea of ability scores dismissed because too complex or because the visualization on its own does not suffice to grasp the meaning Visualization deemed subpar to other types of visualizations such as bar charts Disproportionate focus on identifying individual students in the absolute numbers Distribution disregarded to interpret overall school performance or to interpret unclarities with regard to mean ability score Percentage(s) interpreted as a proportion of AT that were reached Percentage(s) interpreted as a test score
Cutoff between 4 and 5	ESA	Interpreted as corresponding to students scoring half of the items correctly Interpreted (correctly) as test norm, but norm is interpreted as “scoring 50%” Actual cutoff disregarded, sample’s mean ability score interpreted as “norm”
Mean ability scores	ESA	Non sequitur attempts to calculate a directly corresponding relationship between mean ability score and number/percentage of students reaching the AT (e.g. 60% of students reach the AT, therefore the mean ability score is 5.9”) Confusion / sensed discrepancy between high/low mean ability score and small/large percentage of students reaching the AT Mean explained as the median
Row-level		
Reference Group	BSP	‘Blue bar on top’ actively disregarded because unclear in se / unclear how the reference group was composed Mistaken for school-level information, particularly when looking at total percentage of students reaching the AT and at mean ability scores

School	ESA/BSP	Interpreted without comparison to reference group Confusion with regard to different school locations that form an administrative or functional unit (“vestigingsplaatsen” in Dutch)
Classes	ESA/BSP	Teachers: focus on own class blurs interpretation of general school results (in cases where multiple classes participated)

Notes.

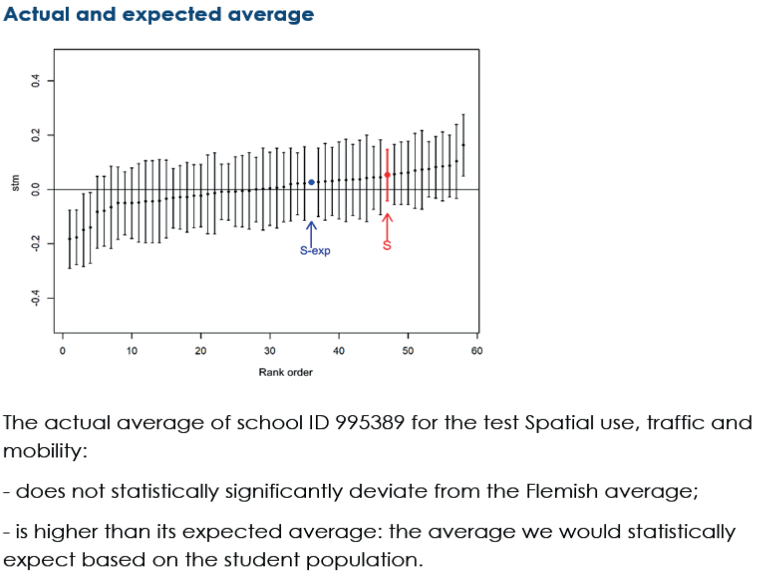
AT = attainment targets.

[a] Conceptual dimension informed predominantly by this component. ESA = expression of student achievement. BSP = benchmarks of school performance.

Caterpillar plots positioning the school’s performance

Figure A2

Caterpillar plot positioning the school’s actual and expected score



Actual and expected average

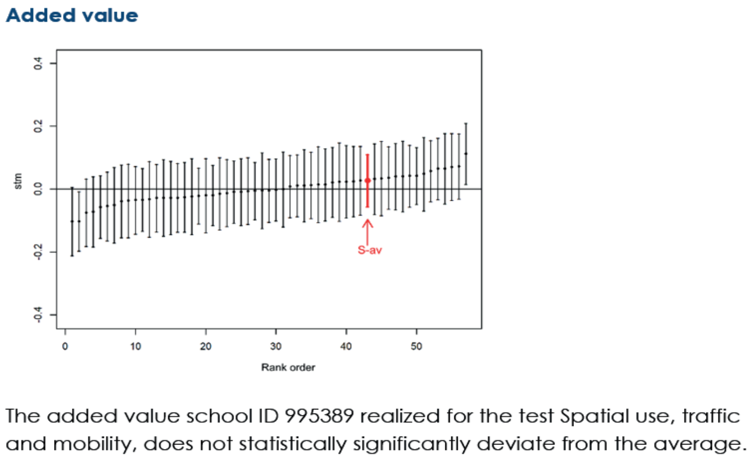
The actual average of school ID 995389 for the test Spatial use, traffic and mobility:

- does not statistically significantly deviate from the Flemish average;

- is higher than its expected average: the average we would statistically expect based on the student population.

Figure A3

Caterpillar plot positioning the school's value-added



Added value

The added value school ID 995389 realized for the test Spatial use, traffic and mobility, does not statistically significantly deviate from the average.

Two caterpillar plots position the school's performance compared to other Flemish schools' performance in the NA.

The first plot (see Figure A2) focuses on the school's raw or 'actual' score and its expected score. On the X-axis, all participating schools are ranked in order of increasing scores. These raw scores are based on mean ability scores and are represented as dots. A red dot indicates the school's own raw score (termed 'actual average'), labeled with the letter 'S'. The blue dot is a theoretical calculation of the school's expected score (termed 'expected average'): the score that would be statistically expected based on a number of pupil background characteristics (i.e. the average NA school with a similar population). It is labeled as 'S-exp' ('S-verw' in Dutch: verw for 'verwacht' i.e. 'expected' in English). On the Y-axis, a horizontal line (the zero line) indicates the "national average", i.e. the mean ability score of the reference group. This allows for a visual comparison of schools' performance to the mean performance in the NA. Furthermore, all score-dots have a vertical line indicating the 95%-confidence interval. If this confidence interval intersects with the zero line, the school's performance does not statistically significantly deviate from the average. Below the plot, auxiliary text is included to verbally express, first, whether the school's actual score significantly deviates from the average (and if so in what direction), and second, whether the schools' actual score is higher or lower than the expected score.

The second plot (see Figure A3) has a very similar setup, but here the dots express value-added i.e. the difference between actual and expected score or the difference a school has made for their student population. Confidence intervals are included here as well. Schools are ranked in order of increasing value-added, the zero line indicates the average value-added. The school's own position is again marked with a red dot, here with the label 'S-av' ('S-tw' in Dutch: tw for 'toegevoegde waarde' i.e. 'added value' in English). Below the plot, auxiliary text is included to verbally express whether the school's added value significantly deviates from the average added value (and if so in what direction).

The interpretive guide in the general part of the report includes annotated fictionalized examples of these caterpillar plots. Also, the specific characteristics that were taken into account to calculate the expected score are listed, and the concept of value-added is explained. Furthermore, explanation is provided about the concept and representation of statistical significance. This explanation describes the confidence interval as a measure of statistical uncertainty i.e. that it is 95% certain that a school's actual performance lies between the upper and lower limits of the vertical line. The shorter the vertical line, the smaller the confidence interval and thus the more reliable the result. The length of the vertical line and, consequently, the degree of certainty are strongly determined by the number of students participating in test. The higher the number of participating students, the smaller the vertical line and the more reliable the result.

Table A2 contains a number of examples of 'unclarities' that we recorded during the interviews, all with regard to the caterpillar plots in the SPF report. Like Table A1, this is not an exhaustive overview nor is the overview intended to indicate the prevalence of specific issues.

Table A2

Examples of problematic aspects and misinterpretations pertaining to the caterpillar plots

Component	Dimension [a]	Examples
Ranking of schools (left to right)	BSP	Interpreted as an absolute classification (i.e. left of the graph being low scorers in absolute terms, instead of lower than the average): "we should all be above the line"
Horizontal line (zero-line, average)	BSP	Interpreted as a normative expectation, sometimes equalled with the cutoff, instead of as an indication of the average: "we should all strive to score above the average" Misinterpretation of the line exacerbates terminological confusion between averages and norms

Position Actual Score (red dot)	ESA/BSP	Mistaken for school's expected score Mistaken for the Flemish average score instead of the school's score
Position Expected Score (blue dot)	BSP	Interpreted as a normative expectation In cases where schools' expected score happens to be positioned on the horizontal line, both are interpreted to refer to the same thing Mistaken for the Flemish average score Mistaken for an expected score for the population instead of for the school Confusion because the dot is blue, and so is the row for the reference group in the table
Position Value-Added	BSP	Hard to grasp even when reading/hearing the explanation Regarded as irrelevant or just nice-to-know Actively disregarded when school's value-added position mirrors school's actual score in the above plot – therefore interpreted as referring to the same thing
Confidence Intervals (vertical lines)	ESA/BSP	Hard to grasp even when reading/hearing the explanation Regarded as irrelevant or just nice-to-know Actively disregarded in favor of the dots Actively disregarded because "they are almost the same for all schools anyway" Some participants (can) reiterate verbal cues below caterpillar plots about statistical significance but cannot relate this to the confidence intervals Some participants (can) reiterate that confidence intervals "have something to do with reliability" but cannot explain further Interpreted as the scoring range between the highest and lowest scoring pupil in a school

Note.

[a] Conceptual dimension informed predominantly by this component. ESA = expression of student achievement. BSP = benchmarks of school performance.