EVELYN GOFFIN

FROM SENSEMAKING TO SCHOOL IMPROVEMENT?

Exploring educational professionals' use of school performance feedback

Dissertation offered to obtain the joint degree of
Doctor of Education Sciences University of Antwerp
and Doctor of Educational Sciences KU Leuven

Faculty of Social Sciences
Training and Education Sciences

Faculty of Psychology and Educational Sciences
Educational Effectiveness and Evaluation

From sensemaking to school improvement?

Exploring educational professionals' use of school performance feedback

Evelyn Goffin

Dissertation offered to obtain the joint degree of
Doctor of Education Sciences and Doctor of Educational Sciences

Supervisors:
Prof. dr. Jan Vanhoof (University of Antwerp)
Prof. dr. Rianne Janssen (KU Leuven)

Antwerpen, 2023

Supervisors

Prof. dr. Jan Vanhoof	University of Antwerp
Prof. dr. Rianne Janssen	KU Leuven

Doctoral jury

Prof. dr. Sven De Maeyer	University of Antwerp	*chair doctoral committee*
Dr. Roos Van Gasse	University of Antwerp	*member doctoral committee*
Prof. dr. Koen Aesaert	KU Leuven	*member doctoral committee*
Prof. dr. Kim Schildkamp	University of Twente	
Prof. dr. Jana Groß Ophoff	University College of Teacher Education, Vorarlberg	

ISBN number 9789057287992

Depotnummer D/2023/12.293/23

STEUNPUNT TOETSONTWIKKELING EN PEILINGEN

Vlaanderen is onderwijs & vorming

This dissertation was supported by the Policy Research Center for Test Development and Assessments (STEP) and was funded by the Flemish Ministry of Education and Training.

Summary

From sensemaking to school improvement? Exploring educational professionals' use of school performance feedback | Dissertation offered to obtain the joint degree of Doctor of Education Sciences and Doctor of Educational Sciences | **Evelyn Goffin** | Supervisors: prof. dr. Jan Vanhoof (UAntwerpen) & prof. dr. Rianne Janssen (KU Leuven)

This dissertation examines how teachers and school leaders *make sense* and *make use* of school performance feedback (SPF) from external standardized assessments, and explores factors that *promote or hinder* these processes. The work is rooted in – and aims to contribute to – research on data-based decision making in education, SPF systems, and score reporting. The conceptual framework has been extended to include perspectives from social and cognitive psychology, such as attribution theory, the theory of planned behavior, and, particularly, sensemaking. Theoretical insights are presented, as well as findings from original empirical research conducted within the context of the Flemish national assessments and parallel tests.

Study 1 reports on an extensive conceptual exploration of 'sensemaking' and of what sensemaking entails when applying this perspective to teachers' and school leaders' engagement with formal achievement data such as SPF. Whereas it has been established that SPF and SPF-like data can be a powerful tool for data-based decision making and school improvement, the complexity of factors that influence educational professionals' data use of these data is not yet fully understood. Building on a review of 25 empirical and theoretical studies, a framework was constructed that integrates insights on the level of the data themselves, the data use process, the individual data user, the social context of the user, users' interactions, as well as the broader system level.

For Studies 2, 3 and 4, qualitative and quantitative inquiries were undertaken to investigate specific aspects of sensemaking and data use, and shed light on what happens when a SPF report comes through a school's proverbial letterbox. Based on data from 22 semi-structured interviews with teachers and school leaders, Study 2 and Study 3 consider user interpretations and the general interpretability of authentic SPF reports. Study 2 critically discusses the user validity of these reports by detecting misconceptions in users' explanations of SPF elements. A disconnect between users' and providers' frames of reference is identified as a source of these misconceptions. Study 3 unpacks the causal attributions that educational professionals make when interpreting their school's results. Findings include that SPF users address a wide range of factors when trying to formulate a diagnosis and tend to turn to external causes for school performance to a great extent. Finally, for Study 4, survey data from 470 educational professionals were used in a path analysis in order to unravel how user-level and school-level factors influence SPF use, and how these factors interplay. Cognitive attitude, perceived expectations of others, and voluntariness in feedback pursuit were found to have positive effects on engagement with SPF in schools. An SPF-oriented school culture, in particular, emerges as an important driver.

By examining the use of SPF from external standardized assessments by teachers and school leaders, this dissertation is at the nexus of educational effectiveness and school improvement research. The four different studies highlight that making sense and making use of educational data in general, and SPF in particular, is not a rational, linear, predictable endeavor. Findings confirm that the mere availability of data such as SPF does not necessarily *drive* data-driven or data-based decision making in schools, and illustrate that SPF use is no one-size-fits-all phenomenon. Implications for policy and practice include that more emphasis should be placed on data cultures and collective sensemaking in schools, on (the balance between) ownership and clear expectations regarding data use, and on the provision of sufficient 'cues' in SPF to aid sensemaking. By involving users and taking into account actual user interpretations to a greater extent, (the delivery of) SPF can be improved. Such measures are necessary to ensure that SPF lives up to its potential: to effectively inform educational decisions and truly contribute to school improvement.

Samenvatting

Betekenis geven aan data om tot schoolontwikkeling te komen. Hoe gaan onderwijsprofessionals aan de slag met schoolprestatiefeedback? | Proefschrift aangeboden tot het verkrijgen van de gezamenlijke graad van Doctor in de Onderwijswetenschappen en Doctor in de Pedagogische Wetenschappen | **Evelyn Goffin** | Promotoren: prof. dr. Jan Vanhoof (UAntwerpen) & prof. dr. Rianne Janssen (KU Leuven)

In dit proefschrift wordt onderzocht hoe leerkrachten en schoolleiders *betekenis geven aan* en *gebruikmaken van* schoolfeedback (SFB) op externe gestandaardiseerde toetsen, en wordt nagegaan welke factoren deze processen *bevorderen of belemmeren*. De kennisbasis waarop dit werk voortbouwt en waaraan het wil bijdragen, betreft onderzoek over geïnformeerde besluitvorming in scholen, schoolfeedbacksystemen en score-rapportage. Het conceptuele kader werd verder aangevuld met perspectieven uit de sociale en cognitieve psychologie, zoals attributietheorie, de theorie van gepland gedrag, en in het bijzonder 'sensemaking' of betekenisgeving. Er worden theoretische inzichten gepresenteerd, alsook resultaten van empirisch onderzoek dat werd uitgevoerd in de context van het Vlaamse peilingsonderzoek.

Studie 1 is een uitgebreide conceptuele verkenning van het betekenisgevingsperspectief en van wat dit perspectief inhoudt wanneer het wordt toegepast op het gebruik van formele prestatiegegevens zoals SFB door leerkrachten en schoolleiders. Hoewel data zoals SFB een krachtig instrument kunnen zijn voor geïnformeerde besluitvorming en schoolontwikkeling, doorgronden we de complexiteit nog niet volledig van factoren die een rol spelen wanneer onderwijsprofessionals aan de slag gaan met zulke data. Op basis van een review van 25 empirische en theoretische studies werd een geïntegreerd conceptueel raamwerk samengesteld met inzichten op het niveau van de data zelf, het verwerkingsproces, de individuele datagebruiker, de context van de gebruiker, de interacties tussen gebruikers, en het bredere systeemniveau.

Voor Studies 2, 3 en 4 werd kwalitatief en kwantitatief onderzoek uitgevoerd om specifieke aspecten van betekenisgeving en informatiegebruik te belichten, en na te gaan wat er gebeurt wanneer een SFB-rapport op de spreekwoordelijke mat valt in scholen. In Studie 2 en Studie 3, die gebaseerd zijn op 22 semigestructureerde interviews met leerkrachten en schoolleiders, wordt ingegaan op gebruikersinterpretaties en op de 'interpreteerbaarheid' van authentieke SFB-rapporten. Studie 2 focust op gebruikersvaliditeit. In de uitleg die gebruikers geven over SFB-elementen worden misconcepties geïdentificeerd. Een discrepantie tussen het referentiekader van de gebruikers enerzijds en dat van de aanbieders anderzijds wordt besproken als een bron van deze misvattingen. Studie 3 onderzoekt causale uitspraken die onderwijsprofessionals doen wanneer zij de resultaten van hun school duiden. SFB-gebruikers blijken een breed scala aan factoren in aanmerking te nemen wanneer zij een diagnose formuleren, en zijn sterk gericht op externe oorzaken. Voor Studie 4, ten slotte, werden surveydata verzameld bij 470 onderwijsprofessionals. Met behulp van een padanalyse werd onderzocht hoe factoren op gebruikersniveau en op schoolniveau het gebruik van SFB beïnvloeden, en hoe deze factoren met elkaar in verbinding staan. Cognitieve attitude, gepercipieerde verwachtingen van anderen, en het actief opzoeken van SFB, blijken positieve effecten te hebben op het gebruik van SFB in scholen. Met name een SFB-georiënteerde schoolcultuur komt naar voren als een belangrijke drijfveer.

Aangezien dit proefschrift onderzoekt hoe leerkrachten en schoolleiders aan de slag gaan met SFB op externe gestandaardiseerde toetsen, bevindt het zich op het raakvlak van effectiviteitsonderzoek en onderzoek rond schoolontwikkeling. De vier verschillende studies wijzen erop dat informatiegebruik, en SFB-gebruik in het bijzonder, geen rationele, lineaire en voorspelbare aangelegenheid is. De onderzoeksresultaten bevestigen dat de loutere beschikbaarheid van gegevens zoals SFB niet noodzakelijkerwijs 'vanzelf' leidt tot geïnformeerde besluitvorming in scholen, en ze illustreren dat het gebruik van SFB geen one-size-fits-all gegeven is. Implicaties voor beleid en praktijk zijn onder meer dat meer nadruk moet worden gelegd op informatieculturen en collectieve betekenisgeving in scholen, op (het evenwicht tussen) eigenaarschap en heldere verwachtingen rond informatiegebruik, en op het verstrekken van voldoende 'aanwijzingen' in SFB om het betekenisgevings-proces te ondersteunen. (Het aanbieden van) SFB kan worden geoptimaliseerd door eindgebruikers meer te betrekken en alleszins meer voeling te krijgen met de feitelijke interpretaties die deze gebruikers maken. Dergelijke ingrepen zijn nodig om ervoor te zorgen dat SFB zijn potentieel waarmaakt: het effectief informeren van onderwijsbeslissingen en het daadwerkelijk bijdragen aan schoolverbetering.

Acknowledgments | Dankwoord

All you've got is this moment
(INXS, *Need You Tonight*, 1987)

June, 2023. Well, I thought very long and very hard about this acknowledgments section, but I put off actually *writing* it until the very last moment, when the deadline was staring me right in the face. Much like many other chapters and sections in this dissertation, come to think of it. You could call it procrastination, or poor planning skills, or crippling self-doubt – and you would probably be correct. But the fact of the matter is that it is a Huge Deal to convey things to the world that you are passionate about, things you find important, things you want to see carefully worded before you Put Them Out There. I felt that way about my research findings and I most definitely feel that way about the gratitude I want to express towards the people who have accompanied me on this crazy, tough, magnificent journey that is now coming to a close.

Dear Sir or Madam, will you read my book?
It took me years to write, will you take a look?
(The Beatles, *Paperback Writer*, 1966)

Jan and Rianne. Where to begin? I have enjoyed working with you so very much. Thank you both for your mentorship, your warm support, your sharp expertise. Thank you for all the laughs we had, and most of all for believing in me when I was the last person to do so myself. (When you're anxious about your work, and you look forward to discussing it with your supervisors because you know that will set your mind at ease, you know you are lucky to be part of a dream team.) Thank you, Jan, for welcoming me in Antwerp with open arms, for your subtle but much-needed mindfulness tips, for your persistence when I was stuck in the umpteenth status-quo, for always being there as a steadfast guide. Thank you, Rianne, for thinking of me when this PhD-opportunity arose, for your never-failing confidence, for your delightfully detailed feedback, for relishing in our shared love of linguistic whimsy. Jan and Rianne, when the doctoral road was rocky, you lifted me up; when the path was dark, you lit the beacons. I did not, and will never, take that for granted.

Speaking of beacons: Sven, Roos, and Koen, thank you for accompanying us by being part of my doctoral committee. You followed my progress with much appreciated interest and empathy, and your constructive feedback helped me see things clearer on more than one occasion. Sven, I admire your kindness and your know-how, and I am so happy to have had you as chair. Roos, I know a few things about data-based decision making and data use in the Flemish context, but you are the queen, that is for sure. Thank you for being an inspiration and a shining example! You truly are second to none in terms of getting all ducks in a row. Koen, I seem to remember some joking banter about your plan to 'not go easy on me' during the defense. As I am writing this, I do not yet know how that panned out – but I do know that, *like Frankie said, you'll do it your way*. As that is usually a thorough, well-seasoned, witty way, I am genuinely looking forward to it.

Kim and Jana, my heartfelt thanks for joining my doctoral jury. Kim, I initially knew you only as an authority on data use, and I was genuinely starstruck when I first heard you speak at a conference back in 2019. Along the way, as we met on other occasions, I also got to know you as a super kind-hearted person. Yet any time we meet, I am still starstruck all over again! Jana, I also remember the first time we met: at a conference, in a small group talking passionately about data use and research literacy, and as it happens, if memory serves, all wearing black Dr. Martens boots. That was one of those moments in the past four and a half years when I knew I had found my tribe. Thank you for staying in touch. Kim and Jana, to have you both weigh in on my research is an honor, a privilege and a delight.

As this research carries the STEP banner, I would also like express my gratitude to the Flemish department of Education and Training, without whose support none of this would have been possible, and to the educational partners who took an interest in my work. I hope that the insights presented in this dissertation may be of use. And of course, I am indebted to the many Flemish teachers and school leaders who so kindly took the time to participate in my studies. Your daily endeavors are, in the end, what it is all about, and I hope I have done you justice with this research project. Last but not least, I would also like to address the masters' students I got to work with over the years. I enjoyed how we searched and found our way together in new material. In particular, I want to thank you, Gila, for allowing me to be part of your journey.

And you may ask yourself,
"Well, how did I get here?"
(Talking Heads, *Once in a Lifetime*, 1980)

The village that helped me raise the research child that lies before you, is very well populated indeed. I have been lucky to have landed in a warm nest of coworkers not once, but twice!

First of all, being part of the Edubron research group at the University of Antwerp is a blessing. It is a daily pleasure to talk to all of you lovely, talented people! It made being together-apart during the pandemic in 2020 and 2021 even more of a challenge, but we pulled through. (Leen, Margot, I am specifically looking at you here. Even when strictly speaking we were no longer colleagues, we remained *bureaubuddies* over the internet and/or over afternoon tea.) Edubronners, I enjoy our meetings, our team activities, and I love that going to a conference with you sort of feels like summer camp. Special acknowledgments go out to my friends at the Policy and Quality Assurance *speerpunt*, past and present. To name but a few: thank you *allerliefste* Lies, Amy and Randi, thank you Ruud (I cherish the day you introduced me to that Morph transition in PowerPoint), Dries (my NVivo Jedi master), and Glen (we chat way too much when we are both at the office but I would not want it any other way).

Of course, the trip that ultimately led to this doctoral degree, started in the autumn of 2007 when Sarah hired me at KU Leuven. Thank you, Sarah – I guess neither of us could have dreamed back then that I would be standing here today. For the ensuing fifteen-plus years, I had the best of times with the best of people at the Centre for Educational Effectiveness and Evaluation. Thank you, each and every one of you! A special shout-out goes out to Anne, Marjan and Sabine (the superpower core of *cel organisatie*), Marijke, Lien, Katrijn and all the other *anciens* (oh, we shared all the joys and sorrows of life at that lunch table), Daniël (you were actually the first person ever to explain basic principles of IRT and multilevel modeling to me, dr. D, and I am so very happy we still keep in touch – be it without discussing IRT or multilevel modeling *ever*), Isabel (thank you so much for being the Thelma to my Louise, or the Louise to my Thelma – I forget who's who, but we are a dynamic duo nonetheless), and Margo (it was lovely sitting next to you on the PhD rollercoaster, taking the lows with the highs along the way).

Finally, I am grateful to the people who helped take care of the organizational aspects of preparing a dissertation and a defense. Claudia, Lieve, Tinne, Caitlin, Ingrid, Marc and many others: your help and patience is crucial, does not go unnoticed, and, rest assured, is very much appreciated!

Always time for good conversation
There's an ear for what you say
(Creedence Clearwater Revival, *Up Around the Bend*, 1970)

In the narrative of this dissertation, sensemaking is a central topic. One of the notions I explored, is that the ideas and beliefs that you have and that you use to make sense of the world around you, are shaped to a great extent by your social context. As I was reading and writing about this, I often thought about my own social context, my circles, the people that I am connected to in my personal life – and it made me appreciate them even more (if that was even possible).

Nathalie, Kim, Els and Antonique, thank you for providing a firm foundation, a bedrock. I would be lost without it for sure. Els, we share a special connection – you are an amazing person and I am so proud that you made me Janne's godmother. (Janne, honey, may you grow up to be a happy, confident and well-grounded individual. But do not lose the 'diva'. *Who run the world?* That's right. *Girls.*) Many thanks as well to the very wonderful *Ronquières* gang, to Kim and Jeroen and the whole extended *Fakbar Letteren framily,* and of course to *de mannen (m/v) van de Raaskalderij*. You all make the world a better, funnier, friendlier place. My profound gratitude also goes out to Sam, Maud and Michelle, who always had an ear to lend when times were tough. It made all the difference.

Mama, thank you for everything. Thank you in particular for listening to my worries, for celebrating my victories with me, and for providing ample opportunity to talk about stuff that has absolutely nothing to do with the PhD. I can imagine it is a little weird to see your kid graduating (again) when she is already a fortysomething, but hey, age is just a number, right? I am sure you will agree. I am also sure you share my regret that papa is no longer with us to celebrate this special occasion (the photography metaphors in this dissertation are for you, dad) and that neither is *bonne* (the matriarch who gifted us both with a kind heart and a no-nonsense attitude). Let us not mourn them today, but cherish the fond memories instead.

Finally, I also want to thank my in-laws *de Koxjes* for their support. Thank you for the fun get-togethers, for the passionate discussions at times, for the joy of seeing Ilias and Leander grow up. Those things are the spice of life. And thank you for making Tom, and making him into the person that he is today.

It's gonna take a lot to drag me away from you –
There's nothing that a hundred men or more could ever do
(Toto, *Africa*, 1982)

Tom, baby, we have been together for over twenty-one years now, and Lord knows you had front row seats when I was climbing the PhD mountain. In the words of Goethe, you saw Evelyn *himmelhoch jauchzend* and Evelyn *zum Tode betrübt*. Thank you for not taking offense when I wore the same track pants for, well, let's say, multiple days in a row, for doing housework when I was lost in a different universe, and for sternly steering me away from my desk whenever things got a little too crazy. You are my best mate, my partner in crime, my number one crush. I love that we laugh every single day. I think I want to keep that up for a couple more decades – in fact, I am sure, *as sure as Kilimanjaro rises like Olympus above the Serengeti*. Thank you for being you.

I know we can make it
I know darn well, we can work it out
(The Pointer Sisters, *Yes We Can Can,* 1973)

A final note. There are, and have been, many magnificent people in my life to whom I could dedicate this body of work. But I choose to dedicate it first and foremost to all my fellow overthinkers and second-guessers out there. If you are reading this: Do it anyway. Seize the opportunity. Take the leap. Sure, you might fall... but as the saying goes: what if you fly?

Table of contents

Summary .... **i**

Samenvatting .... **iii**

Acknowledgments | Dankwoord .... **v**

Table of contents .... **xi**

List of tables .... **xiii**

List of figures .... **xiii**

General introduction **1**

1 Theoretical framework and central concepts 3

2 Research context 6

3 Project setup and research design 12

Study 1 Teachers' and school leaders' sensemaking of formal achievement data: A conceptual review **19**

1 Introduction 21

2 Making sense of sensemaking 23

3 Methodology 27

4 Teachers' and school leaders' sensemaking of formal achievement data: towards an integrated conceptual framework 35

5 Framework and applicability 47

6 Conclusion and Discussion 50

Study 2 Principals' and teachers' comprehension of school performance feedback reports. Exploring misconceptions from a user validity perspective **53**

1 Introduction 55

2 Theoretical framework 56

3 Research context and case 59

4 Methodology 60

5 Findings 64

6 Conclusion and discussion 69

Study 3 Feathers in our cap? Mapping educational professionals' internal and external attributions of school performance feedback 75

1 Introduction 77

2 Theoretical framework...78

3 Methodology...85

4 Findings...94

5 Conclusions and discussion 102

Study 4 The interplay of user beliefs and situated characteristics in explaining school performance feedback use...109

1 Introduction 111

2 Theoretical framework...114

3 Method 118

4 Findings...122

5 Discussion and conclusion..127

General discussion 131

1 Main findings 133

2 Strengths and significance 140

3 Limitations and suggestions for further research.................144

4 Recommendations for policy and practice...........................149

References 151

Appendices 173

Appendix A - Attainment targets Spatial use, Traffic and Mobility 175

Appendix B - SPF report elements with annotation (Study 2)....................176

Appendix C - Regression parameters of the path model (Study 4) 184

Appendix D - Covariance parameters of the path model (Study 4)............185

Author contributions...187

List of tables

Table 1. Outline of school performance feedback reports 11
Table 2. Database queries 29
Table 3. Selected studies 31
Table 4. Participants 62
Table 5. Interpretive scheme for assessing conceptual comprehension 63
Table 6. Participants 87
Table 7. Thematic coding scheme including examples and frequency counts 90
Table 8. Overall prevalence of attributions in the dataset (frequency counts) 97
Table 9. Intensity on participant level 98
Table 10. Overview of the survey scales 121
Table 11. Descriptive statistics of the survey scales 122
Table 12. Examples of problematic aspects and misinterpretations pertaining to the table 178
Table 13. Examples of problematic aspects and misinterpretations pertaining to the caterpillar plots 182

List of figures

Figure 1. Theoretical (left) and empirical component (right) of the doctoral project on educational professionals' use of school performance feedback (SPF) 13
Figure 2. Overview of the four studies in this dissertation 17
Figure 3. Framework for teachers' and school leaders' sensemaking of formal achievement data 49
Figure 4. Conceptual model.. 114
Figure 5. Path model 126
Figure 6. Simplified representation of conceptual transfer in school performance feedback (SPF) reports... 134
Figure 7. Report table: Reaching the attainment targets 177
Figure 8. Caterpillar plot positioning the school's actual and expected score 181
Figure 9. Caterpillar plot positioning the school's value-added.......................... 181

1

General introduction

Practitioners, policymakers and scholars increasingly attest to the importance of data use in education. Educational professionals are encouraged to fully exploit all information sources available to them when shaping their policy and practice, in order to improve student achievement. However, in the literature on data-based or data-driven decision making it has been established that the data themselves do not necessarily drive (C. Brown et al., 2017; Dowd, 2005; Lockton et al., 2019). On the road to data-based school improvement, it is not enough to make high quality data available to schools (Hulpia & Valcke, 2004; Schildkamp & Kuiper, 2010). It also is crucial to understand what will activate efficient data use (Schildkamp et al., 2017) and how influencing factors are interlinked (Coburn & Turner, 2011). Moreover, it is necessary to unpack users' (interpretive) processes as they unfold in reality (Coburn & Turner, 2012; Mandinach & Schildkamp, 2021a; Schildkamp, 2019; Spillane, 2012).

In this dissertation, we zoom in on a number of ways teachers and school leaders engage with one specific type of data: formal achievement data from external standardized assessments. In order to better understand how this 'engagement' takes shape and how this type of data can contribute to schools' developmental goals, we investigate how Flemish educational professionals make sense of school performance feedback from low-stakes national assessments, and how they make use of these data for school improvement. We also explore conditions and factors that foster or complicate these processes.

The general aims that motivated us to undertake this project, are academic and, in second order, operational in nature. First and foremost, we intended to make a contribution to the international knowledge base on data-based decision making in education. Additionally, we aimed to gather insights for future developments and implementations of school performance feedback in the Flemish context, specifically with regard to the national assessments. In order to inform research, policy and practice, empirical findings and theoretical insights were gathered throughout four separate but interrelated studies.

In the next sections of this introductory chapter, we first set the scene by briefly discussing a number of theoretical cornerstones. Next, we provide background information about the research context we operated in. Subsequently, we elaborate on the genesis and the design of the research project. Each of the four different studies presented in the following chapters is introduced with a short preview.

1 Theoretical framework and central concepts

Data-based decision making in education

Data-driven decision making "pertains to the systematic collection, analysis, examination, and interpretation of data to inform practice and policy in educational settings" (Mandinach, 2012, p. 71). Over the years, the term has increasingly been replaced with terms like data-based decision making and data-informed decision making, in order to acknowledge that (quantitative) data cannot constitute, and should not constitute, the sole basis for decisions (C. Brown et al., 2017; Rankin, 2016; Schildkamp, 2019). Rather: data can be a tool, a cue, a piece of the puzzle.

In order for data to effectively inform actions and decisions in schools, a number of prerequisites have to be met (Hoogland et al., 2016). Those prerequisites pertain to the very nature of effective data use, for one. Data use has a beneficial impact on the quality of educational decisions, provided that it is goal-oriented and systematic (Rossi et al., 2004; Schildkamp & Lai, 2013a).

A fair number of theories of action on data-based decision making have been established and discussed over the past decades (Coburn & Turner, 2011; Groß Ophoff et al., 2023; Gummer, 2021; Ikemoto & Marsh, 2007; Mandinach et al., 2008; Mandinach & Schildkamp, 2021a; Marsh, 2012; Marsh et al., 2006; Schildkamp, 2019; Schildkamp & Poortman, 2015). Data-based decision making is generally described as an iterative, cyclical process that originates in goal setting. While this goal is generally formulated in terms of improving student achievement, data can be used for accountability, improvement or instructional purposes (C. Brown et al., 2017). Concrete steps include collecting data, analyzing and interpreting data, formulating improvement actions, and evaluating those actions. At the heart of this process is the transfer of raw data into information, and subsequently into actionable knowledge (Mandinach, 2012). It is implied that data users need to understand the data they are presented with, interpret it within their own context, hold it up to the light of the goals they are pursuing, and decide whether they deem it necessary to act upon the wisdom it has brought them (and if so, *how* they will act and *why*).

So, data-based decision making is not a one-shot, quick-fire or quick-fix task to be "checked off", but a complex process. This complexity is further underscored by the vast array of potential barriers and enablers to data-based decision making that have been identified in research (Schildkamp et al., 2014), including organizational properties of schools, personal attributes of data users within those schools, contextual features and qualities of the data (systems) themselves (e.g., Bolhuis et al., 2016; Groß Ophoff et al., 2023; Hoogland et al., 2016; Schildkamp et al., 2017;

Schildkamp & Poortman, 2015; Van Gasse et al., 2015; Verhaeghe et al., 2010; Visscher & Coe, 2003).

Sensemaking

Theoretical insights on data-based decision making are gaining ever more ground, and in parallel, empirical research has described worked examples of data use processes in numerous contexts. Yet, variability in effectiveness of data-based decision making (C. Brown et al., 2017) remains unexplained to a certain extent. For instance, data users have been found to lack sufficient knowledge, skills and (self) efficacy for processing (certain types of) data (Chick & Pierce, 2013; Datnow & Hubbard, 2016; Hellrung & Hartig, 2013). However, more insight is needed on how to build this capacity. Additionally, we are seeing that data use is not necessarily a rational process (Bertrand & Marsh, 2015; Vanlommel et al., 2017). Yet, we don't yet fully grasp intuitive and belief-driven mechanisms at play. In order to further advance the field, there is an active call for more research on data use in practice (Coburn & Turner, 2012; Spillane, 2012), so that we may better understand how these processes manifest themselves in real life and how data users 'of flesh and blood' can be maximally empowered to make sound decisions based on data (Mandinach & Schildkamp, 2021a; Schildkamp, 2019).

From this perspective, sensemaking (also called meaning-making: the process of "going beyond the numbers and their statistical properties", Mandinach, 2012, p. 73) has become a central theme in recent data use research (Schildkamp, 2019). Sensemaking has been conceptualized in other theoretical traditions that have tangents with educational research (Gummer, 2021; Penuel & Shepard, 2016) such as organizational studies, research on naturalistic decision making, and studies on reform an innovation, as a "deliberate effort to understand events, […] triggered by unexpected changes or other surprises that make us doubt our prior understanding" (Klein et al., 2007) and as a way of "structuring the unknown" (Weick, 1995). In the context of data-based decision making, a sensemaking logic posits that the meaning of data is not given but constructed by data users, and it accounts for the fact that users' personal lenses and their context strongly impact how data use takes shape (Bertrand & Marsh, 2015; Coburn & Talbert, 2006; Datnow et al., 2012; Farrell & Marsh, 2016; Ikemoto & Marsh, 2007; Schildkamp, 2019; Spillane, 2012).

School performance feedback systems

In this dissertation, we focus on the use of school performance feedback: formal data about a school's functioning, collected by an external party, and confidentially fed back to the school for self-evaluation, with the explicit intention to inform the school's decision making process and provide input for school improvement (Coe & Visscher, 2002b; Hellrung & Hartig, 2013; Hulpia & Valcke, 2004; Schildkamp & Teddlie, 2008; Visscher & Coe, 2003). Typically, school performance feedback contains some sort of measurement of student achievement, but it can also feature other (nonacademic) outcomes or information about school processes (Coe & Visscher, 2002b). Examples of school performance feedback systems range from designated self-evaluation tools, over pupil monitoring systems, to (inter)national assessment programs and central examinations (Verhaeghe et al., 2015). Educational research projects offering feedback to participants on school- or student-level are considered school performance feedback systems as well (Verhaeghe et al., 2015).

The assumption underlying school performance feedback systems is that educational professionals can and will use the data as a mirror in order to identify strengths and weaknesses (Coe & Visscher, 2002b; Hulpia & Valcke, 2004; Schildkamp & Teddlie, 2008). In practice, however, the data often remain underused, or they are used in ways that do not align with the intentions of those who develop or mandate the tests or assessments (Mandinach & Gummer, 2013; Spillane, 2012; Visscher & Coe, 2003). This may be caused by strategic considerations, for instance, but also by a lack of capacity to turn data into actionable knowledge (Coe & Visscher, 2002a; Datnow & Hubbard, 2016; Hellrung & Hartig, 2013; Mandinach & Gummer, 2016). However, it can also result from the data not being sufficiently geared to recipients' information needs, or from the provider and the user essentially speaking different "languages" (Breiter & Light, 2006; Gunnulfsen, 2017; Hopster-den Otter et al., 2017; O'Leary et al., 2017).

So, while school performance feedback systems may have the theoretical potential to empower educational professionals with robust information and offer them a unique perspective on student outcomes, it is clear that they do not always realize that potential. If we want to develop school performance feedback systems that educational professionals *want* to use and are *able* to use in a valid manner, we need to learn more about actual feedback use in schools, about the usability of systems, and about the way users construct an understanding of the data they are provided with (Coe & Visscher, 2002a; Hellrung & Hartig, 2013). Only then can we really provide data that aligns with the needs and capacities of intended audiences, i.e., data that will contribute to school improvement.

2 Research context

Primary and secondary education in Flanders

This research is based in Flanders, the northern, Dutch-speaking region of Belgium. In this first subsection, we provide some general background information for readers who are not familiar with the Flemish education system.

Flemish children aged 5 to 18 are subject to compulsory education (Eurydice, 2023). Pre-primary education, while not mandatory, can be attended by children aged 2,5 to 6. Mainstream[1] primary education covers six grades and is aimed at children from 6 to 12 years old. Mainstream secondary education comprises six grades as well, divided into three stages of two grades each. The first stage of secondary school (typical ages 12-14) has an A-stream and a B-stream, the latter intended for students who have not received a certificate from primary school and/or want to pursue vocational training. The second stage (typical ages 14-16) and third stage (typical ages 16-18) are organized according to different tracks and programs within which students can select specific study areas (Eurydice, 2023; Nusche et al., 2015; Vlaams Ministerie van Onderwijs en Vorming, 2008). In order to be able to work as a teacher in Flanders, both a content-specific diploma and a pedagogical qualification are required. In most cases these qualifications are integrated into one degree (e.g. primary school teacher training) (Vlaams Ministerie van Onderwijs en Vorming, n.d.-d).

Freedom of education is a right enshrined in the Belgian constitution and it is a leading principle in the Flemish educational system. Flemish education is highly decentralized, and primary and secondary schools enjoy great autonomy in terms of shaping their pedagogical project, appointing staff and awarding certificates and diplomas (Nusche et al., 2015; Vanlommel, 2022). In order to receive funding from the Flemish community, however, educational institutions need to be officially recognized. In practice, officially recognized schools (which make up the large majority of Flemish schools) are organized into three educational networks. These networks are associations of governing bodies (i.e., school boards) and unite one or more umbrella organizations. The networks and umbrella organizations provide representation, offer pedagogical counseling and professional development, develop curricula and timetables, et cetera (Eurydice, 2023; Nusche et al., 2015).

The Flemish Inspectorate of Education monitors whether schools comply with regulations, pay sufficient attention to internal quality assurance, and work towards reaching the attainment targets and developmental goals that the government

[1] Apart from mainstream education, there are also primary and secondary schools that offer special needs education. Furthermore, in secondary education, part-time education can be combined with workplace learning for students from the age of 15 onwards, on the condition that the student has completed the first stage of secondary education.

formulates for different levels and programs of primary and secondary education. The Flemish attainment targets are minimum goals set for different subjects and learning areas, in order to ensure a minimal desired degree of educational quality (Nusche et al., 2015; Vlaams Ministerie van Onderwijs en Vorming, n.d.-b).

Standardized assessment in Flanders

From 2002 onwards, the Flemish government has commissioned the Policy Research Center for Test Development and Assessments STEP to periodically conduct large-scale assessments in a range of subjects and domains. STEP's research aims are, on the one hand, to determine whether attainment targets are met in mainstream primary and in secondary education. On the other hand, STEP investigates whether there are systematic differences between groups of students and between schools in terms of reaching the attainment targets, and examines which school-, class- and student-level variables are associated with better or poorer performance (e.g., Steunpunt Toetsontwikkeling en Peilingen & Vlaams Ministerie van Onderwijs en Vorming, 2020). The center is led by researchers from KU Leuven who, since 2018, structurally collaborate with researchers from the University of Antwerp.

STEP's main assignment is rooted in an educational measurement and educational effectiveness paradigm. Snapshot data are provided for quality monitoring on system-level by developing standardized assessments and questionnaires, and administering these to representative samples of Flemish schools and students with proctors present. Data analyses are based on Item Response Theory, standard setting by way of the Bookmark procedure (Mitzel et al., 2001), and multilevel contextualized achievement modeling. In addition to its system-level objective, STEP also has a school-level objective: to provide individual participating schools with feedback about their performance in order to inform self-evaluation and internal quality assurance. The team also develops parallel tests, i.e., tests that are equivalent to those administered in the national assessments in terms of content and difficulty. Schools can take these parallel tests voluntarily and free of charge (Steunpunt Toetsontwikkeling en Peilingen, n.d.-a, n.d.-b; Vanlommel, 2022).

Flemish national assessments and parallel tests are highly standardized in terms of content, administration and scoring (American Educational Research Association et al., 2014; Education Resources Information Center, n.d.), but, unlike in government-mandated standardized testing in many other educational contexts, participation holds no stakes for schools and students. School performance feedback reports are strictly confidential and are only imparted onto the participating school. Individual schools' and students' results are never made public, nor are they communicated to the commissioning government bodies, to the inspectorate, or to policymakers.

In 2022, the last Flemish national assessments were conducted, and the parallel tests started to be phased out. From 2024 onwards, another policy research center will start administering central tests for mathematics and Dutch to all Flemish students at the end of the fourth and sixth grades of primary school, and the second and sixth grades of secondary school. This will provide information about the extent to which *all* students in Flanders reach educational goals, and the resulting school performance feedback is intended to support internal quality assurance and decision-making in schools system-wide. The tests will provide school- and student-level data. Student results can be taken into account by the school team when evaluating individual students, but they are not to be used as the sole criterion for evaluation (Steunpunt Centrale Toetsen in Onderwijs, 2022; Vlaams Ministerie van Onderwijs en Vorming, n.d.-c, 2023).

In part, the introduction of these central tests was motivated by Flanders' steadily declining results on national and particularly international large-scale assessments (Steunpunt Centrale Toetsen in Onderwijs, 2022). Flanders regularly participates in international comparative studies such as PISA, TIMSS, PIRLS (Vlaams Ministerie van Onderwijs en Vorming, n.d.-a). Here as well, sample schools receive feedback after participation.

Finally, two umbrella organizations offer their own annual tests to measure student achievement, be it only in primary education: Katholiek Onderwijs Vlaanderen and OVSG (Onderwijsvereniging van Vlaamse Steden en Gemeenten). These tests are based on the organizations' curricula, i.e., their 'translation' of the attainment targets. Feedback is focused primarily on the school-level, and includes a cohort-based comparison with similar schools (in terms of student population) that took the same test during the administration period, but also comprises student-level results. These tests have become part of the Flemish assessment canon and reach about 90 percent of Flemish primary schools, including schools that are part of GO! (Gemeenschapsonderwijs) (Janssen et al., 2017).

Data use by Flemish educational professionals

Flemish educational policy documents express expectations towards schools in terms of using (performance) data for internal quality assurance (Dierick et al., 2021; Vlaams Ministerie van Onderwijs en Vorming & Onderwijsinspectie, 2016a, 2016b, 2019, 2020, 2021, 2022) but there are few formal obligations in place (Vanhoof et al., 2012). In recent years, a decree has been implemented that requires primary schools to make use of validated tests for different domains at the end of the sixth grade (Vanlommel, 2022; Vlaams Ministerie van Onderwijs en Vorming, 2017). According to their own preferences and goals, they can choose parallel tests, or standardized tests offered by the umbrella organizations (Janssen et al., 2017). However, it is not stipulated how schools should make use of the resulting school performance data: it is the individual schools' responsibility to determine an appropriate approach (Vanhoof et al., 2012; Vanlommel, 2022; Vlaams Ministerie van Onderwijs en Vorming, 2017).

Research has shown that Flemish educational professionals rely heavily on their own expertise, and on ad hoc, individual and intuitive decision making in order to shape their policy and practice (Van Gasse et al., 2015; Vanlommel, 2022). Although Flemish educators tend to 'test' a lot, in terms of classroom assessment, they generally make limited use of external output data, benchmarks, feedback from scientific studies, and feedback from existing standardized tests. There are substantial differences between schools in terms of data use and perceptions about data use (Van Gasse et al., 2015; Vanhoof et al., 2012; Vanlommel, 2022; Vanlommel et al., 2016). Overall, educational professionals in primary education perceive a stronger expectation to engage in data use compared to their colleagues in secondary education (Van Gasse et al., 2015). This might be explained, in part, by the fact that the tests from the umbrella organizations, which are only available for primary education, have a long standing tradition of being administered (Janssen et al., 2017). Furthermore, there are demonstrable differences according to work role. School leaders tend to engage more with data as opposed to teachers, feel somewhat more confident in the abilities to do so, perceive stronger expectations to use data, and generally express stronger information needs (Van Gasse et al., 2015; Verhaeghe et al., 2010).

Limited (performance) data use by Flemish educational professionals has been interpreted in light of mutually reinforcing factors on individual, collective and contextual levels. Teachers and school leaders (feel they) lack the necessary knowledge and skills to interpret performance data and they perceive a lack of time to do so (Van Gasse et al., 2015; Vanhoof et al., 2013). Additionally, collaboration in data use is still limited in most school teams (Van Gasse et al., 2015, 2016, 2017; Vanlommel, 2022). Furthermore, available resources are perceived as complicated, insufficiently user-friendly, and insufficiently geared towards existing information needs (Vanhoof et al., 2013; Verhaeghe et al., 2010). Finally, as described above, the Flemish educational context is characterized by high degrees of decentralization and autonomy, and a focus on school development over accountability. Traditionally, this has gone hand in hand with a resistance to (overreliance on) standardized measures from external sources, which has perhaps resulted in a general attitude of – to put it colloquially – "if it ain't broke, don't fix it" (Penninckx et al., 2017; Vanlommel, 2022).

Research case:
school performance feedback from Flemish national assessments and parallel tests

STEP provides tailored, confidential school performance feedback reports to the schools that were part of the representative sample tested in a national assessment, and to schools taking parallel tests (cf. supra). The reports contain results for each written test[2] that was administered in the school for a specific cluster of attainment targets (e.g., a ratio and scale test in a mathematics assessment, or a test on reading comprehension in an assessment of French as a foreign language). Consistent with the most prevalent frames of reference in external standardized assessments (American Educational Research Association et al., 2014; Hellrung & Hartig, 2013), the school's results have a criterion-referenced and a norm-referenced component. The criterion-referenced component of the results pertains to the extent to which attainment targets were reached in the school (i.e., how many students surpassed the cutoff, and to what extent they did or did not). The norm-referenced component of the feedback depicts how the schools' performance relates to the average national results, and how it compares to that of other schools with a similar student population. This correction of school performance for input characteristics is also presented as 'value added' information[3], intended as a measure for fair comparison (Visscher & Coe, 2003) in order to 'compare like with like' (Verhaeghe et al., 2015).

The general structure of the reports is outlined in Table 1. The feedback focuses on school-level results, but criterion-referenced information is also presented per class if applicable[4]. In secondary education results are also included per participating study option. Note that in the parallel test reports, the reference group is (also) the representative sample of schools that participated in the corresponding national assessment.

The doctoral research project presented in this dissertation ran parallel to a STEP user study about school performance feedback, initiated in 2018. Quantitative and qualitative findings, collected and analyzed by way of a multi-method approach (Tashakkori & Teddlie, 2002) were intended to contribute, in part, to future developments of school performance feedback. As discussed earlier, Flemish educational professionals make little use of available school performance feedback data. The STEP user study was set up in order to inform policymakers and feedback developers about users' genuine perceptions about the school performance feedback that is offered, and to explore their (unanswered) information needs. By examining in what manner and to what extent users (actually) engage with the feedback reports, the goal was to find ways in stimulating and supporting users to (better) use such data for internal quality assurance. In the meantime, STEP has been discontinued (cf.

[2] Some national assessments also include performance assessments of practical skills (e.g., oral proficiency in a language assessment, or following a step-by-step plan in a technology assessment). For these performance assessments, (descriptive) results are provided only on system-level.

[3] The term *value added* (*toegevoegde waarde* in Dutch) is used in the STEP feedback reports to express the contribution that schools make to student achievement after correction for input variables. It is a contextualized snapshot, based on a one-time measurement. Note that this interpretation differs from strict interpretations of value added. Value added modeling generally implies a correction for students' prior achievement (growth modeling in order to measure learning gain) in addition to – of even before – correcting for background characteristics (measuring net progress) (American Educational Research Association et al., 2014; Janssens et al., 2014; Levy et al., 2019; Rowe et al., 2002).

[4] In recent years, an insert with limited student-level results has also been added to the parallel test feedback (Steunpunt Toetsontwikkeling en Peilingen, n.d.-a).

supra), but especially since Flanders is on the brink of implementing central tests, research findings will continue to be relevant in a practical sense.

Table 1. Outline of school performance feedback reports

Report section	Content	
Introduction	General information about the Flemish national assessment program and parallel tests	
	Information about the assessment of the subject at hand, including the national results	
Interpretive guide	Detailed overview of the structure of the results chapters	
	Explanation of how the schools' results were calculated	
	Clarification of statistical concepts and graphical representations	
	Guidelines for using the results, including where to turn to for support	
School results	Overview of the tests taken in the school	
	Per test:	
	General information	
	Attainment targets tested	
	Benchmark (assessment sample) size	
	School-specific information	
	Overview of participating classes and number of students	
	School results	
	Table [a]	Distribution and mean of ability scores, juxtaposed to the mean and distribution in the national sample
		Percentage of students reaching the attainment targets, juxtaposed to the percentage recorded nationally
	Caterpillar plots [b]	Plot highlighting the school's raw mean score and expected mean score after correction for input and context factors (against the full sample's mean scores)
		Plot highlighting the school's added value (against the full sample's added values)

Notes.

[a] The table presents results on the level of the school, the level of study option (only in secondary education), and the level of the class. An example is included in Appendix B (an appendix to Study 2).

[b] Caterpillar plots are offered on the level of the school, and per study option (only in secondary education). Examples are included in Appendix B (an appendix to Study 2).

3 Project setup and research design

Objectives

Against the backdrop of crucial theoretical and contextual insights (as sketched in the preceding sections) the academic and operational aims that motivated our research endeavor (cf. supra) crystallized into three concrete research objectives. On a descriptive level, we wanted to explore whether and how educational professionals (physically) make use of school performance feedback for school improvement (Objective 1). It is assumed that schools and users within schools engage with the school performance feedback that is distributed to them, but is that the case? How does this take shape? In parallel, we wanted to gain insight into how educational professionals make sense of school performance feedback (Objective 2). Providers of school performance feedback relay information to feedback recipients, but how do the latter construct a message from the data they receive? Apart from merely describing these mechanisms, we also wanted to gain a deeper understanding of enablers and barriers. After all, we know that making use and making sense of school performance feedback is not self-evident. Therefore, on an explanatory level, we aspired to identify conditions and factors that foster or complicate educational professionals' sensemaking and use of school performance feedback (Objective 3).

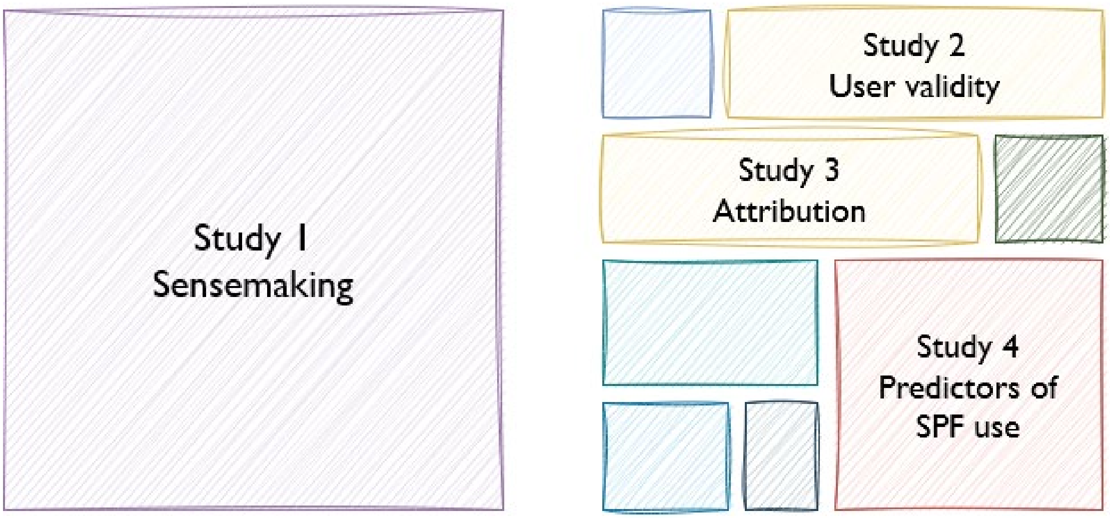
Building blocks

Guided by our research objectives, we conducted four studies. A conceptual exploration of 'sensemaking' (Study 1) was undertaken in order to shed light on the kaleidoscope that is real-life data use and data literacy. Given the nature of our aims and objectives, we focused primarily on educational professionals' use of formal achievement data such as school performance feedback. Furthermore, qualitative data (Study 2 & 3) and quantitative data (Study 4) were collected in the Flemish context[5] in order to broaden and deepen our understanding of the ways educational professionals engage with school performance feedback in practice, by zooming in on a number of specific aspects of sensemaking and data use. As such, this doctoral research project has a theoretical and an empirical component, as shown in Figure 1. While both are strongly interrelated, each component served as a lens to study and better understand (aspects of) teachers' and school leaders' engagement with school performance feedback. By including empty blocks in the part of the figure that

[5] Data collection took place before and during the COVID19-pandemic, but pertained entirely to school performance feedback for assessments and tests that were *administered at least six months before the start of the pandemic* and earlier. Furthermore, the Flemish attainment targets are currently gradually being updated, starting in secondary education (Vlaams Ministerie van Onderwijs en Vorming, n.d.-b). During data collection for the present dissertation, however, all former/existing attainment targets were still in place at the educational levels concerned.

represents the empirical component, we want to make it clear that we are focusing on a number of specific aspects and processes. In order to capture the whole breadth of 'sensemaking' and 'feedback use', more research is obviously needed.

Figure 1. Theoretical (left) and empirical component (right) of the doctoral project on educational professionals' use of school performance feedback (SPF)



Main theoretical foundations

In order to investigate processes and (pre)conditions that play a role in educational professionals' use of school performance feedback, all four studies were embedded in the existing knowledge base. First of all, we explored insights on data-based decision making in general (e.g., Beck & Nunnaley, 2021; Bertrand & Marsh, 2015; Mandinach, 2012; Schildkamp, 2019; Schildkamp & Poortman, 2015; Van Gasse et al., 2015; Vanlommel et al., 2016). In the studies, we expand upon well-known process models describing the motions and emotions related to data use (e.g., Mandinach et al., 2008; Marsh, 2012; Marsh et al., 2006; Schildkamp & Poortman, 2015) and on previously identified enablers and barriers that play a role in data use by educational professionals (e.g., Bolhuis et al., 2016; Schildkamp et al., 2014).

As previously explained, contemporary views on data-based decision making urge researchers to investigate data use in practice in order to understand what it entails and how it can be improved. This includes the need to frame and examine data use as an act of sensemaking. Therefore, in all four studies, we have emphasized the need to take on this sensemaking perspective (cf. Bertrand & Marsh, 2015; Coburn, 2001;

Datnow et al., 2012; Schildkamp, 2019; Spillane et al., 2002). In Study 1 we explored the sensemaking perspective in depth, in order to ground our own research and inspire research that might follow in our footsteps. In this exercise we also explored how existing insights and findings fit in with sensemaking research from other scholarly fields (Klein et al., 2006b, 2006a; Maitlis & Christianson, 2014; Weick, 1995; Weick et al., 2005).

Given our objectives, we took specific guidance from international as well as local research on the design, implementation and use of school performance feedback systems in particular (e.g., Coe & Visscher, 2002a; Schildkamp & Teddlie, 2008; Vanhoof et al., 2011, 2012; Verhaeghe, 2011; Verhaeghe et al., 2015; Verhaeghe et al., 2010; Visscher & Coe, 2003). In Study 2, we specifically zoomed in on (user) validity issues relating to score reporting practices and users' actual comprehension of reports (Hattie, 2009; Kane, 2013b, 2013a; MacIver et al., 2014; O'Leary et al., 2017).

In Studies 3 and 4, we applied frameworks from social and cognitive psychology. Attribution theory served to examine the causal ascriptions that educational professionals make when interpreting school performance feedback (Wang & Hall, 2018; Weiner, 1985, 2010). The Theory of Planned Behavior served as a framework to examine the behavioral quality of school-level feedback use and the way user beliefs relate to contextual factors in steering this behavior (Ajzen, 1991, 2011; Prenger & Schildkamp, 2018). Throughout the studies, we also attended to factors that are presumed to influence the use of school performance feedback, but tend to remain understudied, such as feedback valence and voluntariness in feedback pursuit.

Overview of the four studies

Figure 2 gives a concise overview of the four separate studies[6] of the dissertation, including the rationale behind each of them, the guiding theoretical frameworks and the approaches chosen. All four studies depart from the notion that achievement data such as school performance feedback from standardized assessments *can* inform school improvement, but that we need more insight into the mechanisms of engagement in order to be able to identify challenges and opportunities.

Study 1 - Sensemaking of formal achievement data

Whether data such as school performance feedback are used effectively and as intended, depends on how local actors engage with the data in their daily practice, from their own subjective backgrounds and within their own contexts. In order to illuminate these processes and identify influencing factors, data use researchers are increasingly adopting a sensemaking perspective. Other than the technical-rational

[6] Chronologically, Study 4 was the first study we conducted. The need to undertake a literature study about educational professionals' sensemaking of formal achievement data, presented here as Study 1, crystallized as we were working on Study 4 and preparing for the qualitative inquiries for Studies 2 and 3. In this dissertation, we elect to present the studies in a logical rather than chronological order.

perspectives that underlie many established theoretical models of data-based decision making in education, a sensemaking perspective puts the data user front and center, rather than the data. For Study 1, we systematically conducted a conceptual review of how sensemaking paradigms and vocabularies are applied in existing studies on educational professionals' use of performance and achievement data. Based on a thematic analysis of 25 empirical and theoretical studies we discuss crucial insights and work towards an integrated conceptual framework.

Study 2 - User validity of school performance feedback

When educational professionals do not understand the feedback data they are provided with, or when they misconstrue a message from what they read into the data, they will not arrive at sound inferences, sensible diagnoses or meaningful decisions further along the line. Since comprehension is key in assuring the user validity of school performance feedback, Study 2 zooms in on interpretations and interpretability of authentic feedback reports. Data were collected by way of 22 semi-structured interviews that also included a think-aloud procedure. We identified misconceptions in school leaders' and teachers' reading of the school performance feedback they received on one focal test[7] after their school took part in the 2019 national assessment of People and Society in the sixth grade of primary school. Misconceptions are critically discussed from an information-processing perspective: in what way do users construct meaning from the tangible representations they are provided with, and does this meaning make sense from a provider's point of view? If not, where lie the stumbling blocks? We take on a sensemaking perspective in order to see whether we can explain misconceptions as stemming from a disconnect between users' and providers' frames of reference.

Study 3 - Attribution of school performance

School performance feedback offers a snapshot of student achievement, but it is its interpretation by local actors that informs subsequent decisions for policy and practice. Formulating a diagnosis by reflecting about causes of favorable or unfavorable outcomes, is an integral part of this process. Based on interview data that were collected from the same participants as for Study 2, Study 3 is about the attributions teachers and school leaders make when interpreting their schools' and classes' results. We looked specifically at the locus of causality of these attributions, and examined patterns according to users' work roles and the favorability of the school performance feedback they received.

Study 4 - Predictors of school performance feedback use

In Study 4 we zoom back out, by looking at what happens when a school performance feedback report comes through a schools' proverbial letterbox. From the idea that

[7] Spatial use, Traffic and Mobility. The cluster of attainment targets that was tested, is included in Appendix A.

data use is done by a person, with their very own perceptions about data and data use, but that they do not operate in isolation, as the context plays a role as well, we investigated the relative impact of belief-driven, user-level and situated, school-level factors, on school-level data use. An online survey was administered to 470 educational professionals in primary and secondary education, who had been presented with a school performance feedback report after participating in a national assessment or administering parallel tests. The Theory of Planned Behavior served as a lens to examine potential drivers of school performance feedback use from a dual perspective.

Structure of the dissertation

The four studies introduced in the previous subsections, form the next four chapters of this dissertation. Please note that each chapter is based on a research paper that has been published or submitted for publication in an academic journal. Since each chapter can be read as a separate unit, a certain degree of overlap between these units is inevitable. In a final chapter, we will summarize our main findings and reflect on their significance and their implications.

Figure 2. Overview of the four studies in this dissertation

Research objectives	**To explore whether and how educational professionals *make use* of school performance feedback for school improvement** **To gain insight into how educational professionals *make sense* of school performance feedback from external standardized assessments** **To identify conditions and factors that *foster or complicate* educational professionals' sensemaking and use of school performance feedback**			
Main theoretical foundations	Data-based decision making in education School performance feedback systems Sensemaking			
	Sensemaking perspectives	User Validity	Attribution theory	Theory of Planned Behavior
Four studies	**Study 1** **Sensemaking of** **formal achievement data**	**Study 2** **User validity of** **school performance feedback**	**Study 3** **Attribution of** **school performance**	**Study 4** **Predictors of** **school performance feedback use**
	Technical-rational perspectives cannot fully account for the fact that real-life data use is not a linear or straightforward process. *More insight is needed into sensemaking mechanisms.*	*Data use starts with analyzing the data.* *Users' analysis of the data determines the validity of their further interpretations.*	*Hypothesizing about potential causes for outcomes is a fundamental part of sensemaking.* *Diagnoses shape subsequent responses.*	*Data use takes shape in the hands of individual actors.* *Data users do not operate in isolation.*
	Which insights are crucial in order to (better) understand educational professionals' sensemaking of formal achievement data?	Do educational professionals comprehend central concepts in school performance feedback? How can we explain misconceptions?	Do educational professionals attribute school performance to internal or external factors? Do users' work roles and feedback favorability play a role?	How do user beliefs and situated characteristics interplay in explaining school-level school performance feedback use?
Methodological approaches	Literature review	Semi-structured interviews Think-aloud procedure Framework Analysis	Semi-structured interviews Framework Analysis Quantitizing	Survey Path analysis
	Conceptual	Qualitative		Quantitative
	Theoretical	Empirical		

Study 1

Teachers' and school leaders' sensemaking of formal achievement data: A conceptual review

BASED ON 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. https://doi.org/10.1002/rev3.3334

ABSTRACT Formal achievement data such as test scores and school performance feedback from standardized assessments can be a powerful tool for data-based decision making and school improvement. However, teachers' and school leaders' usage of these data is not necessarily straightforward or predictable. In order to illuminate how educational professionals engage with data in their daily practice, from their own subjective backgrounds and within their own contexts, data use researchers increasingly adopt a sensemaking perspective. Sensemaking, a theoretical construct grounded in psychological and organizational scholarship, offers a framework and a vocabulary to explain how cues such as educational output data are processed in real-life educational settings. As such, sensemaking research sheds light on reasons why educational professionals' use of these formal achievement data may deviate from normative expectations.

The present study is a conceptual review of how sensemaking is conceived and applied in literature on educational professionals' use of formal achievement data. In total, 25 empirical and theoretical studies were selected and subjected to thematic analysis. Findings include that sensemaking is used as a lens to study data use, as well as a label for interpretive micro-processes of data analysis and interpretation, and that formal achievement data can be regarded as sensemaking resources. An integrated conceptual framework on educational professionals' sensemaking of formal achievement data is presented, including a discussion of critical insights that may inspire future research on data-based decision making in education.

1 Introduction

Educational professionals are increasingly expected to use data in order to inform, shape and strengthen school policy and instructional practice. A host of sources can and should serve data-based decision making (DBDM) in education, ranging from informal data such as classroom observations, over formal (i.e., systematically collected) data such as test scores and information about school composition, to research findings and big data (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). However, teachers and school leaders often struggle to effectively engage with these data. In the past decades, the DBDM research field has been unravelling data use dynamics in order to find ways to address those struggles. Still, in order to truly empower educational professionals as data-based decision makers, a more thorough theoretical understanding is needed (Mandinach & Schildkamp, 2021a; Schildkamp, 2019).

The present study contributes to the DBDM knowledge base by zooming in on educational professionals' engagement with formal data that provide insight into individual student outcomes and school performance – which we will henceforth refer to as "formal achievement data" for short. We particularly want to inform the debate on affordances and challenges related to educational professionals' use of formal achievement data that hail from school-external systems and standardized testing. Well-known examples are state-level or national assessments such as those organized in the USA, and certification examinations such as the UK's GSCE's. However, output data offered by school performance feedback systems that are designated self-evaluation tools or that give achievement-based feedback to schools that participated in a research project also fall into this category (Coe & Visscher, 2002b; Schildkamp & Teddlie, 2008; Verhaeghe et al., 2015).

The formal achievement data provided by external assessments and feedback systems are generally regarded as a powerful resource for school improvement. The assumption is that educational professionals (teachers, school leaders, supporting staff) can and will use the data as a mirror in order to inform subsequent policy and instructional decisions (Coe & Visscher, 2002b; Hulpia & Valcke, 2004; Schildkamp & Teddlie, 2008). In practice, however, such data are often underused or misused (Coe & Visscher, 2002a) because of how stakeholders approach and engage with these data. Variability in data use can be contributed to a great extent to the way stakeholders understand, explain, position and value the data that they have at their disposal, and the way they determine a course of action based on what they take away from the data – for short: the way they *make sense* of the data.

Numerous sensemaking challenges have been identified in DBDM research. For instance, educational professionals may overly rely on their intuition when interpreting results and making decisions based on formal achievement data (Vanlommel et al., 2017; Vanlommel & Schildkamp, 2019) or lack the capacity to understand reports and turn data into actionable information (Mandinach & Gummer, 2016; van der Kleij & Eggen, 2013; Vanhoof et al., 2011). As a result, critical cues may not be picked up on, or worse, inaccurate or invalid inferences may lead to misguided decisions. Misuse or unintended use of formal achievement data can also be a result of conflated purposes, especially in situations where output data (also) serve to formally hold schools accountable for their outcomes in a high-stakes manner (Coe & Visscher, 2002a; Datnow & Park, 2018; Mandinach & Schildkamp, 2021a; Vanhoof & Van Petegem, 2007). Furthermore, while advancements in the past decades have contributed to a growing availability of robust tools and data sources for schools, it has been found that educational professionals are at risk of “drowning in data” (Mandinach & Schildkamp, 2021a; Schildkamp et al., 2014). Effective data use requires substantiated prioritization, interpretation and triangulation, which in turn require sophisticated knowledge and skills. However, it seems like the more snapshots educational professionals receive, the greater the risk of losing sight of the big picture.

In the interest of addressing these challenges and finding ways to support schools in data use for school improvement, DBDM researchers urgently call for more insight into real-life data use sensemaking mechanisms (Schildkamp, 2019). Consequently, sensemaking is gradually becoming a central theme in recent research on DBDM in education. Sensemaking perspectives provide a more human-centered outlook on data use than the technical-rational perspectives that underlie many theoretical models of DBDM. Those models are largely based on the assumptions that information borne in data is somehow set and unequivocal, and that the mere availability of data will enable educators to diagnose problems and guide them towards improvement (Datnow et al., 2012; Farrell & Marsh, 2016; Horn et al., 2015). While such models provide a clear, normative framework and as such a vital baseline for studying data use in education, explanations for the fact that data use in reality is “messy” (Bertrand & Marsh, 2015) need to be sought elsewhere. A sensemaking perspective accounts for the fact that stakeholders’ personal lenses and their context strongly impact how data use takes shape. It acknowledges that DBDM is what happens when people of flesh and blood encounter data, deal with data, and decide how to move forward based on what they take away from these data (Coburn & Talbert, 2006; Datnow et al., 2012; Farrell & Marsh, 2016; Ikemoto & Marsh, 2007; Schildkamp, 2019; Spillane, 2012). A sensemaking perspective fits in with insights that data use in practice is not a linear or straightforward process (Ikemoto & Marsh, 2007; Mandinach & Gummer, 2016; Mandinach & Schildkamp, 2021a) as one of its central tenets is that the meaning of data is not given but constructed by data users (Spillane, 2012).

Because sensemaking is essential in data use for school improvement, and more insight is needed into how users make sense of data in reality (Mandinach & Schildkamp, 2021a; Schildkamp, 2019) we argue that further conceptual exploration of what sensemaking entails would benefit DBDM research. In research accounts on DBDM in education, much like in other areas such as organizational research, work on (human) sensemaking of (environmental) cues is currently proliferating. At the same time, usage of the term 'sensemaking' is diffuse and inspired by different theoretical paradigms. We therefore propose that the field needs to work towards an integrated conceptual framework. The present study contributes to this endeavor by taking stock of how different (DBDM) scholars interpret and apply the concept, and specifically by exploring how sensemaking is conceptualized in research on educational professionals' engagement with formal achievement data. We report on a "systematically conducted conceptual review" (cf. Amundsen & Wilson, 2012, p. 91; Kennedy, 2007) that takes guidance from the following questions: How is sensemaking conceptualized in relation to teachers' and school leaders' sensemaking of formal achievement data? What are central components of sensemaking in this line of research? And (how) can existing insights be combined into an integrated framework for future scholarship?

In order to give direction to our own exploration of what sensemaking means in relation to educational professionals' engagement with formal achievement data, we first set the scene by exploring the theoretical roots of the sensemaking construct. Next, we present our methodological approach for searching and reviewing the literature teachers' and school leaders' sensemaking of formal achievement data. Subsequently, we present the results of this process by drafting an integrated conceptual framework based on the themes we have identified. We conclude with a discussion of how the present study contributes to the knowledge base, and a reflection on potential further advancements.

2 Making sense of sensemaking

Sensemaking is an abstract but semantically rich word in the English language. In scientific research, however, it is not a neutral term but a theoretical construct that is employed in specific ways. Before we set out to explore sensemaking with regard to educational professionals' engagement with formal achievement data, the theoretical framework presented in this section is needed to appreciate the complexity of the sensemaking phenomenon and to establish a sensemaking vocabulary. In the following paragraphs, we first briefly introduce a number of prominent takes on sensemaking. Subsequently we take a bird's-eye view to salient leitmotivs in sensemaking theory and research.

2.1 Sensemaking perspectives

The concept of sensemaking originates in cognitive and social psychology and features in numerous scholarly traditions. Sensemaking is generally characterized as a process people engage in when they find something novel and/or unexpected on their path (Klein et al., 2007; Maitlis & Christianson, 2014; Weick, 1995). They figure out what this means to them, if they need to deal with it and in what way, and how to move forward (Klein et al., 2007; Weick et al., 2005). Throughout, the (social) context of the sensemaker and their prior experiences shape how this process unfolds (Weick, 1995; Weick et al., 2005).

While there is common ground in conceptualizations, there is no one unified ‘sensemaking theory’ but rather a wide range of ‘sensemaking perspectives’ (Maitlis & Christianson, 2014; Sandberg & Tsoukas, 2015; Weick, 1995). In the past decades, such perspectives have particularly thrived in organizational literature (Maitlis & Christianson, 2014). Organizational work by Weick and colleagues (Weick, 1995; Weick et al., 2005) constitutes one of the most influential sensemaking perspectives to date. Another notable take hails from research on naturalistic decision making. Klein and colleagues developed the Data-Frame theory of sensemaking in order to (empirically) study sensemaking as a macro-cognitive process, i.e., a set of mental activities that people perform in complex real-life situations (Klein et al., 2006b, 2007b, 2010). In educational research, the work of Spillane and Coburn on sensemaking may be considered canonical. Both authors apply a sensemaking perspective to educational policy implementation (e.g., Coburn, 2001, 2006; Spillane, Diamond, et al., 2002; Spillane, Reiser, et al., 2002), for instance in the case of instructional reform. Both Coburn and Spillane have also used their frameworks and insights to discuss data use or ‘evidence use’ as an aspect or materialization of policy implementation (e.g., Coburn et al., 2009; Coburn & Talbert, 2006; Coburn & Turner, 2011; Spillane, 2012; Spillane & Miele, 2007). Whereas the work of Spillane and colleagues is geared more towards the cognitive aspects of sensemaking, the work of Coburn et al. is more focused on mechanisms of co-construction (Walls, 2017). Nevertheless, as is the case throughout the sensemaking literature, and given the very nature of sensemaking, interpretive and social or context-related aspects are intertwined in both bodies of work.

2.2 Sensemaking leitmotivs

Sensemaking begins with a sensemaker and is triggered by ambiguity

People engage in sensemaking when they encounter some sort of “disruptive ambiguity” (Weick et al., 2005, p. 413), an interruption that makes them “doubt [their] prior understanding” (Klein et al., 2007, p. 114). Sensemaking then means actively trying to figure out what this interruption means in light of what is known and believed. Small cues can trigger sensemaking just as well as larger cues and disruptions. In fact, in Weick’s conceptualization, people are continually shaping and enacting their own reality by making sense of cues, thereby creating “a more ordered environment from which further cues can be drawn” (Maitlis & Christianson, 2014, p. 67). However, in order for sensemaking to take place, it is not sufficient that a novelty or a surprise is merely present. There has to be a sufficient sense of discrepancy between what one experiences and what one would have expected (Maitlis & Christianson, 2014; Weick, 1995; Weick et al., 2005).

The fact that sensemaking is “prompted by violated expectations” (Maitlis & Christianson, 2014, p. 67) illustrates how sensemaking “begins with a sensemaker” (Weick, 1995, p. 18). Cues or stimuli trigger sensemaking only when they are perceived as triggers. Personal lenses guide attention and determine what is noticed and bracketed as ambiguous (Klein et al., 2007; Maitlis & Christianson, 2014; Weick, 1995) and they also shape the further sensemaking process. Personal lenses explain why different people recognize and notice different things from one and the same event or issue (Klein et al., 2007). Thus, in terms of DBDM, they explain why different data users might come to different conclusions and observations based on the same data or score set.

Klein and colleagues conceptualize these personal lenses as “frames” that reflect “a person’s compiled experiences” (Klein et al., 2007, p. 118) and serve as explanatory structures. Frames consist of knowledge structures (Attfield et al., 2018), such as schemata and mental models, i.e., personal (causal) beliefs about and understanding of how the world works (Klein et al., 2007; Spillane & Miele, 2007). However, personal lenses can also refer to attitudes and interests. Values and goals determine whether a sensemaker is motivated to engage in sensemaking in the first place (Attfield et al., 2018). Furthermore, emotion fuels and shapes sensemaking. Sensemakers need to be “energized” in order to engage in sensemaking (Maitlis et al., 2013). Conceptualizing sensemaking as concerned with identity-construction, Weick finds that identity threat can be a particularly important trigger for sensemaking (Maitlis & Christianson, 2014; Weick, 1995).

Sensemaking is an active search for coherence, aimed at understanding and action

The idea of seeking fit between salient cues from the environment and personal pre-existing frames (Maitlis & Christianson, 2014; Weick, 1995; Weick et al., 2005) is a central tenet in sensemaking conceptualizations, also in the cognitivist work on sensemaking that preceded Weick‘s advancements (e.g., Starbuck & Milliken, 1988). Klein et al.’s Data-Frame theory or Data-Frame model elaborates on this aspect, by characterizing sensemaking as a “process of fitting data into a frame and fitting a frame around the data” (Klein et al., 2007, p. 120) and by zooming in on the deliberate and iterative acts of framing and reframing (Klein et al., 2006b, 2007b). A central proposition is that data are never given, but always constructed. They are “the interpreted signals of events” (Klein et al., 2007, p. 120).

Sensemaking scholars stress that sensemaking may be interpretive, but is not synonymous to interpretation, as it is more active, deliberate and creative (Maitlis & Christianson, 2014; Weick, 1995). In any case, simply connecting data to a frame based on recognition does not constitute sensemaking (Klein et al., 2007). Sensemaking is a process of constructing meaning, forming an understanding, attributing significance (Weick et al., 2005: "what's the story?"), as well as formulating or taking action (Weick et al., 2005: "now what?"). Sensemaking can be purely explanatory, i.e., aimed at abstract understanding (e.g., making a diagnosis, identifying a problem), and/or anticipatory, i.e., aimed at functional understanding (e.g., preparing a scenario for preventing accidents) (Klein et al., 2007, 2010). Ultimately, sensemaking leads to some sort of change, in understanding or behavior, in beliefs or in actions. In the context of DBDM, this duality is reflected, for instance, in that between conceptual and instrumental uses of data.

The Weickian perspective on sensemaking focuses on its retrospective nature (Weick, 1995), i.e., explaining something that *has* occurred by comparing it to *prior* experience. Other accounts debate the prospective elements of sensemaking, precisely because it is aimed at formulating a future course of action (Maitlis & Christianson, 2014; Sandberg & Tsoukas, 2015). However, since it is "more likely to see sense that has already been made than to see the actual making of it" (Weick, 1995, p. 49), Weick proposes that it makes sense to study sensemaking in relation to "prolonged puzzles" in order to unravel what happens during the process (Weick, 1995, p. 49). While Weick characterizes sensemaking as an ongoing dynamic (Weick, 1995), with people continually making sense of their environment by seeking order in chaos, Klein's take on sensemaking is more episodic. In Klein's interpretation, sensemaking does have an endpoint. When a sensemaker arrives at an understanding that they deem satisfactory, the process of framing and reframing ceases (Klein et al., 2007). Theorists nevertheless agree that sensemaking is a search for coherence, and not for an objective truth. The aim is to arrive at the feeling that one has found congruence: a sensible explanation from which to move forward (Klein et al., 2007; Weick, 1995).

Given that sensemaking is "driven by plausibility rather than accuracy" (Weick, 1995), people will be motivated to move on once they feel they have built a satisfactory account. Based on this idea, Weick et al. (2005) propose that (organizational) sensemaking is, at least in part, a skill that can be developed. In order to grow and move forward, people need the drive and the confidence to act upon their interpretations. Furthermore, Klein et al. (2007) propose that expertise in sensemaking is not a question of more sophisticated reasoning, but of having a richer and more elaborate "repertoire of frames" than novices. Given that frames change over time as people gather or encounter more data (Klein et al., 2006b), sensemaking can be "developed through experience and learning through reflection" (Kahneman & Klein, 2009, as interpreted by Cook & Gregory, 2020, p. 11). In any case, sensemaking is not straightforward and can be biased. As people seek fit between cues and frames, it may be hard to determine which frame is the "right" one to explain what is going on. People are inclined to frame new information within what is familiar or expected, and what resonates with them in terms of values and norms (Klein et al., 2007; Spillane, Reiser, et al., 2002; Starbuck & Milliken, 1988).

Sensemaking is individual as well as social, cognitive as well as discursive

Regarding the "ontology of sensemaking", Maitlis and Christianson (2014) point out that it can be studied as something that occurs "in people's heads" as well as "in conversations between people". So, sensemaking is not only a matter of cognition but also of language and discourse. Language is a primary locus of sensemaking, particularly in the Weickian perspective. Here, sensemaking is seen as putting comprehension into words: people draft narrative accounts that enables them to rationalize what they are thinking and doing (Weick et al., 2005). It is constructing and revising a plausible story in which the central questions are what is going on and what to do next (Weick et al., 2005). By emphasizing the discursive nature of (organizational) sensemaking, Weick is credited for placing the sensemaking concept into a social constructivist paradigm (Maitlis & Christianson, 2014; Sandberg & Tsoukas, 2015). Consequently, the Weickian perspective devotes a lot of attention to intersubjectivity and sensegiving mechanisms that act as a precursor to sensemaking (Maitlis & Christianson, 2014; Sandberg & Tsoukas, 2015; Weick, 1995). However, as Weick (1995, p. 40) points out, "even monologues and one-way communications presume an audience".

Regardless of whether sensemaking is studied at the individual, interpretive level, or at the collective level, e.g., in team settings, it is always acknowledged to be a situated and social phenomenon. People derive their identities from social groups they belong to, and every individual has a "parliament of selves" (Mead, 1934, as quoted in Weick, 1995, p. 18). Both individuals and organizations or other types of groupings have histories that shape their beliefs, values, norms and expectations (Spillane, Reiser, et al., 2002). Consequently, cognitive frames can be situated as well (Spillane, Reiser, et al., 2002). Such frames can be general or specific (Sandberg & Tsoukas, 2015). The former refer to frames grounded in sociocultural templates such as corporate/industrial or regional/national, or in ideologies, for instance gender or politics. The latter refer to tacit knowledge and internalized theories of action. Furthermore, sense is always made in situ (Spillane, 2012; Spillane & Miele, 2007). This means that formal and informal routines and tools shape sensemaking interactions between people (Spillane, 2012; Spillane & Miele, 2007). It also means that people negotiate meaning in such interactions, that they co-construct accounts and narratives, and that the nature of interactions determine how sensemaking unfolds. Work roles, leadership, and (organizational) (sub)cultures play a part in shaping sensemaking, as do institutional and political forces, authority relationships and mechanisms of problem framing (Coburn, 2001, 2006; Coburn et al., 2009; Coburn & Talbert, 2006; Spillane, Diamond, et al., 2002).

3 Methodology

Guided by three methodological frameworks (conceptual reviews, scoping studies and theoretical reviews), we conducted a systematic query of two research databases in pursuit of studies that would help us understand educational professionals' engagement with formal achievement data from a sensemaking perspective. We performed a thematic analysis on the selected studies. In the following paragraphs, we discuss the different steps we undertook in more detail. We end this section by giving an overview of the selected studies.

3.1 General approach

Our primary goal was to take stock of how sensemaking is conceptualized and applied in studies on educational professionals' engagement with formal achievement data. We did not set out to answer a narrowed down, empirical research question in order to find evidence for causes and effects for a specific phenomenon. We rather envisioned a fluid yet methodical exercise in mapping out what is of interest to our research field. Therefore, we found a conceptual review (Kennedy, 2007) to be best suited to our purposes. Contrary to a traditional systematic review, a conceptual review has some "flexibility to address the complexity of the substantive issues we care about" (Kennedy, 2007, p. 146). It allows to refine and extend guiding questions in the course of the review process. Moreover, it also accommodates the application of rigor and transparency to the database search, the study selection and the analyses.

Kennedy's (2007) broad distinction between systematic and conceptual reviews was useful to articulate our epistemological outlook, but we also turned to other frameworks and typologies for further methodological guidance. In terms of approach we position our review as a scoping study. Like systematic reviews, scoping studies or scoping reviews follow "a structured process" but instead of exploring a delineated empirical research question they aim "to identify knowledge gaps, scope a body of literature, clarify concepts or to investigate research conduct" (Munn et al., 2018, p. 1). Scoping studies are descriptive and do not necessarily include a critical appraisal of the quality of the included literature (Arksey & O'Malley, 2005). Furthermore, our research aim – to map and possibly integrate current conceptualizations of sensemaking in a specific context – corresponds to that of a theoretical review, as we intended to "[draw] on existing conceptual and empirical studies to provide a context for identifying, describing, and transforming into a higher order of theoretical structure and various concepts, constructs or relationships" (Paré et al., 2015, p. 188).

3.2 Database query

The Web of Science (WoS) and ERIC databases were queried in July 2020. As listed in Table 2, we used Boolean operators to combine search terms referring to (1) sensemaking or sensegiving; (2) outcome measures; and (3) educational professionals. We allowed these terms to be present in all search fields. We also repeated the query with search terms referring to DBDM at the second position, but this only yielded one additional reference. The ERIC search query was filtered to include peer reviewed sources only, a prerequisite fulfilled by default in WoS. The systematic database query yielded 1426 unique records that were put into a spreadsheet.

Table 2. Database queries

Database query	Search term combination
1	*sensemak* OR sense-mak* OR sensegiv* OR sense-giv** *AND* *data* OR output OR outcome OR score OR feedback OR performance OR assess* OR evaluat** *AND* *edu* school* OR teacher OR principal*
2	*sensemak* OR sense-mak* OR sensegiv* OR sense-giv** *AND* *data use OR data-based OR data-driven OR data-informed* *AND* *edu* school* OR teacher OR principal*

3.3 Study selection

The selection process comprised four phases. The first one involved comprehensive coding of all 1426 titles and abstracts. During this first phase, a review diary was also kept in which themes for further exploration and striking quotes were noted. This phase allowed us to grasp the breadth of application of the sensemaking construct, yet served primarily to assess the records' eligibility for inclusion. We were specifically in pursuit of theoretical and empirical studies that discuss teachers' and school leaders' engagement with student achievement or school performance data, and use the term 'sensemaking' to refer to this process or phenomenon. Sensemaking did not need to be the (sole) focus of the paper in order for it to be included. Moreover, as our aim was a broad conceptual exploration, we did not a priori exclude papers based on whether they referred to a specific theoretical sensemaking paradigm or used 'sensemaking' simply as a descriptive label.

We assigned each record one or more open codes to describe its scope or main focus, and a code for the research methodology employed. Where possible and relevant, we also assigned in vivo thematic codes describing how sensemaking was defined in the study, or what aspects of or perspectives on sensemaking were particularly salient (examples of such codes are *intersubjectivity, Weick, sociocultural).* In line with our baseline inclusion criteria, we also coded the records for whether or not they pertained to research in an educational setting, and if so, which level. Furthermore, where possible, we coded for the sensemaking actor (examples include *health professionals, business executives,* and in educational studies, *students, district leaders, principals, pre-service teachers*). Based on these codes we assigned each record a preliminary mark for inclusion. In total, 114 records were marked for further review.

In a second phase, we reread all 114 abstracts that had received a preliminary mark. This second reading served for further refinement. Studies or papers in which (based upon more thorough review) the sensemaking actor was clearly not an educational professional in a K-12 school context were, for instance, excluded. The same holds for studies in which the object of sensemaking was policy implementation, rather than performance or achievement data. In a small number of cases we diverged from these criteria when we felt the paper in question provided relevant insights needed to enrich our understanding of sensemaking in educational DBDM. At the end of the second phase, our selection was narrowed down to 28 records.

In a third phase, the full text papers of the 28 selected records were read and reviewed. Again, we critically assessed the studies for sensemaking object and actor, as we further refined our selection. Papers with a policy orientation, papers situated more at the district levels, and papers that discussed accountability pressures in general, were only retained if deemed crucial to inform our understanding of the way sensemaking perspectives are employed in data use research in a K-12 context. In total, 20 papers were retained from the database query.

As a sensemaking perspective is on the rise in DBDM research, so are DBDM studies specifically discussing, or at least mentioning sensemaking. Therefore, in a fourth and final phase we included five recent papers that had not been returned by the database query, yet provided insights on sensemaking themes directly relevant to our research aim, and that fit in well with the selection we were making. These papers were available as advance online copies of articles to be published in 2021 in *School Effectiveness and School Improvement* and the 2021 special issue on DBDM of *Studies in Educational Evaluation*. These additions bring the total to 25 papers, as listed in order of publication date in Table 3.

The selected papers were ordered into a matrix that would allow for both horizontal and vertical analyses of recurring themes (Miles et al., 2014). As our aim was indeed to identify themes and patterns, and as we acknowledged our own active role in this process, we used a thematic analysis approach (Braun & Clarke, 2006). A combination of charting the data and subsequently performing thematic analysis is considered suitable for the purposes of scoping studies (Levac et al., 2010). The approach facilitates the production of a narrative account, and it allows or even forces the researcher "to prioritize certain aspects of the literature" throughout the analysis process (Arksey & O'Malley, 2005, p. 28). Charting and coding were primarily guided by the leitmotivs identified in the theoretical framework (cf. supra) and the thematic clusters that served to group the selected papers (cf. infra). These clusters were connected to (and based on) underlying dimensions of sensemaking (micro)processes, actors, locus, outcomes and triggers.

Table 3. Selected studies

Reference	Selection	Source	Type	Short description	Cluster
Even, 2005	DB Query	Mathematics Education Research Journal	theoretical	problem analysis of mathematics teachers' use of contemporary assessment approaches and techniques	A
Knight & Yorke, 2008	DB Query	International Journal of Educational Research	theoretical	(high-stakes) assessment and (public and formal) reporting practices about achievement are "contexted acts of sense-making about fluxional social practices"	C
Coburn et al., 2009	DB Query	Teachers College Record	qualitative	the unfolding of evidence use (decision trajectories) at the district level	B
Coburn & Turner, 2011	DB Query	Measurement: Interdisciplinary Research & Perspective	theoretical	data use involves interpretive processes, social and organizational conditions, and politics	B
Cosner, 2011	DB Query	Educational Management Administration and Leadership	qualitative	development of teacher knowledge and instructional considerations through data-based collaboration as part of a literacy instructional reform; the influence of principal communication (as a sensegiving mechanism) on teams' sensemaking of standardized student literacy assessment (and other) data	B
Spillane, 2012	DB Query	American Journal of Education	theoretical	conceptual and analytical tools for studying data in practice	B
Datnow et al., 2012	DB Query	Journal of Education for Students Placed at Risk	qualitative	high school teachers' understanding and implementation of data use to improve instruction	B
Jennings, 2012	DB Query	Teachers College Record	theoretical	features of accountability systems influence data use but this influence is mediated by individual and organizational characteristics	B
Park et al., 2013	DB Query	Educational Policy	qualitative	strategic framing by formal leaders (at the local level) mediates the implementation of data-driven decision making	B
Cho & Wayman, 2014	DB Query	Teachers College Record	qualitative	role of sensemaking in data system implementation	C
Bertrand & Marsh, 2015	DB Query	American Educational Research Journal	qualitative	middle school teachers' sensemaking of student outcome data (with a focus on English language learners and special education students)	A
Sellar, 2015	DB Query	Critical Studies in Education	theoretical	relationship between commensuration/datafication and affect	C

Table 3 (Continued)

Reference	Selection	Source	Type	Short description	Cluster
Christman et al., 2016	DB Query	Teachers College Record	qualitative	one primary school teacher's sensemaking of instructional and assessment data in the context of a PLC (professional learning community) intervention on mathematics, resulting instructional changes	A
Farrell & Marsh, 2016	DB Query	Educational Administration Quarterly	qualitative	properties and perceptions of data shape teacher sensemaking of data and instructional responses	C
Wardrip & Herman, 2018	DB Query	Teacher Development	qualitative	teachers' collaborative sensemaking of student data, drawing upon informal data	A
Schildkamp, 2019	DB Query	Educational Research	theoretical	iterative model of school improvement in which data use (with sensemaking as a central phase) plays an important role	A
Vanlommel & Schildkamp, 2019	DB Query	American Educational Research Journal	qualitative	primary school teachers' high-stakes decision making: intuition v rationality	A
Snodgrass Rangel et al., 2019	DB Query	Education and Urban Society	qualitative	science teachers decide how to use data (and decide which data to use) based on policies and expectations, which they balance with their own understanding of science education	B
Falabella, 2020	DB Query	Journal of Education Policy	qualitative	enactment (not mere implementation) of accountability policy including school members' sensemaking of outcome data	B
Sutherland, 2020	DB Query	Educational Policy	qualitative	local stakeholders' interpretation and implementation of accountability policies	B
Farley-Ripple et al., 2021	* Added	School Effectiveness and School Improvement	mixed methods	assessment systems (information systems) and their features mediate educators' sensemaking i.e. knowledge creation and deciding on actions	C
Mandinach & Schildkamp, 2021	* Added	Studies in Educational Evaluation	theoretical	commentary paper that addresses misconceptions regarding data-based decision making (research)	A
Vanlommel et al., 2021	* Added	Studies in Educational Evaluation	qualitative	primary school teachers' high-stakes decision making process from a dual process perspective	A
Fjørtoft & Lai, 2021	* Added	Studies in Educational Evaluation	theoretical	narrative and numerical data have different modal affordances (sensemaking resources)	C
Lasater et al., 2021	* Added	Studies in Educational Evaluation	qualitative	organizational aspects of data use influence sensemaking of data and may induce deficit thinking in educational professionals	B

3.5 Overview of the selected papers

Type and scope

As indicated in Table 3, 15 out of the 25 selected papers are qualitative studies. For the most part these are (comparative) case studies, sometimes longitudinal, that offer in-depth insight in sensemaking processes in schools, teams or individuals. For some studies, participants had been purposefully selected because their schools are known to face specific challenges, or, on the contrary, because they can be regarded as good-practice schools, or because they were already participating in an intervention. One other empirical paper discusses a mixed-methods study, in which log data from a computer data system were analyzed to capture data use patterns (Farley-Ripple et al., 2021). The remaining nine papers are theoretical in nature, some with illustrations based on qualitative data from prior research.

Table 3 includes a short description capturing the general scope of each paper. Achievement or performance data is an object of sensemaking in all the selected papers, but not always the sole object. Some of the papers for instance address DBDM in a broader sense, also including use of informal data such as classroom observations. The subject matter to which the data pertain (mathematics, science, language, arts) varies over those studies that explicitly mention it because it is relevant and where there was a specific focus in the first place. In studies about grade level sensemaking in primary schools, for instance, a distinction is not always made. The sensemaking actors involved or discussed in the research are mainly teachers, and school leaders and (other) administrators (e.g., district leaders) to a smaller extent. The theoretical papers by Knight and Yorke (2008) and Sellar (2015) somewhat stand out from the selection in the sense that they are not typical DBDM-studies and do not (specifically) address achievement data in K-12 contexts. They were included because they discuss datafication in education.

Thematic clusters

Based on patterns observed during analysis, we grouped the 25 selected papers into three thematic clusters. A first group of papers, marked as cluster A in Table 3, discuss the (micro)process of sensemaking, i.e., sensemaking as a phase in the DBDM cycle of turning raw data into actionable knowledge. Overall, these papers tend to zoom in on the interpretive nature of sensemaking, attending to processes such as attribution and the role of intuition and pre-existing mental models. Some of these papers address how collaborative data use contributes to individual meaning making by making cognitive processes explicit.

A second group of papers (cluster B) apply a sensemaking perspective to data use in schools. They generally attend to both the cognitive/interpretive and social/situated dimensions of sensemaking. They do so by describing how interpretive processes are shaped by individual teachers’ and school leaders’ knowledge and beliefs, but also by social interactions and by contextual factors, and how sensemaking can be a question of power and politics. Overall, use of achievement data tends to be approached as an instance of policy enactment here. Consequently, a number of papers in this cluster take a closer look at local reception and interpretation of policy messages. Together, the papers in this cluster provide insight into the interplay of individual and collective sensemaking (sensemaking actors), the interplay of cognition and language (as the locus of sensemaking), and factors that potentially contribute to (un)desirable outcomes of data use.

A third and final group of papers (cluster C) looks at data (systems) as sensemaking resources, i.e., as triggers and tools for sensemaking. These papers discuss (interpretive) processes and responses associated with different types of data and representational qualities of data, and the “interpretive flexibility” of data use technology.

Theoretical underpinnings

When referring to “sensemaking theory”, a number of authors refer at least cursorily to the seminal Weickian perspective on sensemaking. Furthermore, applied and conceptual work by Coburn and Spillane on policy enactment features as a source in a majority of the selected papers. In this particular selection, Vanlommel and colleagues (2021; 2019) are chronologically the first to explicitly link to Klein’s Data-Frame theory. A number of papers also build on other sensemaking interpretations, for instance from the information systems literature (e.g., Cho & Wayman, 2014; Sellar, 2015), or do not explicitly refer to a sensemaking paradigm at all but use the term descriptively (Even, 2005; Falabella, 2020; Knight & Yorke, 2008).

Sensemaking perspectives are combined with insights from other psychological scholarship such as attribution theory, self-affirmation theory, heuristics and affect theory, and from organizational studies, e.g., organizational learning, organizational decision making, naturalistic decision making. Furthermore, co-construction paradigms, situative theory and political theory serve to study sensemaking as it comes about in interaction and in daily practice, while social semiotics serves to shed light on modal affordances of data.

4 Teachers' and school leaders' sensemaking of formal achievement data: towards an integrated conceptual framework

In this section we bring together conceptualizations and insights drawn from the review, with the aim to get a better and more integrated grasp of what educational professionals' sensemaking of formal achievement data entails. The themes we identified during analysis serve to structure our discussion of our findings. First, we reflect on the place of sensemaking in data use models, and next we zoom in on the interpretive nature of sensemaking (micro)processes. Subsequently, we consider individual and collective aspects of educational professionals' sensemaking of formal achievement data, and the way sensemaking processes interplay with sensemakers' contexts. Finally, we shift our attention to the data and data systems themselves. The 'vocabulary' drawn from our general theoretical framework serves to discuss these various themes and components.

4.1 Sensemaking is regarded as a core aspect of DBDM

Sensemaking is a phase in the data use cycle

Sensemaking is a prominent phase in contemporary theories of action on data use, such as Schildkamp's (2019) iterative model of DBDM for school improvement. This model is based on the premise that school improvement endeavors start with educational professionals formulating improvement goals. Subsequently they collect or access different types of data, and *make sense* of those data in order to gauge or monitor whether they are achieving the goals. The aim is to be able to formulate and follow up on decisions and actions that will help them (further) realize the goals (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). The essence of sensemaking in this cyclical and iterative process is figuring out what data mean in relation to the goals (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). Which problems do the data bring to the surface (why are certain goals not met)? How can those problems be explained (where are the gaps)? And, how should one proceed from there in order to realize the goals (how can gaps be closed)? Sensemaking is characterized in this view as a complex problem-solving process (Wardrip & Herman, 2018) involving problem definition, diagnosis and judgement (Coburn et al., 2009; Coburn & Turner, 2011; Datnow et al., 2012; Vanlommel et al., 2021; Vanlommel & Schildkamp, 2019).

What raw data mean is not given but constructed

Sensemaking is seen as a crucial phase in the data use cycle, because data simply do not tell a story by themselves. What raw data mean, and specifically what they mean in relation to goals, is not given: answering the aforementioned questions is rarely self-evident (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). Pinpointing problems, hypothesizing about causes, and designing solutions, requires educational professionals to actively 'check' a number of things. What do they infer from the data? How does this information fit in with what they already know, understand and assume about their pupils and their organization? How does it fit in what they have learnt throughout time about learning and instruction, about educational practice and policy? In other words, making sense of data means looking at those data through the lens of individual and local knowledge, prior experience, and professional expertise (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). The outcomes of ongoing sensemaking in daily practice, but also the way consecutive and concurrent sensemaking processes unfold, shape individual and organizational thinking. Consequently, it shapes the implementation of change (or preservation) in teacher practice, school policy, and ultimately student learning (Coburn & Turner, 2011; Datnow et al., 2012; Spillane, 2012).

4.2 Sensemaking involves interpretive processes

Sensemaking is fundamentally interpretive

Educational professionals' sensemaking of (achievement) data is fundamentally interpretive (Bertrand & Marsh, 2015; Coburn et al., 2009; Coburn & Turner, 2011; Farrell & Marsh, 2016; Vanlommel & Schildkamp, 2019; Wardrip & Herman, 2018). It comprises different steps and interrelated micro-operations, of which we find different characterizations in the reviewed papers. Some authors (implicitly or explicitly) focus on the deliberate nature of sensemaking and the skills needed to "effectively" make sense of data – effective in the sense that sensemaking and decision-making will result in school improvement. They do so by making a broad distinction between data analysis on the one hand and interpretation on the other hand, i.e., being able to read and understand the data, and being able to make valid inferences based on the data (Mandinach & Schildkamp, 2021a; Schildkamp, 2019; Vanlommel & Schildkamp, 2019). Authors following this line of thought (e.g., Schildkamp, 2019) explicitly apply the global label of sensemaking to operations unraveled in more detail in data use frameworks put forward by authors such as Marsh et al. (2006), Mandinach et al. (2008), Marsh (2012) and Schildkamp and Poortman (2015). Those models tend to be rooted in a waterfall logic of turning data into information into actionable knowledge, through micro-operations, such as data organization, analysis, and synthesis. However, we also find characterizations of the interpretive sensemaking steps that lean more towards what occurs naturally and automatically (at least to some extent) when people process what they encounter in their environment. Coburn and Turner (2011), for instance, distinguish between phases of "noticing", "making meaning", and "constructing implications". This classification is in concordance with "sensemaking moves" as described in organizational sensemaking literature: sensemaking being triggered by an ambiguous issue or event, people constructing an understanding of this issue or event, and them taking action (Maitlis & Christianson, 2014).

Together, in terms of steps and micro-operations, the studies address a number of questions that educational professionals might ask themselves when making sense of test scores and other types of (achievement) data. Examples of questions include: Which aspects of performance have been captured in the presented measure and what is an appropriate way to analyze the symbolic representations of those measures (Fjørtoft & Lai, 2021; Mandinach & Schildkamp, 2021a; Spillane, 2012)? Which valence do I attach to this result, i.e., am I satisfied or not (Coburn & Turner, 2011)? What could have contributed to this result (Bertrand & Marsh, 2015; Lasater et al., 2021)? Should I (should we) respond and if so, in what manner (Coburn & Turner, 2011)? While it is possible and necessary to distinguish and study different sensemaking steps and micro-operations, sensemaking is not a demarcated process in reality. It is fluid and complex and therefore difficult to fit into a prescription.

4.3 Sensemaking is a product of individual lenses

As established, sensemaking does not start with data (external cues or stimuli), but with a sensemaker. Since data need to be processed by human beings in order to convey meaning, and meaning is a subjective construction instead of an objective truth, personal lenses inevitably act as filters. As Even (2005) puts it, instructional decision making involves “hearing through”: teachers will use their own knowledge, beliefs and dispositions – their “personal and social resources” – to interpret student outcomes. This entails that data can come to mean different things to different people (Schildkamp, 2019). Also, when personal and social resources are limited or overemphasized, this can problematize sensemaking (Even, 2005).

Sensemaking requires human capacity

Sensemaking involves competence and capacity (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). In order to go through the sensemaking motions in a meaningful way (e.g., being able to analyze data appropriately, sift through and prioritize information, recognize and articulate problems, formulate workable improvement actions), educational professionals need competence and a certain degree of expertise (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). This idea is closely related to contemporary and wide-spanning conceptualizations of data literacy (see for instance Beck & Nunnaley, 2021; Mandinach & Gummer, 2016) which acknowledge that data literacy is multilayered. Data literacy comprises knowledge and skills pertaining to appropriate data analysis, but also for instance to learning and instruction and subject matter. Besides knowledge and skills, human capacity for sensemaking also has an attitudinal dimension in terms of confidence, safety and motivation (Coburn & Turner, 2011; Even, 2005; Mandinach & Schildkamp, 2021a; Schildkamp, 2019). Educational professionals need to be able to feel they can overcome potential struggles in interpreting data (Mandinach & Schildkamp, 2021a; Schildkamp, 2019), that they can use data in a healthy and safe professional environment (Falabella, 2020; Lasater et al., 2021) and that data use is geared towards their own values and those of their organization (Schildkamp, 2019). Sensemaking can also trigger affective, emotional responses that serve as either an impediment or as a springboard into action (Falabella, 2020; Sellar, 2015).

Personal beliefs and assumptions shape sensemaking

The interpretive nature of sensemaking and the fact that sensemakers' feelings, attitudes and motivation guide how their sensemaking unfolds, point to the impact of personal beliefs and assumptions (Coburn & Turner, 2011; Schildkamp, 2019). Several studies describe how new information is assessed, coded and used by fitting it in with their own cognitive frameworks or mental models (e.g., Bertrand & Marsh, 2015; Spillane, 2012; Vanlommel et al., 2021; also see earlier work by Spillane & Miele, 2007). A wide range of beliefs and epistemological stances is addressed in the literature. These include beliefs about students, learning and instruction, and assessment (Bertrand & Marsh, 2015; Lasater et al., 2021), about the nature, utility, relevance and validity of (certain types of) data (Bertrand & Marsh, 2015; Coburn & Turner, 2011; Farrell & Marsh, 2016; Jennings, 2012; Wardrip & Herman, 2018), but also about data use and data-based inquiry in general (Cho & Wayman, 2014; Datnow et al., 2012; also see Jimerson, 2014).

Educational professionals' beliefs are intertwined with the end to which achievement data are ultimately used. Data can be used in order to determine "how teachers view their schools, students, and themselves ([when test data used as a] lens); how they determine what's working, what's going wrong, and why ([tool for]diagnosis); what they should do in response (compass); how they establish whether it worked (monitoring); and how they justify decisions to themselves or to others (legitimizer)" (Jennings, 2012, p. 4). On a micro-level, filtering through personal beliefs and assumptions occurs even in the nuclear stages of sensemaking. People tend to notice and select those cues that accord with their prior experiences and assumptions (Coburn & Turner, 2011; Spillane, 2012). A recognition-primed paradigm describes how people take those familiar elements in order to form quick explanations and conclusions, disregarding potential ambiguity, without thoroughly and truly making sense of data (Klein et al., 2007; Vanlommel et al., 2021). Building on prior sensemaking research (e.g., Coburn & Turner, 2011; Datnow et al., 2012; Spillane & Miele, 2007), Bertrand and Marsh (2015) empirically illustrate how teachers' mental models, i.e., their implicit or explicit beliefs about causality, guide attributions of student achievement data and subsequent decision-making. In line with attribution theory, the authors find that the nature of causal inferences influences teachers' motivation to make subsequent changes or improvements (Bertrand & Marsh, 2015). In this respect, they particularly zoom in on the levels of control and malleability associated with perceived causes of outcomes: are scores perceived as the product of instruction, of student understanding, of the nature of the test, or of student background characteristics?

Sensemaking is rooted in identity-construction

The essence of achievement data is to get a picture of how individuals or teams do compared to standards, to others, or to prior performance (the classic typology of criterion-, norm- and self-referenced feedback). Interpreting such data, inherently involves evaluation and judgement. Did we (did you, did they) do well or not, and who is responsible? This can be uncomfortable when the proposed answers to these questions challenge one's self-image and/or self-efficacy, or when the conclusion is undesirable. Sometimes, such friction will increase people's motivation to thoroughly process the information (Coburn & Turner, 2011, referring to Spillane, Reiser, et al., 2002) and get to work on achieving their goals. Typically, however, people have a tendency to dismiss or downplay undesirable information out of self-preservation. This is when mechanisms of confirmation bias and self-affirmation bias come into play (Lasater et al., 2021; Vanlommel & Schildkamp, 2019; Coburn & Turner, 2011, referring to Spillane, Reiser, et al., 2002). Sensemaking of data "prompts a constant reexamination of identity" and people are naturally inclined to try and validate their pre-existing beliefs and preserve their identity (Lasater et al., 2021).

Sensemaking is not necessarily a rational affair

(Over)reliance on pre-existing assumptions thus explains to a certain extent why sensemaking of data is not necessarily a rational affair (Mandinach & Schildkamp, 2021a; Schildkamp, 2019), even when those data have been collected deliberately and systematically. When making inferences and judgements, educational professionals – like other decision makers – are inclined to use mental shortcuts and rely on their intuition (Mandinach & Schildkamp, 2021 and Schildkamp, 2019, referring to Bertrand & Marsh, 2015; Kahneman & Frederick, 2005; Vanlommel et al., 2017). Vanlommel and colleagues enlighten how data use, and particularly teachers' construction of "interpretive arguments" in high-stakes DBDM, involves dual processing (Vanlommel et al., 2021). They find that teachers sometimes base conclusions on personal criteria and quick-fire "judgmental heuristics", instead of considering multiple data sources and evaluating competing explanations (Vanlommel & Schildkamp, 2019).

Intuitive expertise (as described in DBDM research but also by authors on naturalistic decision making such as Kahneman and Klein) has its merits and has even been put on a pedestal in educational decision making for years (Vanlommel et al., 2021). Nevertheless, a critical stance is required. After all, sensemaking can lead to incorrect readings, invalid inferences or biased decisions when personal lenses become blinders. Consciously or unconsciously favoring certain types of data or conjectures that validate one's prior views and assumptions, blocks alternative explanations and incongruent information from vision (Mandinach & Schildkamp, 2021a; Schildkamp, 2019; Vanlommel et al., 2021; Vanlommel & Schildkamp, 2019).

4.4 Sensemaking is a collective endeavor

Collective sensemaking entails meaning negotiation and co-construction

In a school improvement logic, it is generally considered paramount for educational professionals (in different roles) to collectively make sense of data (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). Collective sensemaking, for instance in data discussions, broadens the lens through which data are interpreted and problems are framed, provokes debate, and yields new insights, both on an individual and on a shared level. It also entails meaning negotiation and the co-construction of frames and narratives (Coburn & Turner, 2011; Datnow et al., 2012; Park et al., 2013; Spillane, 2012). Interacting with colleagues and coaches, and participating in professional learning communities with peers and/or facilitators can enhance sensemaking because voicing inferences helps to surface and highlight beliefs and affectivities, as well as ambiguities and "intuitive pitfalls" (Bertrand & Marsh, 2015; Christman et al., 2016; Even, 2005; Vanlommel & Schildkamp, 2019). Educational professionals' individual sensemaking may benefit from training and coaching concerning both the "mechanics of data use" and ways of translating insights into daily practice (Coburn & Turner, 2011).

Collective sensemaking takes shape in routines

Coburn and Turner (2011) and Spillane (2012) discuss the dynamics and affordances of (organizational) data use routines in practice: educational professionals' day-to-day interactions with data and with each other. Data use routines, both formalized routines and more informal interactions, give direction to sensemaking as they bring "a particular configuration of people together around a particular set of data and structure their interactions in specific ways" (Coburn & Turner, 2011, p. 181). Consequently, the interpretations made in data interactions are greatly influenced (a) by the participants involved (which prior experiences, views, interests do they bring to the table?); (b) by the data they use as a starting point (what data are seen as valuable and informative by participants, which data can participants bring in to help with contextualization?); and (c) by the dynamics of these interactions (whose voice is heard, which observations do they make, and who wields the proverbial gavel?) (Coburn & Turner, 2011).

In terms of participants, individual sensemakers have different positions within a school or system, paired with different vantage points and interests. This enriches the dialogue but can also provoke debate. Moreover, relationships of power and authority, which are often derived from formal roles and structures within an organization or a system, impact the influence specific individual actors can exert in meaning negotiation (Coburn & Turner, 2011; Spillane, 2012). Furthermore, data interactions, conversations and routines often focus on specific types of data (Coburn & Turner, 2011). For instance, in a study of teacher discussions in grade-level teams, Wardrip and Herman (2018) find that standardized test scores tend to initiate such conversations, but that educational professionals also use a host of other sources, including informal ones such as observations from daily practice, to explain and contextualize those scores. This points to the importance of knowledge management in schools. The fact that teachers interpret achievement data using a host of other data, as well as their own intuition, is expanded upon as well by authors such as Datnow et al. (2012).

Finally, the mechanisms of data interactions pertain to reasoning and negotiating (Wardrip & Herman, 2018) and can introduce new levels of ambiguity and friction. Christman et al. (2016) use a situative theory perspective, which characterizes instructional growth as a form of learning-by-doing and focusses on the interplay of individual cognitive processes and dynamics of co-construction. They find that collective sensemaking of teaching practice and student outcomes in professional learning communities can stimulate productive dissonance. Articulating individual sensemaking and personal views in collegial discussions, deliberations and recurring feedback cycles provokes cognitive conflict (Cobb et al., 1990, as cited by Christman et al., 2016) as new information sometimes challenges prevailing assumptions and practices. When participants commit to taking up this new information to weigh up and potentially revise their held beliefs, instead of simply dismissing it, the experienced dissonance becomes productive. This allows teachers to grow in their reasoning and in their pedagogical expertise and produces instructional change and improvement (Christman et al., 2016).

4.5 Sensemaking is embedded in a social and organizational context

Interpretive sensemaking processes are shaped by the social environment of the sensemaker

Making sense of data does not happen in isolation (Mandinach & Schildkamp, 2021a). Interpretive sensemaking processes shape and are shaped by the social and contextual surroundings of the sensemaker(s) (Coburn et al., 2009; Coburn & Turner, 2011; Lasater et al., 2021; Spillane, 2012). In order to get a sense of how sensemaking unfolds, both on an individual and on a collective level, and how it contributes to school improvement, it needs to be studied as it takes place in day-to-day practice (Datnow et al., 2012; Spillane, 2012), embedded in a specific organizational and political context (Coburn & Turner, 2011).

Factors that influence individual sensemaking, also play a role in group-level sensemaking

Many of the same factors that influence individual sensemaking, are also at play in group-level sensemaking that takes place at the local level (in schools and for instance also in districts). Data interactions involve cognition, as they guide which elements participants notice and focus on, and how they frame information (Spillane, 2012). Drawing on insights pertaining to the relationship between individual cognition and situated and distributed cognition (cf. the cognitive framework developed by Spillane, Reiser, et al., 2002), Spillane (2012) explains how day-to-day educational practice within a community of practice, such as a school (or a system in the wider sense), is shaped by individual mental models, but also by shared, intermental models. In the case of evaluation and assessment, for instance, these can be intermental models about what constitutes "successful performance". Thus, organizational sensemaking is belief- and value-driven as well, for instance because it needs to be geared towards (improvement) goals that have been agreed upon as being important (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). Culture, norms and values influence sensemaking at the level of the school, but also within subgroups, such as departments (Coburn & Turner, 2011; Datnow et al., 2012). Furthermore, issues of professional safety and responsibility, but also of collective identity are born and embedded in the (organizational) narratives of groups, schools and systems (Falabella, 2020; Lasater et al., 2021). Christman et al. (2016) argue that sufficient human capital (knowledge and expertise among individual participants) but also social capital (trust and willingness to engage in and commit to interpersonal exchanges) are important preconditions for productive dissonance to occur in data conversations. In describing how culture and interaction shape the mechanics of sensemaking in schools, a number of authors also refer to the work of Supovitz and of Horn and colleagues on educational professionals' organizational learning (e.g., Horn et al., 2015; Horn & Little, 2010; Supovitz, 2010).

Organizational conditions impact individual and collective sensemaking at the local level

A number of more tangible organizational conditions also impact individual and collective sensemaking at the local level. Educational professionals need time and resources for sensemaking (Coburn et al., 2009; Coburn & Turner, 2011; Datnow et al., 2012) and access to data and technology (Coburn & Turner, 2011). Ideally, there is also a system of knowledge management in place (Wardrip & Herman, 2018), productive data use routines (Coburn & Turner, 2011; Spillane, 2012) and of course "human infrastructure" (Coburn & Turner, 2011).

Key actors in shaping both tangible and intangible conditions for sensemaking are formal and informal leaders, such as school leaders, but also district leaders (Coburn et al., 2009), or school board members (D. H. Sutherland, 2020). They not only participate in collective sensemaking and data use routines, but also shape how these processes unfold within their organization (Coburn & Turner, 2011; Cosner, 2011; Schildkamp, 2019). Firstly, in terms of management and coordination, leaders facilitate sensemaking processes for their team (Schildkamp, 2019). They do so by making sure there are structures, resources and supports (Coburn & Turner, 2011; Cosner, 2011; Datnow et al., 2012; Mandinach & Schildkamp, 2021a), by designing and facilitating data use routines (Coburn & Turner, 2011; Cosner, 2011), and by deciding whether and how improvement actions are implemented (Schildkamp, 2019). Secondly, in terms of culture building and leadership, the way leaders establish norms and values, implement data use policies, and set priorities, impacts sensemaking and decision making to a great extent (Coburn et al., 2009; Datnow et al., 2012; Mandinach & Schildkamp, 2021a). Finally, in terms of negotiation and sensegiving, local leaders also mediate policy messages and pressures from other levels in the school system and educational system (Coburn & Turner, 2011).

Leaders act as sensegivers

Coburn et al. (2009), Cosner (2011), and Park et al. (2013) combine sensemaking insights with frame analysis to take a detailed look at how local leaders act as sensegivers by framing data and data use within their organization, how this framing affects the actual implementation of data policies in schools, and how the meaning of data such as test scores is subsequently constructed and negotiated among educational professionals. Leaders construct narratives to frame problems (diagnostic framing) and potential solutions (prognostic framing), but also to create resonance and buy-in (motivating framing) (Coburn et al., 2009; Park et al., 2013).

This framing is both interpretive and strategic. Firstly, in terms of interpretation, the (content) knowledge of sensegivers determines how frames for articulating problems and designing solutions come into existence and are substantiated (Coburn et al., 2009). Additionally, leaders' own, evolving conceptions of data analysis and collective sensemaking shape the expectations they formulate towards their team and the conditions they create with regard to collaborative data use (Cosner, 2011). Furthermore, in order for framing to be sufficiently credible, sensegivers need to be aware of and "slightly stretch" beliefs and practices prevalent in their organizations (Park et al., 2013). Secondly, in terms of strategy, authority and politics also play crucial roles (Coburn et al., 2009). Attribution and appraisal can be a delicate story of articulating responsibility for certain result, for instance. Moreover, framing not only creates conditions for sensemaking but it is also a means of persuasion – instilling the idea in a local team that data use can be a meaningful practice (Park et al., 2013). Overall, leaders are “key communicators” in data-based reform by articulating goals and expectations (Cosner, 2011; Park et al., 2013).

4.6 Sensemaking interplays with the broader (policy) context

School-external protocols, interventions and policies shape sensemaking

Much like data use routines that burgeon school-internally, *school-externally* devised protocols and interventions harbor the potential to focus attention on specific issues and give direction to data use conversations. However, they are mediated by sensemakers at the local level (Coburn & Turner, 2011). The same holds for educational policy initiatives. Their enactment or implementation is shaped by the individual and collective interpretations of local educational professionals, as well as sensegiving efforts by local leaders. Contextual factors impact teachers' data use in practice by interplaying with local beliefs and practices (Coburn & Turner, 2011; Datnow et al., 2012; Falabella, 2020; Jennings, 2012; Snodgrass Rangel et al., 2019; D. H. Sutherland, 2020). The implementation of DBDM and data use policies is a (micro)political act to a certain extent (Coburn et al., 2009; Park et al., 2013).

Sensemaking is a key to understand why data produce unexpected and non-normative outcomes

Multiple authors use a sensemaking outlook and contrast it with techno-rational perspectives on data use in order to explain why (achievement) data are sometimes used in non-normative ways or at least in ways unexpected or unintended by policymakers (e.g., Datnow et al., 2012; Jennings, 2012). Data use policies – and the papers reviewed here tend to specifically discuss accountability-based policies – are often based on the assumption that the availability of data will enforce change and improvement. However, in practice data "do not objectively guide decisions on their own—people do" (Spillane, 2012, p. 114). Sutherland's (2020) study of school board members' enactment of mandated assessment policies illustrates well how sensemaking, sensegiving and the construction of narratives are scaffolded within systems. She finds that individual board members might take on different positions towards what can and what cannot be achieved with standardized assessments, yet collectively decide to use such instruments in a way that fits into the local narrative of their organizations. In turn, those local interpretations do not necessarily align with system-level messaging regarding the evaluative purposes and value of standardized assessments (D. H. Sutherland, 2020). Similarly, Snodgrass Rangel et al. (2019) find that teachers balance messages about policy requirements and expectations with their own understandings of education and their personal beliefs about assessment and data use, and as a result, favor some types of data over others.

Jennings (2012) discusses how features of accountability systems may induce productive or distortive use of test score data, depending on how these features are received, understood and interpreted by educational professionals. These features pertain, for instance, to the perceived amount and locus of pressure (who is perceived as being held accountable for the results, and what are the consequences associated with accountability) and to the goals and benchmarks that are perceived as salient (long term versus short term gains, growth versus proficiency, process versus outcome etc.). Highlighting how sensemaking plays a part in the alignment between perceptions and assumptions of data users on the one hand, and of data providers on the other hand, productive use of test scores is defined as encompassing "practices that improve student learning and *do not invalidate the inferences about student and school-level performance that policy makers, educators, and parents hope to make* [emphasis added]" (Jennings, 2012, p. 4).

System-level policies are mediated by local sensemakers and sensegivers

Accountability policies and pressures, and the way they are mediated at the local level, also impact individual and collective beliefs and mindsets in schools (Mandinach & Schildkamp, 2021a). Lasater et al. (2021) demonstrate how enactment of high-stakes accountability policies and framing by local leaders can cause educational professionals to feel threatened in their professional self-integrity and to push them into deficit thinking. Other authors also address how high-stakes standardized testing can raise questions of autonomy, (institutional) identity and (individual) responsibility in schools (Datnow et al., 2012; Falabella, 2020; Spillane, 2012) and impacts individuals' "subjectivities and affectivities". Processing test data can be a struggle for educational professionals, and information that challenges one's self-image can trigger emotional responses and strategic framing (Falabella, 2020).

System-level policies shape local discourse and day-to-day practice

In any case, it is clear that standardized test scores play a role in defining day-to-day instructional practice in schools (Spillane, 2012). Standardized test scores contribute to institutional storytelling, since interpreting such scores triggers questions about the meaning of schooling, about a school's identity and comparative positions to other schools. As such they provide productive discourse: numbers have symbolic power and "not only describe reality, they also produce it" (Falabella, 2020, p. 30). Falabella (2020) discusses datafication trends and what she calls the "accountability trap": a growing emphasis on instrumental logic risks blurring schools' goals and making learning subordinate to measurable outcomes.

4.7 Data and data systems are sensemaking resources

Different types of data require and activate different sensemaking processes

Different types of data require and will activate different sensemaking processes, and consequently, contribute to different (instructional) responses and other forms of educational decision-making (Farrell & Marsh, 2016; Schildkamp, 2019). While raw data may not carry meaning in the absence of a sensemaker, they can trigger connotations and values in the eye of a beholder before actual interpretation is yet to occur. Sensemaking perspectives uncover the mechanisms through which (inter)subjective beliefs shape such perceptions (Farrell & Marsh, 2016).

Data such as state assessment data can be regarded as a manifestation of accountability policy (Jennings, 2012). From an institutional perspective, data such as standardized test scores “embody particular representations of what it means to learn and teach” (Spillane, 2012, p. 131). They are instances of commensuration (fitting attributes into one common metric). As such, they simplify performance into something that can be measured and thereby draw attention to specific aspects of learning and instruction (Sellar, 2015; Spillane, 2012). As artifacts, standardized test scores are symbolic representations of achievement, but it is important to note they are also the result of a conscious transformation (Knight & Yorke, 2008; Sellar, 2015). And, like sensemaking in itself, any form of commensuration or ‘datafication' is creative and adds something to the world (Sellar, 2015). Still, achievement data do not carry stable and general meaning that is received at face value: data are given meaning by local sensemakers, and that meaning can diverge from the meaning intended by those who mandate testing and supply the data (Knight & Yorke, 2008). In this respect, Sellar (2015) also points to the affective nature of both commensuration and ensuing sensemaking. Meaning attributed to data can trigger emotional responses in recipients. Consequently, data can also be used to fuel “perceptual shifts” (with the PISA-shock as a system-level example).

A sensemaking perspective sheds light on why not all types of data are equal, and not even all (numeric) assessment data trigger the same responses (Farrell & Marsh, 2016). Farrell and Marsh (2016) find that educational professionals’ perceptions of data and assessment properties, such as the format in which data are presented, the scope and format of the assessment itself, or their own active involvement in the assessment (design), determine how they will ultimately use the data in their daily practice. Self-designed and self-administered classroom assessments, for instance, are perceived as most closely aligned to daily instruction, and as offering more opportunities for improving rather than proving. State assessments, on the other hand, are presented in such a way that it offers guidance for, for instance, grouping students (Farrell & Marsh, 2016). In that respect, Farrell and Marsh (2016) also find that the logic of classifying students according to proficiency, as is done in state assessments, has found its way to the school and classroom. This points to the fact that data use is performative: policy initiatives and data systems can introduce paradigms that become canon over time, and consequently also start to pervade and shape school-internal discourse. Referring to prior scholarship, Coburn and Turner (2011) give a similar example of how the No Child Left Behind Act in the USA introduced proficiency categories, which became a “system of meaning” in its own right and also entered school- and district-internal narratives.

Different types of data have different modal affordances

Taking on a social-semiotic perspective, Fjørtoft and Lai (2021) explain how different types of data have different “modal affordances” according to the conventions, beliefs and strategies that interpretive communities establish around them. Data are material-semiotic artifacts: their concrete representational properties, e.g., whether they are narrative or numeric, carry meaning and value because people have grown to interpret them and act upon them in specific ways (Fjørtoft & Lai, 2021). Narrative data tend to be associated with evolving storylines and informal, micro-level decision-making, for instance. Numeric data such as test scores and other statistical, psychometric data, on the other hand, have an aura of certainty and objectivity even though recipients sometimes struggle with interpreting them appropriately (Fjørtoft & Lai, 2021). Awareness of these modal affordances – the way specific types of data become associated with specific beliefs and practices – offers a way of looking at why certain data are overemphasized or accepted as valuable and valid (Mandinach & Schildkamp, 2021a). It also enlightens sensemaking challenges and opportunities, for instance in terms of data triangulation.

Data systems have interpretive flexibility

Finally, computer data systems can help educational professionals turn data into usable information. As such, they have been found to be an important mediator for knowledge development and design of improvement actions (Farley-Ripple et al., 2021). Technology can indeed support the human endeavor that is data use (Mandinach & Schildkamp, 2021a). However, much like the availability of data does not automatically lead to school improvement, providing access to systems does not guarantee that those systems will be used, let alone used in unequivocal and productive ways as intended by those who devise them (Cho & Wayman, 2014; Farley-Ripple et al., 2021). Technological determinism can be countered by looking at the "interpretive flexibility" of data systems (Cho & Wayman, 2014). Individual educational professionals differ in their usage of data systems according to the way they make sense of the data systems themselves, according to their personal notions of "data" and "data use", and according to assumptions about the potential affordances of available features and functions (Cho & Wayman, 2014; Farley-Ripple et al., 2021). Opportunities to promote productive use of data technology, for instance through support structures, leadership communication, and professional development or coaching, can only truly succeed when they take into account interpretive sensemaking processes (Cho & Wayman, 2014; Farley-Ripple et al., 2021; Coburn & Turner, 2011, referring to Means et al., 2009).

5 Framework and applicability

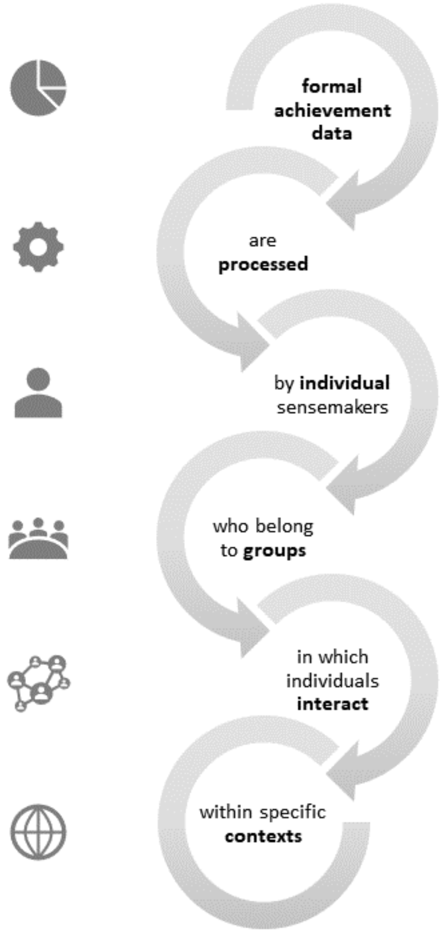
A prevalent assumption in educational policy and test development, is that providing achievement data, such as student test scores and school performance feedback, will successfully inform and drive school improvement endeavors. However, data "are only as good as how they are used" (Coburn & Turner, 2011, p. 173) by individuals and teams at the local level. We reviewed a selection of studies that use or at least echo this perspective when discussing teachers' and school leaders' use of formal achievement data. While the studies all have a sensemaking lens in common, authors vary in their use of the metaphorical function wheels. By focusing, zooming in or out, and applying filters, they capture specific aspects of the sensemaking phenomenon in DBDM.

Figure 3 brings together the crucial insights that we identified in the selected studies and discussed in detail in the previous subsections. The framework outlines a number of considerations that need to be taken into account when seeking to understand what 'happens' when teachers and school leaders make sense of formal achievement data. Because what happens, in its most basic sense, is that formal achievement *data* are *processed* by *individual* sensemakers. Those individual sensemakers in turn belong to *groups* in which individuals *interact*. And sensemaking always occurs within sensemakers' *contexts*.

The leveled structure of this framework does not imply a hierarchy or even a strong sense of linearity. In its entirety, the presented framework is precisely an appeal to keep sight of the bigger picture when investigating how formal achievement data are actually processed and how educational professionals' engagement with these data might produce normative or non-normative outcomes. Nevertheless, the holistic nature of this framework does not preclude its utility to serve as a 'pantry' of leads for educational researchers, policymakers, and test developers. (Niche) research may want to select some of the presented insights in order to zoom in or out on individual sensemakers in schools within educational contexts, for instance in comparative research on the effectiveness of assessment interventions. They may want to look at tangible aspects such as structures and conditions that are in place for sensemaking. Equally, they may choose to shed light on less tangible aspects that permeate and fuel the entire sensemaking process, such as individual and collective beliefs or interpersonal relationships.

We contend that this framework also has the potential to inspire practitioners, provided it is appropriately translated. The framework substantiates, for instance, *why* it is helpful for school leaders and teachers to collaboratively work with (formal achievement) data. It also elucidates why articulating your own assumptions and convictions (to yourself or towards others) can place interpretations in a new light while giving meaning to data. Furthermore, local leaders may benefit from a more thorough and conscious understanding of their role as sensegivers. Thus, a sensemaking perspective may also help to inform data users themselves, as well as those who support and train them.

Figure 3. Framework for teachers' and school leaders' sensemaking of formal achievement data



formal achievement data

- What raw data mean is not given but constructed.
- Different types of data have different modal affordances.
- Different types of data require and activate different sensemaking processes.
- Data systems have interpretive flexibility.

are **processed**

- Sensemaking is a phase in the data use cycle.
- Sensemaking is fundamentally interpretive.
- Sensemaking is not necessarily a rational affair.

by **individual** sensemakers

- Sensemaking is rooted in identity-construction.
- Personal beliefs and assumptions shape sensemaking.
- Sensemaking requires human capacity.

who belong to **groups**

- Interpretive sensemaking processes are shaped by the social environment of the sensemaker.
- Factors that influence individual sensemaking, also play a role in group-level sensemaking.
- Organizational conditions impact individual and collective sensemaking at the local level.

in which individuals **interact**

- Collective sensemaking takes shape in routines.
- Collective sensemaking entails meaning negotiation and co-construction.
- Leaders act as sensegivers.

within specific **contexts**

- School-external protocols, interventions and policies shape sensemaking.
- System-level policies are mediated by local sensemakers and sensegivers.
- System-level policies shape local discourse and day-to-day practice.
- Sensemaking is a key to understand why data produce (un)expected and (non)normative outcomes.

6 Conclusion and Discussion

Taking on a sensemaking perspective opens up the complexity of the DBDM phenomenon, and sheds light on challenges and opportunities. It offers a way of looking at mechanisms and influencing factors that are at play when educational professionals engage with formal achievement data. Sensemaking provides a human-centered key to explain *how* data use is influenced by characteristics of data users, their organizations and their contexts, as well as characteristics of the data(systems) themselves. Characterizing data use as an act of sensemaking provides counterbalance to rational and deterministic models of data use and to “naïve” waterfall accounts of transforming data into knowledge. Those models and accounts assume clear and linear paths which simply do not occur in real-world sensemaking and decision-making (Klein et al., 2010).

We searched and reviewed the literature on sensemaking in DBDM, specifically with regard to formal achievement data, in order to provide some conceptual clarification and take stock of critical insights. While the knowledge base reviewed in the present paper does not capture every possible dimension or niche, we contend it can provide a good starting point to inform further research on educational professionals’ use of formal achievement data.

Our findings reflect the kaleidoscopic nature of sensemaking in DBDM. Firstly, the fact that “sensemaking begins with a sensemaker and is triggered by ambiguity” means that data should be considered as sensemaking resources (Fjørtoft & Lai, 2021). Formal achievement data add something to the world (Sellar, 2015), but they do not make sense on their own (Spillane, 2012). Teachers and school leaders make sense of data, and their own personal lenses and interests guide them in this process. Secondly, “sensemaking is an active search for coherence, aimed at understanding and action”. It entails an array of interpretive (micro)processes (Schildkamp, 2019) that can be active or unconscious, rational or intuitive (Vanlommel & Schildkamp, 2019), such as noticing, interpreting, inferring, valuing, judging, deciding et cetera. And once a sensemaker has found coherence, i.e., an explanation that makes sense to their own belief system about what will work (Bertrand & Marsh, 2015), they move on. Finally, “sensemaking is individual as well as social, cognitive as well as discursive”. Sensemaking is not an isolated affair, it happens in interaction with others and with one’ own multilayered context (Coburn & Turner, 2011). Sensemaking and sensegiving mean that you draw up an explanation and (are able to) articulate that explanation to yourself and/or to others.

Limitations

As we carried out this research, we were sensemakers as well. Selections and patterns did not emerge, they are a product of applying our own personal lenses. The result of our conceptual review is a broad, but by no means exhaustive overview of what sensemaking can mean in relation to educational professionals' use of formal achievement data. One (methodological) limitation we need to acknowledge, for instance, is that we narrowed down our search to studies that explicitly use sensemaking as a term or a keyword. That makes sense, given the fact that we aimed for a rather specific conceptual clarification. Nevertheless, we would like to emphasize that there are also other DBDM studies and lines of research that do not use the term, yet have a distinct sensemaking 'flavor' (for instance when discussing data literacy or organizational learning, or employing concepts that are front and center in sensemaking perspectives, such as mental models). A further exploration of sensemaking in DBDM could therefore include (more) conceptual snowballing. Furthermore, we did not include related or quasi-synonymous terms like 'meaning making', which would have potentially yielded more material. Finally, we limited our theoretical framework to a number of well-known sensemaking perspectives in order to shed light on the roots of sensemaking theory and give direction to our own review work. While this choice allowed us to highlight and substantiate a number of trends and salient themes, it also means we disregarded other bodies of work on sensemaking. Noteworthy examples are Dervin's take on sensemaking as a research methodology (Dervin, 1983, 2015) and work on academic/instructional sensemaking (Fitzgerald & Palincsar, 2019; Odden & Russ, 2019).

Suggestions for further research

Contemporary DBDM research has gradually incorporated a sensemaking logic that builds on foundations laid out by authors such as Coburn and Spillane, over Bertrand and Marsh's reconceptualization of the data use cycle, to recent work on intuition by Vanlommel and colleagues. In order to further advance this field, scholars call for more insight into the DBDM process and the sensemaking phase in particular, for instance through micro-process studies that also take into account how sensemaking unfolds in interaction (Christman et al., 2016; Mandinach & Schildkamp, 2021a; Schildkamp, 2019; Wardrip & Herman, 2018). Research, but also professional development initiatives, would benefit from insight into these micro-processes, for instance when this allows to make assumptions and attributions explicit (Bertrand & Marsh, 2015). In the same vein, more insight is needed into the competences required for sensemaking (Schildkamp, 2019) and into stages of intuitive expertise (Vanlommel & Schildkamp, 2019). Furthermore, future research can further illuminate how specific contexts and external resources affect the sensemaking process (Schildkamp, 2019). Educational professionals' sensemaking of data cannot be disconnected from their sensemaking and enactment of data use policies. Within the context of their schools, educational professionals juggle different expectations, perceptions and interests, emanating from different internal and external stakeholders. The way they balance these expectations and interests with local knowledge, beliefs and structures greatly impact the outcomes of data use (Jennings, 2012). In any case, what is clear from the knowledge base presented here is that when studying data use, we should not only look at outcomes but also at how the process of sensemaking unfolds in practice (Farrell & Marsh, 2016; Spillane, 2012; Vanlommel et al., 2021).

Based on insights from the studies that we reviewed, we propose that a sensemaking perspective will benefit future research on data literacy and user validity in particular. Insight into the malleability of factors that influence sensemaking, including mental models, but also affective responses to data, will benefit professional development and the development of assessment systems that can live up to expectations in terms of promoting school improvement. Methodologically, techniques such as discourse analysis might shed light on the ways different actors make sense of data and where their understandings diverge, for instance between test developers or policymakers versus data users, or teachers versus administrators. Furthermore, more longitudinal research will be able to capture how sensemaking is not only shaped by existing (inter)subjective beliefs, but also shapes future beliefs in an ongoing dynamic of enactment. With regard to the ‘shutter speed’ to employ in conceptual and empirical work on sensemaking (as a phase in the iterative DBDM cycle, cf. Schildkamp, 2019) it does make sense to unravel episodic ‘acts of sensemaking’, but at the same time we need to be mindful that sensemaking is ongoing. In their daily practice, educational professionals continually engage with different types of data, making observations and interpreting them, thereby adjusting their own mental models and growing organizational knowledge bases for future sensemaking along the way (Bertrand & Marsh, 2015; Coburn & Turner, 2011; Datnow et al., 2012; Even, 2005; Spillane, 2012; Wardrip & Herman, 2018).

Finally, the great majority of the papers reviewed in this study hails from high-accountability educational contexts. Exceptions are the both Vanlommel papers set in Belgium, the Norway example in the paper by Fjørtoft and Lai, and the Even paper that is focused more on contemporary assessment techniques. Although formal achievement data and scores from standardized tests do not necessarily need to be associated with accountability, an accountability discourse did permeate the findings and theories discussed in many of the selected papers. Future research on sensemaking of formal achievement data should try to contrast systemic narratives rooted in both low and high accountability paradigms, so it might inform both reciprocally.

Study 2

Principals' and teachers' comprehension of school performance feedback reports. Exploring misconceptions from a user validity perspective

BASED ON Goffin, E., Janssen, R., & Vanhoof, J. (2023). Principals' and teachers' comprehension of school performance feedback reports. Exploring misconceptions from a user validity perspective. *Pedagogische Studiën, 100*(1), 67–97. https://doi.org/10.59302/ps.v100i1.13991

ABSTRACT 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.

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 et al., 2017; Vanhoof et al., 2011; Verhaeghe et al., 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 et al., 2014; Kane, 2013b; O'Leary et al., 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 et al., 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 et al., 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.

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 et al., 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 et al., 2019; Means et al., 2011; Zapata-Rivera et al., 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 et al., 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 et al., 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 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 et al., 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 is 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, 2013b), 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 et al., 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 et al., 2021; Smith et al., 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 et al., 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 7 in Appendix B, 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 Appendix B, see Figure 8 and Figure 9. Please note that the figures and annotations in Appendix B provide background information needed to fully appreciate the setup of the data collection and the findings as presented in the following sections.

4 Methodology

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 et al., 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 7 in Appendix B) and caterpillar plots (see Figure 8 and Figure 9 in Appendix B) in their own time and "as if speaking to a colleague". The interviewer noted which components (see Table 12 and Table 13 in Appendix B) 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 4, 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.

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.

Table 4. 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.

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 12 and Table 13 in Appendix B, 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 5 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.

Table 5. 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 school 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.

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 7 in Appendix B) 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 7 in Appendix B). When interpreting the actual and expected score plot (see Figure 8 in Appendix B), 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.

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 extra’s 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 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 et al., 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.

Study 3

Feathers in our cap? Mapping educational professionals' internal and external attributions of school performance feedback

BASED ON Goffin, E., Janssen, R., & Vanhoof, J. (Submitted for peer review). Feathers in our cap? Mapping educational professionals' internal and external attributions of school performance feedback.

ABSTRACT After participating in external standardized assessments, schools are typically presented with performance feedback intended to inform self-evaluation. In the context of the Flemish national assessments, we conducted a qualitative study in which we investigated which causes teachers and school leaders invoke for their school's results as presented in an authentic school feedback report. We examined the locus of causality of these attributions and explored patterns according to participants' work role and perceived favorability of the feedback. Data were collected through 22 online semi-structured interviews, and subjected to a framework analysis. Attributions at the school-, class-, student-, and test-levels are discussed. In line with previous research, we find that school performance is attributed to external factors to a great extent. We also find that educational professionals make sense of school performance feedback from their own frame of reference. School leaders apply a policy outlook, while teachers reflect more on the input from students. Reservations about (the design of) the assessment emerge primarily to explain negative results. The finding that teachers and school leaders (even within schools) place different emphases to interpret the (same) outcomes highlights the importance of collective sensemaking. The observation that most participants mentioned a whole range of factors illustrates that people see learning outcomes as the product of different building blocks, but also that it is not easy to formulate an unambiguous diagnosis. Implications for practice and suggestions for further research are addressed.

1 Introduction

School performance feedback (SPF) systems present educational professionals with student achievement data in order to support self-evaluation and data-based decision making (Schildkamp & Teddlie, 2008; Visscher & Coe, 2003). However, using such data for school improvement is all but linear and straightforward. Data use involves a sensemaking process in which the raw data need to be analyzed, interpreted and contextualized (Goffin et al., 2022; Schildkamp, 2019; Vanlommel & Schildkamp, 2019) in order to turn those data into information and subsequently into knowledge that is 'actionable' in a particular context (Mandinach et al., 2008; Marsh, 2012; Marsh et al., 2006; Schildkamp & Poortman, 2015). Hypothesizing about potential causes for student outcomes is a fundamental part of this sensemaking process, is highly interpretive, and shapes the subsequent response (Coburn & Turner, 2011; Goffin et al., 2022; Schildkamp et al., 2016; Verhaeghe et al., 2010). In line with the basic propositions of attribution theory (Weiner, 1985, 2010) the nature of causal explanations for student outcomes has been found to affect educators' emotions and their subsequent (instructional) behavior (Wang & Hall, 2018).

Overall, studies demonstrate that educational professionals find it challenging to reflect on causes for student outcomes, especially when these outcomes are unfavorable (Verhaeghe et al., 2010). In a study on SPF use, school leaders indicated that they feel particularly lost in the diagnostic phase, not only because of a perceived lack of support and guidelines, but also due to a perceived lack of identified causes and concrete suggestions for improvement in the reports they received (Verhaeghe et al., 2010). Furthermore, research finds that teachers have a tendency to attribute student achievement to a great extent to external factors, such as what the student brings in, instead of (directly) relating it to matters internal to themselves, such as teaching-related practices (Evans et al., 2019; Van Gasse & Mol, 2021). This is especially apparent in cases of student failure (Wang & Hall, 2018). Even though reflecting about external causes can provide valuable insights, it is often based on incorrect assumptions (Evans et al., 2019; Schildkamp et al., 2016). Looking at internal factors and dynamics is a way of taking responsibility for student outcomes, and thus (believed to be) more productive in identifying areas for improvement (Schildkamp et al., 2016; Wang & Hall, 2018).

In the present study, we examine educational professionals' causal explanations for results presented in a SPF report from a low-stakes Flemish national assessment. We are interested to learn why they think their schools performed the way it did, and we want to particularly zoom in on the locus of causality of their attributions. To what extent is SPF interpreted introspectively and to what extent is performance ascribed to external factors? Put differently: to what extent do educational professionals interpret school performance as an accomplishment, and (good) results as proverbial feathers in their caps?

We will not only focus on teachers' attributions, but also on causal explanations made by school leaders. Our review of the literature, presented in Section 2, suggests that perceptions of school leaders remain underexposed in studies on attribution in educational data use, as the majority of studies appears to be concerned with the role of teachers' causal ascription. A potential reason for this is that teachers have the most direct and observable impact on student learning through instructional (micro) decision making (Schildkamp et al., 2016). However, SPF explicitly intends to inform both school policy and instructional practice.

Additionally, we examine causal explanations for both outcomes perceived as favorable, and those perceived as unfavorable. In line with the very term *diagnosis*, we find that the attributions and attributional processes discussed in empirical literature are predominantly focused on explanations for student failure (Van Gasse & Mol, 2021; Verhaeghe et al., 2010) and not so much for student success. This seems to be in line with educational practice: attribution is more prevalent or explicit when educational professionals seek to understand what goes wrong (and what may be done about it) rather than trying to understand what is going right (Evans et al., 2019; Nabors Oláh et al., 2010; Verhaeghe et al., 2010). However, school improvement is not only a narrative of identifying problems, difficulties and lacunas, but also of fostering what works.

In summary, the research questions we will address, are:

- RQ1 To which internal and external factors do teachers and school leaders attribute their school's performance on an external assessment?
- RQ2 Do attributions differ according to the attributor's work role (i.e., teachers versus school leaders)?
- RQ3 Do attributions differ according to attributors' perceived favorability of the result?

2 Theoretical framework

In order to theoretically inform the present study, we review the literature on causal searches that teachers and school leaders undertake when engaging with student academic achievement data such as school performance feedback (SPF) from external standardized assessments. We first set the scene by exploring the place of attribution in the data use cycle, and discussing attribution theory as a baseline framework (2.1). Next, we discuss locus of causality as a property of attributions, and reflect on why it is relevant for our case (2.2). Subsequently, we explore how perceived favorability can play a role in attribution (2.3). Finally, we present an overview of attributional models discussed in the literature on educational professionals' sensemaking of student achievement data (2.4). This overview will guide and frame how we look at our own empirical data in the present study.

2.1 Attribution and the data use cycle

Most conceptualizations of data use in schools, explicitly treat and study *diagnosis* as a distinct phase that follows data analysis, and feeds into the design of concrete actions and decisions (e.g., Nabors Oláh et al., 2010; Schildkamp et al., 2016; Van Gasse & Mol, 2021; Verhaeghe et al., 2010). Other perspectives, most notably Bertrand and Marsh's (2015) reconceptualization of the data use cycle, unravel attributions and their underlying dimensions to emphasize that attribution is actually present in all data use phases. A central premise underlying the latter view, is that people look at data through their own personal lenses – lenses shaped by individual knowledge and beliefs, prior experiences, and social and organizational contexts (Coburn & Turner, 2011; Goffin et al., 2022; Kelchtermans, 2009; Vanlommel & Schildkamp, 2019). As a micro-process that permeates educational professionals' sensemaking of student outcomes (Bertrand & Marsh, 2015), attribution is shaped by these personal lenses as well. When educational professionals formulate causes for student outcomes they are guided by the way they view and label individual students (Nabors Oláh et al., 2010; Wang & Hall, 2018) and by their general ideas about, for instance, content difficulty, or learning and instruction, or students in general, or even data use itself (Bertrand & Marsh, 2015; Evans et al., 2019; Nabors Oláh et al., 2010; Schildkamp et al., 2016).

An often-used framework to isolate and unpack educators' diagnoses of student outcomes is Weiner's (1985, 2010) attribution-based theory of motivation. Attribution theory describes mechanisms of causal ascription in achievement-related contexts, and positions causal ascriptions of success and failure on three distinct dichotomous properties. The perceived *locus of causality* can be internal or external to the attributor, i.e., the individual undertaking the causal search. Perceived *stability* refers to whether the cause is regarded as permanent or temporary, and perceived *controllability* describes whether or not the attributor thinks it is possible to change something about the identified cause. Together, these properties of attributions predict the psychological effects (e.g., pride, shame, anger, sympathy, gratitude, expectations about future endeavors) and subsequent behavioral effects (e.g., sustaining, trying, giving up, punishing, helping) that follow success or failure (Wang & Hall, 2018; Weiner, 1985, 2010). Attribution or *causal search* can be intrapersonal, when the attributor is the actor and analyzes their own success or failure, or interpersonal, when the attributor is judging the performance of others as an observer (Wang & Hall, 2018; Weiner, 2000). It is particularly likely to be undertaken when an outcome is unexpected, unfavorable, or important (Weiner, 2000, 2010).

In the present study, we regard the causal searches that educational professionals undertake when they process SPF as instances of interpersonal attribution (cf. Wang & Hall, 2018). This means we regard educational professionals as attributors and observers, and students as actors as they are the ones who scored, performed, achieved.

2.2 The locus of causality of attributions

As an aspect of an attribution made in a specific situation, locus of causality (or ‘causal locus’, ‘locus’ for short) describes whether attributors seek the cause of success or failure within themselves or elsewhere. Causal locus is believed to predict pride and self-esteem related to an accomplishment (Weiner, 1985, 2010). Internal attributions can pertain, for instance, to one’s own ability (e.g., “we failed because I don’t have the necessary skills”) or effort made (e.g., “this is a success because I worked really hard”). External attributions, on the other hand, situate the locus of causality with other people or other forces and occurrences.

In the literature, the term *locus of causality* is sometimes used interchangeably with the term *locus of control*. However, while they are related to a certain extent, these concepts are theoretically distinct (Weiner, 2010). Locus of control is a personality trait that can be situated on a continuum. It refers to people’s general convictions about the extent to which they have a hand in their own successes and failures (Rotter, 1966). As a data user characteristic, locus of control has been found to impact the effectiveness of data use (Schildkamp, Rekers-Mombarg, et al., 2012; Schildkamp & Kuiper, 2010), because it influences the locus of causality of attributions of student achievement, and thus the subsequent adjustments educational professionals will (not) make. Thus, locus of control can be regarded as an antecedent of attributions. Teachers with a higher external locus of control have been found to attribute (poor) student achievement and school performance on external assessments to external factors, such as (the motivation of) the specific student population at the time of measurement, or to the setup of the assessment itself, rather than relating it to their own functioning (Schildkamp & Kuiper, 2010). Teachers with a stronger internal locus of control have more faith in their own capacity to change things and in their power to influence student learning for the better, and are said to take cues from external assessment data accordingly, turning them into an impetus for change (Schildkamp, Rekers-Mombarg, et al., 2012).

Both locus of control and properties of concrete attributions can mediate the effect of success or failure on a person’s perception of self-efficacy (Bandura, 1977, 1997). For instance, when a teacher labels students’ successes as fluke (e.g., “they just got lucky by checking random answer options”) such ‘successes’ will not contribute to their own sense of instructional efficacy. Overall, teachers with a very outspoken internal locus of control will find that student success contributes to their job satisfaction, but student failure risks eroding their sense of self-efficacy (Kelchtermans, 2009).

The aforementioned mechanisms elucidate how the causal locus of educators’ attributions is associated with different perceptions of responsibility for student learning (Matteucci & Gosling, 2004). The idea is that internal attributions entail assuming a greater responsibility for outcomes, and will reinforce future behavior such as effort and persistence (Wang & Hall, 2018). From a review of the literature on teacher attributions, Wang and Hall (2018) report mixed findings as to the extent to which teachers assume personal responsibility for student outcomes. Matteucci and Helker (2018) find that teachers feel they are equally responsible as parents for overseeing students' learning *process* and providing a supportive *environment*, but ultimately place the greatest responsibility for learning *outcomes* with the students themselves.

The way teachers navigate (the boundaries of) their professional responsibilities can be interpreted in terms of their task perception, which is the normative component of teacher's professional self-understanding: what do I need to do to be a good teacher, what exactly is my job, to where does my (moral) responsibility towards students extend (Kelchtermans, 2009)? Teachers' professional responsibility involves making value-laden decisions about how best to address the needs and possibilities of their students, and on a day-to-day basis, extends beyond strict accountability for measurable student outcomes in terms of academic achievement (Kelchtermans, 2011, 2018). Both in broad and in narrow interpretations of responsibility, teaching is fundamentally vulnerable: teachers can make a difference to a certain extent, yet there are always external factors that also have an impact on student outcomes (Kelchtermans, 2009, 2011, 2018).

2.3 Effects of (perceived) feedback favorability on attributions

Feedback sign, or feedback valence, is conceptualized as the direction of the discrepancy between the recipients' behavior as described in the feedback messages on the one hand, and the behavioral goals, standards or ideals they uphold on the other hand (Kluger & DeNisi, 1996; Podsakoff & Farh, 1989). A feedback message is generally considered unfavorable by a recipient when the feedback sign is negative, i.e., when there is a perceived negative gap between reported performance and pursued, desired or expected performance. In the present study, following Wang and Hall (2018), we regard success and failure (within in the attributional frame) not as absolute classifications but rather as *satisfactory* versus *unsatisfactory* outcomes. Student achievement data and SPF do not always exhibit a clear-cut pass/fail pattern. Moreover, whether a certain outcome is a success or failure, pertains to the perception of the attributor: appraising a score as high or low is part of sensemaking as well (Coburn & Turner, 2011).

Research has found that people interact differently with feedback messages that are perceived as favorable or unfavorable (Lechermeier & Fassnacht, 2018). Overall, feedback perceived as positive is associated with a better and more detailed recollection of the message, a higher level of perceived feedback credibility, and higher acceptance, at least in part because it positively reinforces recipients' self-image (Anseel & Lievens, 2009; Ilgen et al., 1979; Lechermeier & Fassnacht, 2018; Podsakoff & Farh, 1989). Unfavorable feedback messages are more at risk of being avoided or overlooked because they activate defense mechanisms in the recipient. This is problematic because negative feedback needs to be accurately interpreted in order to correctly and effectively guide or influence subsequent behavior (Ilgen et al., 1979).

In educational professionals' sensemaking of student achievement data, it appears that the perceived favorability of an outcome can determine whether or not attributions are formulated at all. In practice, teachers' instructional decision making is largely focused on detecting weaknesses in order to come up with ways of addressing them (Evans et al., 2019; Nabors Oláh et al., 2010). Verhaeghe et al. (2010) find that school leaders are motivated to undertake a causal search when SPF from an external assessment is unsatisfactory, but not so much when the results are satisfactory. Interestingly, however, when the results are so disconcerting that it would bring down the team, causal search is abandoned (Verhaeghe et al., 2010).

Some studies on (general) feedback effects find that negative feedback boosts response and subsequent performance because recipients make a bigger effort in an attempt to smooth out the negative gap between reported performance and desired performance (Kluger & DeNisi, 1996; Mesch et al., 1994; Podsakoff & Farh, 1989). However, differential effects of feedback sign have been found. They are explained through mechanisms of self-regulation: people who are in promotion focus (seeking to fulfill a certain desire) are more inclined to act upon positive feedback, while those in prevention focus (seeking to prevent negative repercussions) respond more strongly to negative feedback (Van-Dijk & Kluger, 2004).

Finally, research finds that teachers tend to attribute student success to factors internal to themselves, such as the instruction they provided, while ascribing student failure more to external factors, such as a lack of effort or motivation on the students' part (Schildkamp et al., 2016; Wang & Hall, 2018). This appears to be in line with hedonic, self-enhancing or *self-serving bias*: people's tendency to ascribe success to internal factors and failure to external factors, or to assume personal responsibility more readily in case of a favorable outcome (Miller & Ross, 1975; Wang & Hall, 2018).

2.4 Attributional models

Educational professionals' causal explanations of student outcome data are situated on different levels. School leaders and (particularly) teachers have been reported to attribute satisfactory and unsatisfactory student results to aspects of the test (or the assessment, the measurement), to characteristics of the students themselves, and to features of the school and the classroom as well as the people in those schools and classrooms. Attributions on all of these levels can be interpreted as having an external or internal locus of causality, and are sometimes associated with particular types of data use.

Test-level and student-level attributions (external)

When assessment results defy educational professionals' prior expectations, they are sometimes met with suspicions and validity concerns (Nabors Oláh et al., 2010), which can give rise to external attributions pertaining to the nature and setup of the *assessment* itself (Verhaeghe et al., 2010). Such attributions are made with regard to the quality, validity and usefulness of the assessment as a whole (Evans et al., 2019; Verhaeghe et al., 2010), sometimes also to the one shot nature of an assessment (Schildkamp & Kuiper, 2010; Verhaeghe et al., 2010) and to the difficulty or formulation of specific test items (Bertrand & Marsh, 2015; Nabors Oláh et al., 2010; Verhaeghe et al., 2010). In some cases there is overlap with other explanatory frames, as the attributions pertain to the alignment of the test with the curriculum, with the content already taught in class, with the cognitive abilities, vocabulary or language proficiency of the target group students (Bertrand & Marsh, 2015; Evans et al., 2019; Nabors Oláh et al., 2010).

Some authors discuss teachers' 'validity check' of those test items that students performed particularly low on as a step that precedes the diagnostic phase (e.g., Nabors Oláh et al., 2010). Bertrand and Marsh (2015), however, discuss concerns about the nature of the test as a model of attribution in its own right, and find it is often combined with attributions to student understanding (e.g., "the assessment is ill-aligned with what we can expect of students"). They find it seldom contributes to the formulation of a subsequent instructional response.

Attribution of achievement to observed and/or presumed characteristics pertaining to *students*, is ubiquitous in educational professionals' sensemaking of student achievement. Outcomes are explained in terms of student characteristics in multiple ways: (a) student understanding and ability (Bertrand & Marsh, 2015; Lasater et al., 2021; Wang & Hall, 2018), both in general and with regard to specific subjects, domains or skills (Nabors Oláh et al., 2010; Van Gasse & Mol, 2021); (b) student general (cognitive) weaknesses, such as entry level or background knowledge, language proficiency or reading skills; (c) behavioral characteristics of students, such as (learning) attitudes, motivation, focus and effort (Bertrand & Marsh, 2015; Evans et al., 2019; Nabors Oláh et al., 2010; Schildkamp et al., 2016; Van Gasse & Mol, 2021; Wang & Hall, 2018); (d) student personality traits, emotional issues, learning disorders and medical conditions sometimes pinpointed as underlying causes for the aforementioned factors (Evans et al., 2019; Van Gasse & Mol, 2021; Wang & Hall, 2018), and finally (e) students' home environment, for instance with regard to (lack of) resources and (lack of) parental involvement and support (Evans et al., 2019; Lasater et al., 2021; Schildkamp et al., 2016; Van Gasse & Mol, 2021; Wang & Hall, 2018).

As Evans et al. (2019) point out, the observation that teachers tend to turn to external factors to explain student failure (instead of factors internal to themselves), discourages the theoretical ideal that relating student performance to instruction is the essence of effective data-based decision making. However, they state, perhaps we should not discount or dismiss educators' external attributions, yet zoom in on those claims in order to distinguish between those that are *helpful*, because they bear on knowledge of students, assessment and instruction, and those that are *harmful*, because they are biased, ill-founded or unproductive and reinforce bias and inequity (Evans et al., 2019, p. 26).

Class-level and school-level attribution (internal or external)

Attributions of student achievement sometimes relate to factors situated in *schools or classrooms* (i.e., these are things that happen at school), but external to educational professionals as attributors. Examples are the influence of classroom environment (Evans et al., 2019), the perceived difficulty of specific subjects or domains (Nabors Oláh et al., 2010; Schildkamp et al., 2016), the design of (policy-mandated) curricula (Evans et al., 2019), and, notably, instruction provided to the students by previous educators (Nabors Oláh et al., 2010; Schildkamp et al., 2016).

Overall, teachers' internal attributions reported in the literature, mainly pertain to perceptions about (the quality of) their own *instruction* (Bertrand & Marsh, 2015; Wang & Hall, 2018), such as their effectiveness in teaching specific content or skills (Evans et al., 2019; Schildkamp et al., 2016), in offering tailored support and instruction (Schildkamp et al., 2016), in planning and preparing (Evans et al., 2019), or in terms of (other) aspects of their (general) functioning (Schildkamp et al., 2016). *School*-level factors are less prevalent, or in any case less reported in the literature (Bertrand & Marsh, 2015). School-level attributions are, for instance, concerned with general aspects of school policy, policy regarding failing students or absenteeism, or curriculum coherence (Evans et al., 2019; Schildkamp et al., 2016).

Whether these attributions should be regarded as internal or external, depends on (the perception of) the person or the team that utters the attribution. Work roles determine to a great extent the purpose of data use, and consequently the perceived relevance of data, as well as the nature of attributions made. In order to formulate instructional decisions, teachers are used to relying primarily on classroom-level, student-level and item-level data, while school leaders use and seek out aggregated school level data for policy making (Coburn & Turner, 2011; Schildkamp & Kuiper, 2010; Verhaeghe et al., 2010). Furthermore, for teachers, what is considered as 'internal' or 'external' may even be interpreted in terms of Hoyle's (1974, 2008) distinction between *restricted* and *extended* professionalism (or 'professionality', in earlier conceptualizations). Teachers with *restricted* professionalism take on a task-oriented stance towards the teaching profession, rely on their professional intuition, and focus on what happens in their classroom. Teachers with a more or less *extended* professionalism look beyond the classroom, are involved in school policy, seek out input, feedback and collaborations, and overall regard teaching as a rational practice that is open to continuous improvement (Hoyle, 1974, 2008; Jongmans et al., 1998; Jongmans & Beijaard, 1997).

Numerous studies report that teachers focus primarily on external factors when detecting problems relating to student achievement (Evans et al., 2019; Schildkamp et al., 2016; Van Gasse & Mol, 2021; Wang & Hall, 2018) and turn to internal factors (only) when prior hypotheses about external causes for student failure have been invalidated, or when they actively reflect about their own functioning in order to address external factors (Schildkamp et al., 2016). The latter observation suggests that sequential patterns of attribution are important, and make clear that attribution depends (at least to a certain extent) on the purpose of a data discussion as a starting point. For instance, Bertrand and Marsh (2015) focus on educational professionals' sensemaking of the performance of English Language Learners and special needs students. While they find that teachers invoke mental models relating to external factors such student understanding, student characteristics and the nature of tests and assessments, the teachers in their study do relate the majority of their attributions to their own instruction. This should not be remarkable as the very purpose of the sensemaking exercises analyzed in this study, was to figure out how to tackle difficulties.

3 Methodology

3.1 Research context and participants

This study was conducted in Flanders, the Dutch-speaking region of Belgium. In order to optimally inform our research aim, we used a combination of convenience sampling and purposive sampling (Cohen et al., 2018; Patton, 2015; Savin-Baden & Major, 2013). The study took place within the context of the 2019 national assessment (NA) of *People and Society* (formerly a subdomain of the world studies curriculum) in the sixth grade of primary education. The main purpose of this NA was to collect information at the system-level with regard to the proportion of students reaching the attainment targets, and with regard to school-, class- and student-level variables that explain differences in achievement. However, after the NA, participating schools received a personalized school performance feedback (SPF) report describing to what extent attainment targets were realized (criterion-referenced information), as well as how the school performed compared to the national results and to schools with statistically similar student populations (norm-referenced information). The feedback focused on school-level results and did not include individual student results, but criterion-referenced information was presented per class if applicable. Note that school results from a NA are never publicized, and outcomes do not carry any formal consequences for the participating schools.

Recruitment focused on schools that had received a full feedback report (*N*=99). In order to achieve a sufficient degree of standardization, we selected one focal test, namely *Spatial use, traffic and mobility*. In pursuit of maximum variation (Cohen et al., 2018; Patton, 2015; Savin-Baden & Major, 2013), the schools that received feedback on this particular test (*N*=57) were categorized into four profiles based on their criterion-referenced and norm-referenced results, as indicated in Table 6. This categorization of schools into scoring profiles was done exclusively for the purposes of the present study and was not communicated to schools. Also, profile assignment did not necessarily coincide with participants' personal perceptions. However, by making sure that we included schools that systematically varied in terms of achievement in the NA, we did aim to increase the likelihood that individual participants would also differ sufficiently in terms of how they appraised their results. This would allow us to explore patterns according to perceived favorability of the SPF (cf. RQ3). Approximately one week after having received the report, a random selection of schools within each profile was approached. In total, 22 interviews were scheduled with 23 participants from 13 schools.

3.2 Data collection

We opted for a semi-structured interview guide approach in order to set up a conversational interaction that nevertheless allowed to comprehensively and systematically touch upon the topics of interest in depth (Cohen et al., 2018; Savin-Baden & Major, 2013). Open-ended questions about participants' appraisal of the schools' results on the selected focal test, about how they explained these results, and about how they relate the results to their functioning, were embedded in a protocol that, in full, served a broader research interest, i.e., to shed light on the overall individual and collective sensemaking process of SPF. Examples of questions that were intended to gauge perceived favorability included the prompt "Can you tell me in your own words how your school performed on this test?" and probing questions such as "Are you satisfied with this result?". Attributions were elicited with questions such as "How come you [Dutch plural form *jullie*] obtained this result?" (prompt) and "What causes do you see for this result?" (example of a probing question).

Prior to the interviews, participants were informed about the general goals of the study and the ethical clearance obtained. They were told they did not need to prepare. In order to zoom in on individual perceptions and pay heed to the potentially sensitive nature of considerations shared by the participants (Savin-Baden & Major, 2013) all interviews were conducted one-on-one, with the exception of the leadership interview in School 05, which included both Jenny and Melanie. As Jenny was not able to partake in the full interview, and because we found during data analysis (see Subsection 3.3) that only one attributional statement was recorded for this participant, her data were excluded from the analyses.

Table 6. Participants

Profile	Criterion [a]	Norm [b]	School [c]	Participant [c]	Role	Position	Experience current school[d]	Experience education [d]	Age [d]	Gender	Degree
A	high	higher	01	Valerie	leader	principal	2	13	36	female	master
				Sandra	teacher	6th grade teacher	4	6	37	female	bachelor
			02	Rebecca	teacher	6th grade teacher	2	5	53	female	bachelor
			03	Paula	leader	principal	15	15	36	female	bachelor
			04	Frank	leader	principal	31	32	52	male	bachelor
				Natalie	teacher	6th grade teacher + teacher-mentor	13	15	36	female	bachelor
B	high	lower	05	Jenny [e]	leader	principal	9	28	50	female	bachelor
				Melanie	leader	policy support + 5th grade teacher	10	10	33	female	master
				Laura	teacher	6th grade teacher	11	18	39	female	bachelor
			06	Heidi	teacher	6th grade teacher	4	6	26	female	bachelor
			07	Gina	leader	principal	34	34	54	female	bachelor
				Erika	teacher	6th grade teacher	8	15	36	female	bachelor
			08	Isaac	leader	principal	3	16	39	male	bachelor
C	low	higher	09	Ken	leader	principal	32	32	55	male	bachelor
				Oscar	teacher	6th grade teacher	9	9	29	male	bachelor
			10	Denise	leader	principal	21	21	43	female	bachelor
				Quentin	teacher	6th grade teacher + care teacher + IT support	6	7	30	male	bachelor
D	low	lower	11	William	leader	principal	1	21	42	male	bachelor
				Tony	teacher	6th grade teacher + prevention officer	26	26	51	male	bachelor
			12	Brenda	leader	student care coordinator	12	13	55	female	master
				Catherine	teacher	6th grade teacher	18	18	39	female	bachelor
			13	Andrea	leader	principal	25	40	60	female	bachelor
				Xavier	teacher	6th grade teacher	10	10	31	male	bachelor

Notes.

[a] high = >70% of pupils reach the attainment targets assessed in the focal test

[b] higher/lower score on focal test then expected based on student population characteristics

[c] pseudonymized

[d] in years

[e] not included in the data analysis

All interviews were conducted by the first author. Due to societal restrictions relating to the COVID19 pandemic, the interviews were conducted online using video-conferencing software, which provides a comparable level of synchronicity as in-person interviews and has become widely accessible in recent years (Lo Iacono et al., 2016; Savin-Baden & Major, 2013). Video-conferencing tools allow for virtual 'face-to-face' contact, are evaluated as convenient by researchers and participants alike, and provide screen- and file-sharing options to facilitate engagement (Archibald et al., 2019; Sullivan, 2012). The interviews were audio and video recorded with consent of the participant, and transcribed verbatim.

3.3 Data analysis

The transcriptions were coded, analyzed and organized in NVivo. We applied the Framework method, a pragmatic and paradigm-independent analytical approach that is closely liaised to thematic analysis (e.g., Braun & Clarke, 2006) and matrix-based methods of data display (e.g., Miles et al., 2014) (Gale et al., 2013; Parkinson et al., 2016). Framework analysis allowed us to stay close to the raw data, developing an analytic framework by flexibly moving between different levels of abstraction, and charting and mapping data and findings in order to identify and present themes and patterns (Ritchie et al., 2003; Ritchie & Spencer, 1994). The interviews were coded phrase-by-phrase in their entirety, also including relevant (spontaneous) statements made outside of the designated protocol sections. Utterances and statements were isolated for coding. By moving back and forth between open (or initial) coding and axial coding (Cohen et al., 2018; Saldaña, 2013) in a process of constant comparison (Savin-Baden & Major, 2013) we gradually developed the analytical framework.

The core of the analytical framework pertained to participants' attributions of the schools' outcome. Inspired by the range of attributions identified in our review of the literature (see Subsection 2.4) we coded towards a typology of factors that describe what happens in schools, what happens in classrooms, what students bring in, and what pertains to the assessment. In some cases, segments could be interpreted in multiple ways. We did not double-code but assigned one code deemed most appropriate. We applied analytical codes (Cohen et al., 2018) for locus of causality: student-level and test-level attributions were coded as external, class-level and school-level attributions were coded as internal or external based on participant role and on the nature and the context of the statement.

In a separate step, statements about participants' perceptions of the SPF's favorability were isolated. They were coded as positive (e.g., "So we're doing well, right! We're doing fine compared to what's expected", Valerie, leader, School 01), negative (e.g., "So our school is actually underperforming?", Andrea, leader, School 13) or mixed (e.g., "In the end I think it's a nice result, and yet I do feel frustrated", Heidi, teacher, School 06). This information was put into a case-based matrix in order to assign each participant one overall code for perceived favorability (positive, negative or mixed).

When designing the study and assigning prospective schools to profiles according to two scoring dimensions (see Subsection 3.2 and Table 6), we expected to be able to record different perceptions and attributions for criterion-referenced aspects of the SPF on the one hand, and norm-referenced aspects on the other hand. However, participants seldomly discussed both dimensions separately, even when we tried to stimulate this with focused prompts during the interviews. Overall, participants' appraisal of the results and their reflection on potential causes tended to pertain to the SPF as a whole. Since the distinction between criterion-referenced and norm-referenced results was not sufficiently clear-cut in the data, we decided it would be inappropriate to artificially pursue this distinction in the analyses.

Table 7. Thematic coding scheme including examples and frequency counts

Category	Code	Locus	Example	Attributions made by:						Mentioned at least once by:					
				All participants (total: 258)		School leaders (total: 122)		Teachers (total: 136)		All participants (total: 22)		School leaders (total: 11)		Teachers (total: 11)	
Test				**43**	**17%**	**15**	**12%**	**28**	**21%**	**18**	**82%**	**8**	**73%**	**10**	**91%**
	One-shot nature	external	*But I do think it is unfortunate - even though such an assessment is just a snapshot, maybe half of your class was having a bad day. (Oscar, teacher, School 09)*	12	5%	4	3%	8	6%	9	41%	3	27%	6	55%
	Item formulation	external	*Are the children familiar with the way the questions are asked? Are they being offered that [by the teacher] and do they practice it? (Andrea, leader, School 13)*	11	4%	4	3%	7	5%	9	41%	3	27%	6	55%
	Content	external	*That it has to do with the way the content was provided, because maybe that doesn't quite fit with what we do in the classroom. (Tony, teacher, School 11)*	10	4%	5	4%	5	4%	8	36%	4	36%	4	36%
	Conditions	external	*Plus there was that lady who acted super mysterious. [...] The one who comes to administer the tests, she is a very serious person. [She] stands there with the box saying "yes, now, we may open it". Completely different from how I do it and the children aren't used to that either. (Laura, teacher, School 05)*	10	4%	2	2%	8	6%	6	27%	2	18%	4	36%
Student				**83**	**32%**	**25**	**20%**	**58**	**43%**	**20**	**91%**	**10**	**91%**	**10**	**91%**
	Capacity	external	*Well, with good, clever kids it's easier to get good grades of course. (Ken, leader, School 09)*	33	13%	8	7%	25	18%	16	73%	6	55%	10	91%
	Home and parents	external	*We have a lot of students [whose] parents aren't very involved and I think that also plays a very big role. In order to get something done from children, there also needs to be a very strong team behind them at home. (Catherine, teacher, School 12)*	17	7%	7	6%	10	7%	10	45%	5	45%	5	45%
	Language	external	*I am looking at the individual students. You might say that this student may have scored less because he's a non-native speaker. (Quentin, teacher, School 10)*	10	4%	1	1%	9	7%	8	36%	1	9%	7	64%

Table 7 (Continued)

				Attributions made by:						Mentioned at least once by:					
Category	Code	Locus	Example	All participants (total: 258)		School leaders (total: 122)		Teachers (total: 136)		All participants (total: 22)		School leaders (total: 11)		Teachers (total: 11)	
	SES	external	*In terms of population at our school, we don't have a lot of non-native students. I know that if you have a lot of non-Dutch-speaking students in the class, that is also a factor that contributes to the fact that the level might go down a bit. (Natalie, teacher, School 04)*	8	3%	3	2%	5	4%	7	32%	3	27%	4	36%
	Wellbeing	external	*We are a school that does not work in the traditional way, and because of that we deal with a lot of children with issues, and that also makes that we score a bit lower. Because we mainly have children with autism, children with concentration and attention problems. (Gina, leader, School 07)*	8	3%	4	3%	4	3%	7	32%	4	36%	3	27%
	Motivation	external	*That class at the time was a very clever class overall, and really wanted to perform, was very performance-oriented. Students did their utmost. So is stress may also part of that somehow? (Heidi, teacher, School 06)*	7	3%	2	2%	5	4%	5	23%	2	18%	3	27%
Class				**55**	**21%**	**31**	**25%**	**24**	**18%**	**17**	**77%**	**9**	**82%**	**8**	**73%**
	Instruction	external	*Well, then we actually need to look at whether our teachers are really working to achieve the attainment targets. (Andrea, leader, School 13)*	6	2%	6	5%	0	0%	4	18%	4	36%		0%
		internal	*Not all children are equally interested but I think you can encourage that in the way you approach your lessons - and for world studies that is easier than for math for example. (Sandra, teacher, School 01)*	20	8%	4	3%	16	12%	8	36%	2	18%	6	55%
	Transfer	external	*In the classroom [...] they go digital very quickly because it's also faster. Kids pull out the Chromebook, type and type and hup! They've got it! While specifically - because, if I seem to remember, there was an assignment in there with an atlas as well - how many times do they actually get their hands on an atlas? Just in those lessons when they really have to. (Denise, leader, School 10)*	1	0%	1	1%	0	0%	1	5%	1	9%	0	0%

Table 7 (Continued)

				Attributions made by:						Mentioned at least once by:					
Category	**Code**	**Locus**	**Example**	**All participants (total: 258)**		**School leaders (total: 122)**		**Teachers (total: 136)**		**All participants (total: 22)**		**School leaders (total: 11)**		**Teachers (total: 11)**	
	(Transfer)	internal	*When we go somewhere we do it by bike and we make stops to give some explanation. But that's not with books, that's not with theory. (Tony, teacher, School 11)*	13	5%	7	6%	6	4%	9	41%	4	36%	5	45%
	Materials	external	*If I'm honest, in terms of world studies - the method we use in our school, devotes little attention to traffic. [...] We had to put [a line] together but our method doesn't offer that. (Xavier, teacher, School 13)*	6	2%	5	4%	1	1%	4	18%	3	27%	1	9%
		internal	*At the request of the teachers, we searched for another way, or another methodology or method whatever you want to call it, to instruct about traffic, because the teachers found the old method, the old way, less and less up-to-date among other things, and they found themselves less and less comfortable with it. (Paula, leader, School 03)*	3	1%	3	2%	0	0%	3	14%	3	27%	0	0%
	Teacher professio-nalism	external	*We have a teacher team that relies heavily on the method. And then when you combine that with the fact that sometimes the method is not the right tool for our audience… But they find it very difficult to adjust, to deviate. (Brenda, leader, School 12)*	5	2%	5	4%	0	0%	4	18%	4	36%	0	0%
		internal	*Teachers are only human as well, and you prefer to teach some things over others, and what you like to teach, you are going to teach very well and very intensively. The things you don't like to teach, you will teach, but maybe less intensively. (Oscar, teacher, School 09)*	1	0%	0	0%	1	1%	1	5%	0	0%	1	9%
School				**77**	**30%**	**51**	**42%**	**26**	**19%**	**20**	**91%**	**11**	**100%**	**9**	**82%**
	Curricular line and emphasis	external	*I think it's also a process throughout the children's school career, what have they been offered along the way? (Paula, leader, School 03)*	1	0%	1	1%	0	0%	1	5%	1	9%	0	0%
		internal	*World studies is a bit of a, well, not really a second rate subject but ... we do focus a lot on math and language. (Brenda, leader, School 12)*	47	18%	28	23%	19	14%	19	86%	10	91%	9	82%

Table 7 (Continued)

				Attributions made by:						Mentioned at least once by:					
Category	**Code**	**Locus**	**Example**	**All participants (total: 258)**		**School leaders (total: 122)**		**Teachers (total: 136)**		**All participants (total: 22)**		**School leaders (total: 11)**		**Teachers (total: 11)**	
	Community	external	*It's a credit even to the parent council, which is strongly engaged in the traffic work group. I think you can also say that to those people: "Look, all the energy you put into it, pays off in terms of those kids reaching the attainment targets.” (Frank, leader, School 04)*	2	1%	2	2%	0	0%	1	5%	1	9%	0	0%
	Location	external	*We live in a pretty expansive area here, where the kids have a lot of opportunity to practice and see things. (Paula, leader, School 03)*	12	5%	8	7%	4	3%	8	36%	5	45%	3	27%
	Staff and collabo-ration	internal	*Like I said it's really convenient for us now that we have a designated traffic teacher, so more traffic lessons are taught anyway and more thought is put into that. (Erika, teacher, School 07)*	7	3%	6	5%	1	1%	4	18%	3	27%	1	9%
	Infra-structure	internal	*What is not covered in the manual we have, do we have an adequate manual there? Do we need to buy something for that? Are there enough resources? Uhm, things like that. (Isaac, leader, School 08)*	3	1%	3	2%	0	0%	2	9%	2	18%	0	0%
	Other policy aspects	internal	*I think a very strong asset of our school is that we intensely mentor children who are having a very difficult time so we can get them that one step higher - so it doesn't actually surprise me that we score a little bit higher than what they would expect from us. (Oscar, teacher, School 09)*	5	2%	3	2%	2	1%	5	23%	3	27%	2	18%

4 Findings

In the following subsections, we will first narratively present and discuss participants' attributions and the locus of causality of these attributions. Subsequently, the combined based matrix that took shape during analysis, will serve to cautiously 'quantitize' findings by way of frequency counts (Miles et al., 2014; Sandelowski et al., 2009; Teddlie & Tashakkori, 2009). The aim here is not to reduce nor to generalize the qualitative data, but to explore trends within and across categories, attributes and perceptions.

4.1 Educational professionals' attributions of school performance

The first, broad research question we sought to explore, pertains to the factors that educational professionals invoke when reflecting on causes for their school's performance in an external assessment (RQ1). We find that, throughout the dataset, school performance is attributed to both internal and external factors. Attributions come up from all four 'levels' or categories we identified in the literature: student-, class, school-level and the level of the test itself. In Table 7, a detailed overview is given of the different codes and sub-codes assigned to participants' causal ascriptions. This table includes illustrative quotes lifted from the dataset, as well as general frequency counts to give a cautious idea of the prevalence of different themes. Frequency counts are presented for the dataset as a whole, as well as split up by work role.

Locus of causality

Adhering to the logic that causal locus needs to be interpreted as relative to the attributor, student-level attributions are interpreted as instances of external attribution. When teachers and school leaders attribute their school's result to what the students bring in, they are reflecting on an external cause. The same is true for test-level attributions: the NA was externally developed, scored and analyzed, and test administration was overseen by a proctor.

School- and class-level attributions were not a priori coded as internal. In order to determine the locus of causality of school- and class-level attributions, we took into account a number of elements. First, we observed that some things may well happen at or pertain to the school or class, but are contextual and therefore external to the attributor. An example is the *location* of the school, which is sometimes referred to in order to explain why traffic and mobility topics could (not) (sufficiently) be instructed in, or transferred to, real-world settings. Another example is that, on class-level, *materials* can be external (e.g., what does the textbook/method cover) or internal (e.g., to what extent is the method adhered to). Second, we tracked who was the attributor, i.e., the source of the attribution. For instance, was it a teacher or school leader speaking about teachers' instruction? Third, we examined how the statement was formulated. In certain cases, explicit use of the I- and we-forms suggests that attributions were perceived as internal by the person who uttered them.

Test-level attributions

Attributions on test-level pertain to the *one-shot nature* of the NA, the *content* of the test, the *formulation of test items*, and the *conditions* under which the test was administered. As illustrated by some of the quotes in Table 7, a number of these attributions refer to a perceived disconnect between the test and regular classroom practice. Another perceived disconnect relates to the concern that the test does not accommodate students' diverse needs and individual capacities, and can be perceived as daunting. Attributions about the content of the test and the formulation of test items are sometimes formulated more as attributional hypotheses, particularly by participants who have not seen the test at all and/or were not present while it was administered. So, to a certain extent, these attributions pertain to presumed characteristics of the test rather than observed characteristics.

Student-level attributions

Most of the attributions on student-level relate to student *capacity*. Other student-level attributions touch upon students' (more or less stimulating/supportive) *home situation and parents, language* issues*, SES*, *wellbeing*, and *motivation*. The latter concerns students' drive to perform in general or their interest in the subject that was tested. Home environment and parental support are often linked to the specific subject matter of the test at hand: particularly, the traffic and mobility dimensions. Participants reflect on whether or not students are provided by their parents with an additional opportunity to learn in real-world settings and day-to-day contexts (e.g., parents here just drop their kids off by car, parents here take their children on cycling holidays). In some cases, student characteristics are linked to the school (e.g., *we* have motivated students, *we* have a lot of non-native speakers).

Class-level attributions

Class-level attributions pertain to *instruction, transfer*, and (to a lesser extent) *materials*, and *teacher professionalism*. As can be inferred from the example quotes in Table 7, there is some overlap between *instruction* and *transfer*, but the latter code was assigned to statements that specifically address practical aspects of the tested domain that need to be experienced in real-world settings (e.g., we go cycling a lot, we don't just learn these things from books). Furthermore, there is conceptual overlap between some class-level factors and the factors we labeled as school-level. Particularly, the code *materials* refers to the didactical supports that teachers use in class, and therefore we included it on that level. However, it could be argued that this is a school-level variable, leaning on *curricular line and emphasis*. We did notice that school leaders invoke this *materials* factor more often than teachers themselves.

School leaders' class-level attributions can be external, when they talk about wat their teachers do in daily practice and how they assume this does or does not contribute to achievement. Most references to teacher professionalism (e.g., you need teachers with sufficient content knowledge) are for instance external. However, a large proportion of school leaders' class-level attributions are formulated as internal, in the we-form (e.g., we should pay more attention to differentiation in world studies education, we organize cycling exams in different grades).

Whereas school leaders' class-level attributions have some touching points with aspects of school policy, a number of teachers' class-level attributions touch upon student-level factors (e.g., not all students are sufficiently interested or motivated but you can address that in your teaching). Notably, teachers' internal class-level attributions, pertaining particularly to aspects of *instruction* and *transfer*, are seldom formulated in the I-form (e.g., knowing this, I might want to focus on this domain a bit more intensively in the future). Rather, statements do not specify a subject (e.g., as a teacher you sometimes focus too much on those students who aren't on board yet) or are often made in the we-form (e.g., we try to go outside a lot).

School-level attributions

On school-level, a number of participants mention contextual factors, particularly, the *location* of the school (and the perceived opportunity to learn and practice in real-world settings) and in one case (see Table 7) the *community* (referring primarily to parental involvement in school policy). The majority of the other school-level attributions, uttered by school leaders and teachers alike, can be regarded as internal. They pertain primarily to the *curricular line* set out for the domain under scrutiny and the *emphasis* that is put on it within the school's pedagogical process. Other school-level factors concern policy relating to *infrastructure* (e.g., we make sure that we are properly equipped to instruct students in this domain), *staff and collaboration* (e.g., we have a designated traffic teacher, the teacher team configuration was particularly unstable during the period the NA took place), and a general category referring to *other policy aspects* (e.g., how pupils are assigned to classes based on their general cognitive capacity). Almost exclusively, internal school-level attributions are formulated in the we-form, regardless of whether the attributor is a teacher or a school leader.

General trends

For the sake of additional clarity, the frequency counts from Table 7 are summarized for category and locus of causality in Table 8. Based on these simple frequency counts of coded statements, and without taking into account non-unique ideas, we find that attributions on school-level and student-level are overall uttered more frequently than attributions on class-level and test-level. School- and student-level attributions each make up almost one third of all attributions. Referring back to Table 7, we see that school-level factor *curricular line and emphasis* (describing the way world studies is approached within the school, 19% of all coded utterances) and the student-level factor *capacity* (referring to students' cognitive abilities, 13% of all coded utterances) are mentioned most. Furthermore, a majority of attributions (159 out of 258, or 62%) relate to factors external to the attributor.

Table 8. Overall prevalence of attributions in the dataset (frequency counts)

		Attributions made by:					
		All participants (total: 258)		**School leaders (total: 122)**		**Teachers (total: 136)**	
Category							
	Test	43	17%	15	12%	28	21%
	Student	83	32%	25	20%	58	43%
	Class	55	21%	31	25%	24	18%
	School	77	30%	51	42%	26	19%
Locus							
	External	159	62%	68	56%	91	67%
	Internal	99	38%	54	44%	45	33%

In order to facilitate the exploration of differences in attribution according to the attributor's work role (RQ2) and according to the perceived favorability of the SPF (RQ3) in the next subsections, we supplement Table 7 and Table 8 with Table 9. Table 9 visualizes the 'intensity' of attributions on participant level. Data have been sorted first by perceived favorability of the SPF, then by role. Darker colors indicate that certain themes are more prominent than others within the totality of attributional statements coded per participant.

Table 9 allows to make a number of general observations. First of all, the data in the table reflect that individual participants tend to make a fairly wide range of attributions. From more than half of the participants (13 out of 22, or 59%) attributions were recorded in all four pre-defined categories (test, student, class, school) over the course of the interviews. Furthermore, while all participants mention external factors to a smaller or larger extent when hypothesizing about potential causes for their school's performance, a majority of participants also *emphasizes* external factors over internal factors. Notably, two participants do not make any internal attributions at all: teachers Laura (School 05) and Quentin (School 10).

Table 9. Intensity on participant level

					Intensity								
Profile	School	Participant	Role	Perceived favorability	Category				Locus		Statement-level valence		
					Test	Student	Class	School	External	Internal	Positive	Negative	Both
A	School 01	Valerie	leader	positive	0%	13%	0%	88%	13%	88%	75%	25%	0%
A	School 03	Paula	leader	positive	18%	45%	9%	27%	91%	9%	45%	9%	45%
A	School 04	Frank	leader	positive	10%	20%	30%	40%	55%	45%	65%	25%	10%
A	School 01	Sandra	teacher	positive	29%	48%	14%	10%	76%	24%	48%	38%	14%
A	School 02	Rebecca	teacher	positive	7%	40%	33%	20%	53%	47%	60%	20%	20%
A	School 04	Natalie	teacher	positive	20%	20%	20%	40%	40%	60%	70%	0%	30%
B	School 05	Melanie	leader	mixed	75%	0%	0%	25%	75%	25%	25%	50%	25%
B	School 07	Gina	leader	mixed	0%	30%	30%	40%	60%	40%	0%	100%	0%
B	School 08	Isaac	leader	mixed	0%	8%	42%	50%	42%	58%	0%	92%	8%
C	School 09	Ken	leader	mixed	10%	30%	20%	40%	50%	50%	20%	70%	10%
C	School 10	Denise	leader	mixed	11%	33%	11%	44%	67%	33%	0%	100%	0%
B	School 06	Heidi	teacher	mixed	50%	38%	0%	13%	88%	13%	25%	63%	13%
B	School 07	Erika	teacher	mixed	0%	36%	14%	50%	50%	50%	7%	86%	7%
C	School 09	Oscar	teacher	mixed	11%	42%	32%	16%	58%	42%	26%	68%	5%
C	School 10	Quentin	teacher	mixed	25%	75%	0%	0%	100%	0%	0%	100%	0%
D	School 11	William	leader	negative	27%	9%	27%	36%	64%	36%	0%	73%	27%
D	School 12	Brenda	leader	negative	11%	11%	33%	44%	39%	61%	6%	94%	0%
D	School 13	Andrea	leader	negative	11%	22%	44%	22%	78%	22%	0%	100%	0%
B	School 05	Laura	teacher	negative	40%	60%	0%	0%	100%	0%	0%	90%	10%
D	School 11	Tony	teacher	negative	29%	43%	14%	14%	71%	29%	7%	93%	0%
D	School 12	Catherine	teacher	negative	8%	46%	23%	23%	54%	46%	0%	92%	8%
D	School 13	Xavier	teacher	negative	50%	0%	25%	25%	75%	25%	0%	100%	0%

4.2 Patterns in attribution

Work role

In Subsection 4.1, we already discussed that teachers and school leaders approach class- and school-level factors in a different manner when making attributions for SPF, in part of course because their stance on some of these issues determines the locus of causality of their attributions. We also established that both groups link and interrelate factors in a different manner, for instance, when reflecting on aspects of classroom practice and how those aspects relate to school policy or student input. The quantitized findings presented in Table 7, Table 8 and Table 9 further confirm that, overall, the attributor's work role has an observable impact on the nature and prevalence of their attributions (cf. RQ2). School leaders proportionally make more school-level attributions while teachers make more student-level attributions. This is reflected in the fact that, whereas external attributions are more prevalent overall, school leaders make more internal attributions compared to teachers.

A closer look at the individual codes underneath the four different categories (see Table 7), reveals that 91% of school leaders refer (at least once) to the school-level factor curricular line and emphasis when making attributions. A recurring observation seems to be that school leaders largely adopt a policy perspective to interpret SPF.

> "If we would have time to work on world studies, I think we would have to start by working on the vertical line - that's what the sixth grade teachers felt. They had to fill out a questionnaire at the time [of the NA] as well, like: 'what topics had been instructed not only in the sixth grade but also throughout the other grades'. And they couldn't answer that, that wasn't clear to them. So we need to start working more around the learning objectives. [...] I think a lot of it comes down to policy." (Brenda, leader, School 12)

Teachers, on the other hand, are more likely to conjure up a concrete, clear picture of the children they deal with daily, and dealt with at the time of the NA.

> [looking at the table in the report] "I am specifically looking at the children who have not reached the attainment targets and I see three individuals there. Two of them I can pinpoint but I don't really know who the third one would have been… […] I can't tell from this - but you're trying to link it to a specific child I guess. To see who would have fallen short on that." (Rebecca, teacher, School 02)

Overall, 91% of teachers refer to student *capacity*, followed nevertheless by *curricular line and emphasis* on school-level, which is mentioned by 82% of teacher participants.

Furthermore, on class-level, the group of school leaders mentions the whole variety of class-level factors we identified, while the teacher group appears to be predominantly focused on instruction and transfer. This suggests that school leaders take on a broader outlook to interpret school performance issues. Additionally, more teachers than school leaders make attributions on test-level. Perhaps this can be explained by the fact that teachers were more likely present during the administration of the NA test, and by the fact that teachers are more focused on concrete assessment in their daily practice. Finally, our qualitative dataset is too limited to allow for solid within-school observations. However, based on Table 9, we do see that with the exception of School 04 and School 07, school leaders and teachers from the same school tend to emphasize different categories when explaining (or reflecting on) their school's performance in the NA.

One final reflection bears mentioning with regard to the way attributors' work roles relate to (the locus of causality in) their attributions of SPF. In Subsection 4.1, we remarked that school- and class-level attributions are often formulated in the we-form, by both groups alike. We took this into account in order to classify such attributions as internal. So, the data suggest that there is a distinction to make not only between internal and external attribution, but also between *individual* internal and *collective* internal attribution. The suggestion that SPF is strongly associated with a collective perception of internal responsibility is also reflected in answers to follow-up 'who'-questions participants were asked (notably: Whose merit or responsibility is it that your school performed the way it did?) after the 'why'-question in the protocol (e.g., How come your school performed the way it did?). Some participants do address individual responsibility, mainly with regard to the sixth grade teacher.

> "When teachers pick up the word 'traffic', they quickly narrow it all down to traffic education." (William, leader, School 11)

However, nearly all participants emphasize on one or more occasions during the interview that they perceive school performance on the NA as the responsibility of the school team in its entirety.

> "This is the responsibility of the entire team. Ultimately, we never just carry these results with only the sixth-grade teachers." (Xavier, teacher, School 13)

Perceived favorability

The intensity data presented in Table 9 suggest that overall, work role plays a greater part in explaining differences in attribution than perceived favorability of the results.

Participants' overall perceived favorability of the SPF (as we interpreted it, see Subsection 3.3) is in line with the school profiles we pre-identified for sampling purposes (see Subsection 3.2 and Table 6). Participants from profile A, whose schools scored higher overall on the focal test, tend to predominantly give a positive appraisal, while in profile D negative perceptions prevail. In profiles B and C, the appraisal is mixed, be it with one exception: a teacher with an outspokenly negative evaluation of the school results.

At first glance, we do not observe clear patterns of attribution according to perceived favorability: regardless of their appraisal of the SPF, teachers and school leaders tend to turn to the same 'levels' when thinking about causes for their performance. Interestingly, however, it appears that none of the participants that give a positive appraisal, emphasize class-or test-level factors. Test-level factors are emphasized by a few participants – mostly teachers – with mixed or negative perceptions of the feedback's favorability. Class-level factors are emphasized over other factors by one school leader who perceives the school's SPF as unfavorable.

> "Well, then we actually have to examine whether our teachers are really working towards achieving the attainment targets. Because you can work in a class – and I come from the sixth grade myself – and be completely out of touch. But if you involve children in the topics that are hot topics nowadays, and go outside with children to really experience traffic education... Then I do think that those attainment targets should definitely be reached. So I'm definitely going to have a conversation with the teacher about that. Scoring so poorly!" (Andrea, leader, School 13)

A closer look at the individual codes underneath the four different categories, gives more indication as to where perceived favorability of the SPF plays a role. *Curricular line and emphasis* and student *capacity* are discussed by all participants that give a positive appraisal (100%) and by the majority of participants that give a mixed appraisal (80% and 70% respectively). However, while *curricular line and emphasis* is mentioned by 86% of participants that negatively appraise their school's results, an equal proportion of these participants also discuss (problematic) test *item formulation*. This suggests that feelings that the NA test items were not clear enough, are stronger in this latter group of participants. Furthermore, the average number of coded utterances, i.e., recorded attributional statements, is highest in the participant group that positively appraise their SPF. The diversity of attributions, i.e., the number of different thematic codes assigned to attributional statements, is on average the lowest in the participant group from school profile B.

During analysis, we found that that regardless of school profile and overall perceived favorability, all participants utter attributions for positive as well negative aspects of performance, or to try and interpret why some aspects of the results defy their prior expectations. More than half of all coded utterances (65%) are attributions in which participants reflect on negative performance or negative aspects of performance (e.g., the number of pupils not reaching the AT) even when overall perceived favorability was positive. Therefore, we also included this information as 'statement-level valence' in Table 9.

Factors from all four categories are used to explain both positive and negative aspects of performance. However, test-level attributions are uttered predominantly to express concern or to address perceived disconnects when explaining negative (aspects of) the school results, also by participants who overall perceive the feedback as favorable. Additionally, participants who regard the feedback as unfavorable, use student-level attributions exclusively to address negative results. A number of participants (in all school profiles) also reflect on student-level factors for why scores defy their prior expectations for worse or for better. Finally, school-level attributions are most salient in school profile A (high performance overall) when participants account for positive aspects of performance.

When we look at all external attributions in the dataset, we see that 71% of them pertain to negative aspects of performance. Internal attributions, however, pertain for 38% to positive aspects, for 55% to negative aspects, and for a remaining 7% to nuance and reflect on why a result could go or be interpreted either way.

5 Conclusions and discussion

5.1 Interpretation

Our research aim in this study was to explore how school performance feedback (SPF) from an external assessment gets woven into the causal narratives of schools and of individual educators. We collected and interpreted interview data from teachers and school leaders whose schools had participated in a low-stakes national assessment (NA) and had received criterion- and norm-referenced feedback on their performance. We asked them whether they evaluated the results obtained as positive or negative, and asked them to reflect on the causes they saw for performance. In this exercise, we were particularly interested to learn where and with whom they placed the locus of causality. Who or what contributed to the fact that the school scored the way it did? An underlying question being: to where do one’s own responsibilities and realms of impact extend?

In line with prior research, we find that school performance is attributed to external factors to a great extent (cf. RQ1). Arguments include: you have a certain student population to work with, the assessment itself is not flawless, and also, your location and the effort of parents determine whether or not the subject matter in question can be sufficiently ‘experienced’. However, teachers and school leaders do make internal attributions as well, pertaining to instruction and, here particularly, school policy. In internal attributions, participants overwhelmingly adopt a collective stance: *we* made this happen, rather than *I* made this happen. Student achievement and school performance seem to be experienced as very much a collective, rather than a personal endeavor.

Overall, participants tend to turn to a wide variety of internal and external factors when reflecting on causes for school performance. From a data-based decision making perspective, this is encouraging, because it attests to the fact that educational professionals rightfully acknowledge that learning outcomes are the product of many different building blocks. External factors have a place here as well: it is important to be conscious of the context and the abilities of the students you teach, of the instrument that is used to assess achievement, of the contribution of your peers, employees, team mates. On the other hand, however, this finding also exacerbates the cautious suggestion that educational professionals may not readily assume personal responsibility for student outcomes. After all, if everything matters, and everyone plays a part, and we are all in this together – my own role as a cogwheel in the system becomes smaller and smaller. This may be discouraging, but, to a certain extent, it is a common human reflex. Ego-defensive biases seep into attribution, sensemaking and decision making in an attempt to preserve one's values and beliefs, and to shield and maintain one's identity, self-worth and self-integrity (Coburn & Turner, 2011; Goffin et al., 2022; Lasater et al., 2021; Sherman & Cohen, 2002).

Evaluative interpretations aside, the observation that attributions fan out so widely does illustrate that it is not easy or straightforward for users to formulate an unambiguous analysis and an actionable diagnosis based on SPF – as data providers and data use researchers would perhaps hope or assume. Additionally, our data show that attributions sometimes overlap and interlink. For instance, some of school leaders' class-level attributions have touching points with aspects of school policy, some of teachers' class-level attributions touch upon student-level factors, factors pertaining to the test are related to instructional practice and characteristics of the student population, and student population is framed by some attributors as a characteristic of the school. These observations demonstrate that, although we disentangled different levels, categories and codes, real-life attributions are not always nicely siloed and clear cut. This corresponds to theoretical and empirical notes establishing that attributors typically invoke multiple mental models at once, which can help to understand why it is often difficult to arrive at one definitive explanation (Bertrand & Marsh, 2015; Spillane & Miele, 2007).

With regard to work role (cf. RQ2) we find that teachers, generally speaking, tend to emphasize student-level factors, while school leaders focus most on the school-level and address a wider range of potential causes pertaining to the class-level. Test-level attributions, on the other hand, appear to be more top-of-mind with teachers. These findings fit with a sensemaking perspective: users make meaning of the school performance data they were presented with from their own frame of reference . Since we also find clear differences between teachers and school leaders within schools, this attests to the necessity of collective sensemaking. A lot of actionable information can be gained by processing SPF within a team in order to broaden the interpretive frame.

In our research design, we took care to stimulate variability in perceived favorability (cf. RQ3) by recruiting participants from schools in different scoring profiles. The fact that perceived favorability (as we summarized and interpreted it from the raw participant-level data) largely corresponds to these profiles, suggests that this indeed was a good starting point. We find that concerns about the (external) assessment are far more prevalent when negative (aspects of) performance is/are addressed (cf. Nabors Oláh et al., 2010; Verhaeghe et al., 2010). Furthermore, we clearly see that all participants, regardless of overall perceived favorability of the SPF, undertake more intense causal searches with regard to negative aspects of performance. This is in line with assumptions in attribution theory (Weiner, 2000, 2010) and with empirical findings (Van Gasse & Mol, 2021; Verhaeghe et al., 2010). It suggests that it is perhaps easier and more natural to freely hypothesize about potential causes for what is going wrong, than to pinpoint what is already going well and what needs to be sustained.

5.2 Practical implications

Literature on attribution of student achievement varies in terms of terminology (*attribution, diagnosis, causal explanations*,…), focus (formal or informal data use) and methodology (e.g., interviews about data use versus observing data use in situ). Studies also vary in terms of the types of data discussed (e.g., external assessment data, school-internal data) and the goals of data use (e.g., instructional decision making, diagnostic student testing, student guidance, policy reflection). It is likely that differing patterns of attribution need to be interpreted in light of this variability (Bertrand & Marsh, 2015; Evans et al., 2019). For the present study, we made use of authentic SPF from a typical external assessment, with enough benchmark data to allow for standardization and comparison, but devoid of an 'accountability filter' as found in many other educational contexts. We interviewed educational professionals in different roles and from schools dispersed over a scoring continuum, in order to delve into the meaning they make of school results on an atypical subject matter.

Based on our findings, we can propose a number of hands-on avenues to consider for SPF providers who want to help SPF users make sense of school performance data in a productive and sensible manner. A number of these recommendations are tantamount to offering users sufficient clues and cues to aid them in 'filling in the blanks'. First, for instance, without being tempted to offer an 'objective' evaluation, it can still be meaningful to offer users (more or more outspoken) guidance in their appraisal of the favorability of school results. How to unravel the pluses and the minuses? What can be the benchmark *for us* for these pluses and minuses? And why exactly is it important to look at things that seem to be going well, as well?

Second, the different variables and levels that can be distinguished in users' (spontaneous) attributions, can be made explicit in order to help users to unpack what needs fostering and what needs adjusting. An interpretive guide could include designated sections for school-, class-, student- and test-level factors to be reflected on. In cases such as the Flemish NA, in which the impact of some of these variables is analyzed on system-level, research reports can also inspire causal search. That being said, providers should not necessarily aspire to provide checklists for individual schools where blanks are already filled in, as at the end of the day, results are only meaningful when they are interpreted in light of a school's own goals (Schildkamp, 2019). As discussed in the previous subsection (5.1), this interpretation needs the collective outlook of a school team in order to really be rich and informative. Broadening the frame is essential, although we know that, in practice, this is not yet self-evident (Gutwirth et al., 2021).

Third, it may be helpful to have teachers and school leaders reflect explicitly on locus of causality, and also on the other attributional properties (controllability and stability) of the causes they hypothesize about. If the causal locus is external, maybe you or someone else still has some sort of control over what is occurring? Questioning one's own role in addressing detected difficulties is perhaps more an issue of controllability than of locus of causality, and it attests to the depth of inquiry actually needed for data-based decision making (Schildkamp et al., 2016; Van Gasse & Mol, 2021). What do we need to do to help students overcome their difficulties? And ultimately, if some of the hypothesized causes are external and uncontrollable, is it possible to dig deeper, looking for other contributions or adjustments you can make, instead of being tempted into deficit-thinking?

5.3 Academic strengths, limitations and follow-up

With this study, we aimed to explore trends and hope to inspire further research as well as practical developments. However, a number of limitations need addressing. First and foremost, it is important to acknowledge that the interviews we conducted, did not constitute an authentic sensemaking setting. We elicited rather than observed attributions: we *made* our participants make sense. Although we coded the data generously, taking into account spontaneous reflections uttered throughout the interviews, we cannot be certain that participants would hypothesize about the same wide range of factors in daily practice. Additionally, the interviewer was known by participants to be affiliated to the policy research center responsible for the NA, which may have influence the nature and number of test-level attributions they did or did not make. Micro-process studies could overcome these limitations (Little, 2012; Schildkamp, 2019), although this approach lacks some of the informative value our inquiry could provide.

Future research could also build further on other foundations laid out in the present study. For example, we did originally try to disentangle attributions for criterion- and norm-referenced aspects of the SPF, but during the interviews it became apparent that, overall, the boundaries were blurry. This was confirmed when coding and analyzing the data, so we decided not to pursue this path. An approach in which both dimensions are more strictly distinguished, could add depth to perceived favorability as a variable. Additionally, while we succeeded in isolating attributions and examining them in depth, it is also valuable to examine attributions in relation to other data use phases and other aspects of sensemaking, as prior studies have already demonstrated to a certain extent: (How exactly) do attributions relate to the goals that schools and individual educators have set for themselves? (How) do attributions feed into decisions and actions? And also: how do we need to interpret data users' attributions in light of the (validity of) their understanding of the data? As we were interested in participants' authentic causal perceptions, we did not take into account accuracy of interpretation or potential misconceptions for this study.

A contribution to the knowledge base that we made in this study, is that we described patterns in attribution according to users' work roles. There are other characteristics and antecedents worth exploring as well, in order to examine differences *within* user groups. Individual users' data literacy is one example. Difficulties with initial analysis and interpretation of SPF have been reported to complicate and even thwart the diagnostic phase, which in turn impacts the (non)formulation of a response (Verhaeghe et al., 2010). Furthermore, the nature of teachers' attributions has been found to be influenced by personal characteristics such as training level or, interestingly, experience in the profession. Some studies find that early-career teachers have more faith in the role and responsibility of teachers in students' learning processes, while seasoned teachers emphasize external factors to a greater extent in their attributions (Wang & Hall, 2018). It would be interesting to see whether the same holds for school leaders.

On a more collective level, attributional antecedents worth exploring are group dynamics in collective sensemaking and contextual factors. Attribution is shaped by personal lenses, but in collective sensemaking of student achievement data, explanations need to be deliberated and negotiated (Spillane, 2012). The nature of these group interactions can influence attribution (Bertrand & Marsh, 2015). For instance, self-preservation reflex can make it hard to address internal causes for (problematic) student outcomes and to address responsibility issues, and particularly delicate in formal discussions with colleagues and superiors (Van Gasse & Mol, 2021). Trust and a professional environment that is perceived as safe, are important prerequisites as people are not always willing to show their vulnerability (Gutwirth et al., 2021; Van Gasse et al., 2021; Van Gasse & Mol, 2021). Furthermore, cultural differences in attribution have been reported (Matteucci & Gosling, 2004; Wang & Hall, 2018) which attests to the fact that the educational system (and its goals, norms, policies and customs) influences educational professionals' mental models and their sensemaking of achievement data (Goffin et al., 2022; Mandinach & Schildkamp, 2021a). Similarly, school culture and organization play a big part in educational professionals' attributions, as the purposes of data use ("what do we want to learn here?") and the conditions under which (collective) sensemaking can take place (who is involved) are often set out on policy level (Bertrand & Marsh, 2015). Lasater et al. (2021) discuss how data use policies and practices on system level and on school level (high-stakes accountability measures, unsafe work environments and an unhealthy focus on identifying deficiencies) can induce and exacerbate deficit thinking in teachers and school leaders: a sharp emphasis on attributing student failure to student-internal factors, and a relinquishment of personal responsibility, which ultimately creates and sustains inequity.

Finally, we zoomed in on the locus of causality of teachers' and school leaders' attributions of SPF partly because we were interested in learning to what extent school performance is experienced as a personal or collective accomplishment. Are (good) results indeed feathers in educators' caps? In order to do this, we approached educational professionals' causal ascriptions in SPF use as instances of interpersonal attribution, considering educational professionals the observers, and students the actors (see Subsection 2.1). While it is beyond the scope of our interests in the present study, it would be worthwhile to reflect on whether this is accurate, or rather: complete. Particularly in the case of SPF, a question that needs further exploration is to what extent school teams consider the outcomes as personal (individual or collective) successes or failures ("*our school* should be proud of this achievement", "*my class* scored the highest", "*the students* did well on this assessment"). Discourse analysis could provide more insight into this matter.

Study 4

The interplay of user beliefs and situated characteristics in explaining school performance feedback use

BASED ON Goffin, E., Janssen, R., & Vanhoof, J. (2022). The interplay of user beliefs and situated characteristics in explaining school performance feedback use. *School Effectiveness and School Improvement, 33*(3), 456–478. https://doi.org/10.1080/09243453.2022.2041048

ABSTRACT The present study explores predictors of school performance feedback (SPF) use. In total, 470 Flemish educational professionals were surveyed about their use of SPF from school-external, low-stakes standardized assessments. A path analysis was conducted in order to investigate how individual user beliefs impact SPF use on school level and how those beliefs mediate the effects of school-level features pertaining to school organization, performance and voluntariness. Findings include that users' cognitive attitude and perceived expectations of others have a small effect on engagement with SPF in schools, and that these predictors mediate the effects of certain organizational characteristics. Whereas performance levels do not impact school-level feedback use, voluntariness in feedback pursuit and particularly an SPF-oriented school culture emerge as drivers. Implications for practice include the need for stimulating ownership in data-based decision making. Suggestions for further research are also discussed.

1 Introduction

The past decades have been marked by an increasing awareness of the importance of data use in education. Analogous to evidence-based approaches in medicine (Schildkamp, 2019), researchers find that educational professionals endeavoring to improve student achievement need to fully exploit all information sources available to them in order to shape their policy and practice. However, the literature on data-driven decision making (DDDM), or data-based decision making (DBDM), has also established that data themselves do not necessarily drive (Dowd, 2005; Lockton et al., 2019). In order to foster informed school improvement, for instance through interventions, it is not sufficient to make high quality data available (Hulpia & Valcke, 2004; Schildkamp & Kuiper, 2010). It is also crucial to be conscious of the factors that trigger, accommodate or inhibit efficient data use in schools.

Research has identified a wide range of such influencing factors. For one, data use requires human capacity (Mandinach & Gummer, 2013; Mandinach & Schildkamp, 2021b). A fundamental prerequisite to DDDM is that educators are sufficiently data literate. Data literate educators possess the knowledge, skills and dispositions that enable them to transform information into actionable knowledge (Mandinach & Gummer, 2016). Instead of solely relying on intuition, they confidently and critically approach a wide range of information sources, interpret and contextualize this information, and use it to shape their policy and practice in a responsible and appropriate manner (Mandinach & Gummer, 2016; Vanlommel et al., 2017). They are also willing and able to engage in collective sensemaking, as collaboration and co-construction are key in effective DDDM (Mandinach et al., 2011; Mandinach & Gummer, 2016).

Since data use processes always take place within a certain setting and structure, situational characteristics invariably influence how individual data users engage with data (Abrams et al., 2021). For instance, like other organizational processes, the data use process in schools is influenced by the school context (e.g., staff, expertise, professional capacity and resources) and the school organization (e.g., leadership, innovation climate, collaboration) (Abrams et al., 2021; Bryk, 2010; Jimerson et al., 2021; Visscher, 2021).

Moreover, the educational context determines to a great extent how data use processes take shape (Coburn & Turner, 2011; Mandinach & Schildkamp, 2021a). An educational system tends to be characterized by its inclination towards accountability or improvement, two purposes of data use between which there is a duality and often a tension (Datnow & Park, 2018; Schildkamp et al., 2014; Visscher & Coe, 2003). Several authors take a passionate stance against accountability-driven systems, which they propose corrupt the processes they intend to monitor (Nichols & Berliner, 2007), weaken schools because of the pressure they puts on educators (Nichols & Harris, 2016), and do not (or at least not conclusively) enhance student achievement (Nichols et al., 2006, 2012; Nichols & Berliner, 2007). Such systems often rely on high-stakes testing: standardized forms of assessment that serve school accountability or student accountability goals (or both), and typically only cover a limited range of topics, which in turn also raises equity concerns (Datnow & Park, 2018). “High stakes testing” and “standardized testing” are frequently used as synonyms, also in many research accounts on DDDM. The issue lies, however, not entirely with the standardized nature of these tests and assessments, but rather in the stakes, or consequences attached to the outcomes (Nichols & Harris, 2016). Low-stakes external standardized assessments can provide valuable information for school improvement and can meaningfully contribute to a picture of student achievement, when they are part of a balanced system of testing, assessment and process monitoring (Nichols & Berliner, 2007) and when aimed at identifying areas for support and improvement (Datnow & Park, 2018; Nichols & Harris, 2016). In several educational contexts, insights like these are bringing about a gradual shift in focus from data use for accountability to data use for (continuous) improvement and organizational development (Mandinach, 2012; Mandinach & Schildkamp, 2021a).

Nevertheless, data used in DDDM should not be limited to assessment data and test scores (Mandinach & Schildkamp, 2021a). And above all, in data use, alignment of the data with the goals is paramount (Mandinach, 2012; Mandinach & Gummer, 2016). Using data inappropriately or for unintended purposes, raises issues of validity and may lead to poor, unfit or undesirable decisions (Mandinach, 2012; Mandinach & Gummer, 2016; Visscher & Coe, 2003).

Focus of this study

A growing number of descriptive studies is providing in-depth insight into the mechanics of how the aforementioned factors influence data use. However, the field is in need of more explanatory research (Van Gasse et al., 2017) to shed light on the relative impact of influencing factors (Schildkamp et al., 2017) and their interplay (Coburn & Turner, 2011, 2012). In the present study, we address this knowledge gap by investigating the use of school performance feedback (SPF) from school-external, low-stakes standardized assessments as a case. SPF is conceptualized as data about a school’s functioning or performance, provided confidentially to the school by an external agent for self-evaluation, intended to inform the school’s decision making process (Visscher & Coe, 2003). This definition entails a clear school development orientation.

We investigate factors that enable or hinder SPF use in schools by adopting a quantitative approach, and by taking on a dual perspective. We include user-level predictors in order to acknowledge that SPF use, like other forms of data use, takes shape in the hands of individual actors (Coburn & Turner, 2012; Datnow & Hubbard, 2016; Prenger & Schildkamp, 2018; Schildkamp et al., 2014). By also including school-level predictors, we account for the fact that these data users do not operate in isolation (Abrams et al., 2021; Coburn & Talbert, 2006; Coburn & Turner, 2011, 2012; Schildkamp, 2019; Schildkamp et al., 2014).

In order to select user-level predictors of SPF use and in order to hypothesize how these interact with school-level predictors, we take inspiration from the Theory of Planned Behavior (TPB; Ajzen, 1991). The TPB states that the intention to perform a certain behavior is shaped by the strength and favorability of the agent's behavioral, normative and control beliefs (Ajzen, 1991). Operationally, these beliefs form three distinct constructs: attitude, subjective norm and perceived control. Consequently, we hypothesize that educational professionals' attitude, subjective norm and perceived control regarding SPF influence engagement with SPF reports in schools. The present study focuses on users' self-efficacy when investigating perceived control.

The TPB acknowledges that the relative impact of its central predictors varies across different settings (Ajzen, 1991, 2002, 2011; Armitage & Conner, 2001), which justifies situating SPF use and individual user beliefs within in a specific school context. In the present study, we hypothesize that organizational, performance-related and contextual school-level features affect school-level SPF use (cf. Verhaeghe et al., 2010; Visscher & Coe, 2003) because they influence the beliefs of individual SPF users. More specifically, we hypothesize that users' beliefs about SPF will be more favorable or salient when they perceive their school culture to be accommodating of SPF use. Due to the nature of SPF, we also presume users in a coordinating role regard SPF more favorably than those who teach. Furthermore, we hypothesize that a higher performance level of the school has a positive impact on user beliefs regarding the SPF, and consequently, will positively relate to school-level SPF use. Finally, we propose that SPF that was actively and voluntarily requested, kindles more positive user perceptions and boosts engagement to a larger extent.

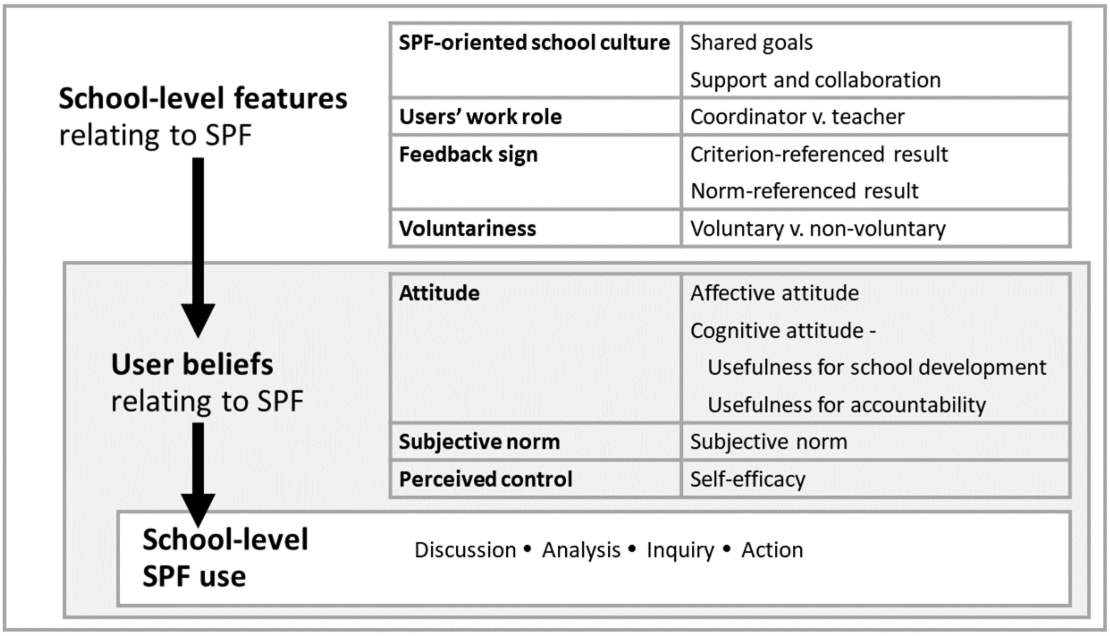
These hypotheses give rise to the following research questions:

- (RQ1) To what extent do users' attitudes, subjective norm and self-efficacy impact SPF use on school level?
- (RQ2a) To what extent do SPF-oriented school culture, users' work role, feedback sign and voluntariness in feedback pursuit impact SPF use on school level, and
- (RQ2b) Are those effects mediated by users' attitudes, subjective norm and self-efficacy?

2 Theoretical framework

In the following paragraphs, we will briefly present theoretical insights pertaining to each of the predictors included in our conceptual model, as pictured in Figure 4, and discuss their operationalization in this study.

Figure 4. Conceptual model



SPF-oriented school culture	Shared goals Support and collaboration
Users' work role	Coordinator v. teacher
Feedback sign	Criterion-referenced result Norm-referenced result
Voluntariness	Voluntary v. non-voluntary

Attitude	Affective attitude Cognitive attitude - Usefulness for school development Usefulness for accountability
Subjective norm	Subjective norm
Perceived control	Self-efficacy

Note. SPF = school performance feedback.

2.1 School-level SPF use

Since implementation is key in effective SPF use (Hellrung & Hartig, 2013; Visscher & Coe, 2003), we investigate school-level SPF use in a tangible, concrete manner. We focus on the policy-making cycle that corresponds to effective use of SPF data for school improvement (Verhaeghe et al., 2010) and on the systematic process of transforming data into actionable knowledge which is central to different data use theories of action (Mandinach, 2012; Marsh, 2012). First of all, SPF reports need to be discussed and analyzed within the school team in order to turn raw data into information. Next, this information needs to be triangulated with prior knowledge and other sources to turn it into knowledge. Finally, this process ideally leads to a decision or action (Visscher & Coe, 2003), which, in the larger policy cycle, is then evaluated. In line with this take on data use, we conceptualize school-level SPF use as SPF having been discussed, thoroughly analyzed, used as an impulse for further inquiry and a basis for formulating actions.

2.2 User beliefs relating to SPF

2.2.1 Attitude

Attitudes are formed by an agent's behavioral beliefs: their judgements of prospective outcomes of the behavior (Ajzen, 1991, 2002). Attitudes have an affective and a cognitive dimension, as beliefs can be based on emotions elicited by the behavior or on an evaluation of its attributes (Prenger & Schildkamp, 2018; Sanbonmatsu & Fazio, 1990). Applied to data use, affective attitude may refer to users feeling comfortable with, excited about, or apprehensive of using data (Jimerson, 2014; Vanhoof et al., 2014) and it has been found that data use anxiety contributes to DDDM resistance in educators (Dunn, Airola, & Garrison, 2013). Cognitive attitude, which is influenced by educators' general views on DDDM (Dunn et al., 2019; Jimerson, 2014), is concerned with "buy-in" (Schildkamp & Kuiper, 2010) or the extent to which users regard data as useful (van der Kleij & Eggen, 2013; Vanhoof et al., 2014). In this study, we will relate perceived usefulness of the SPF to potential outcomes of SPF use. Usefulness for school development refers to acknowledging the potential of SPF for formulating concrete actions or decisions (instrumental use) and for inspiring the decision making process (conceptual use), whereas usefulness for accountability pertains to regarding SPF as a tool for supporting prior decisions (symbolic use) and for self-promotion and legitimatization (strategic use) (Hellrung & Hartig, 2013; Rossi et al., 2004; Visscher & Coe, 2003).

2.2.2 Subjective norm

Subjective norm is produced by an agent's normative beliefs: the way they feel others expect them to engage in the behavior (Ajzen, 1991, 2002). Normative beliefs do not necessarily reflect perceived coercion, but can also refer to perceived encouragement. Data use expectations can emanate from actors within the school, e.g., a school leader advocating the use of a certain instrument, or from external parties, e.g., in systems where data use is associated with compliance (Uiterwijk-Luijk et al., 2017; Vanhoof et al., 2014). School-external expectations can also prompt school leaders to formulate expectations towards teachers (Abrams et al., 2021). Some scholars propose that social pressure is detrimental to data use, because it compromises educators' autonomous motivation to use data (Vanlommel et al., 2016) or their sense of ownership (Schildkamp & Teddlie, 2008). These concerns are raised in particular in contexts with an emphasis on school or student accountability, which typically involve high-stakes standardized testing (Nichols & Harris, 2016). Nevertheless, subjective norm has also been found to positively affect data use in certain circumstances. Social pressure to work in an inquiry-based manner is positively related to teachers displaying an inquiry habit of mind (Uiterwijk-Luijk et al., 2017) and both development-oriented and accountability-based external expectations have been found to motivate principals to make use of data (Vanhoof et al., 2014).

2.2.3 Self-efficacy

Perceived control is rooted in an agent's control beliefs: their perception of factors that help or inhibit them in engaging in the behavior (Ajzen, 1991, 2002). We will operationalize perceived control as self-efficacy. In the context of data use, an educational professional's self-efficacy expresses the extent to which they feel capable of engaging in data use because they possess the necessary competences to do so (Bandura, 1997; Van Gasse et al., 2017). In other words, self-efficacy expresses confidence over one's data literacy, i.e., one's knowledge and skills for processing data and formulating responses accordingly (Mandinach & Gummer, 2016). The literature paints a rather pessimistic picture with regard to educators' data literacy (e.g., van der Kleij & Eggen, 2013; Vanhoof et al., 2011) and finds that many still profess to feeling insecure in this respect (Datnow & Hubbard, 2016; Earl & Fullan, 2003). Nevertheless, a sense of self-efficacy is an important determinant of data use (Dunn, Airola, Lo, et al., 2013a). In this study, we conceptualize self-efficacy as SPF users feeling they are able to understand SPF, interpret it, and translate it into concrete actions. We thus acknowledge that data literacy surpasses mere statistical literacy, but also entails the transformation of information into actionable knowledge (Mandinach & Gummer, 2016).

2.3 School-level features relating to SPF

2.3.1 SPF-oriented school culture

A strong data culture in schools is a DDDM enabler and fosters data literate educators. Strong data cultures are grounded in a clear vision and common goals, and collaborative structures (Bryk, 2010; Hamilton et al., 2009; Jimerson, 2014; Jimerson et al., 2021; Jimerson & Wayman, 2015). Leadership is key in shaping and facilitating these data cultures (Bryk, 2010; Hamilton et al., 2009; Jimerson, 2014; Jimerson et al., 2021) as is a mindset of continuous improvement (S. Sutherland, 2004). In order to assess the prevailing school culture regarding SPF, we will first focus on users' perception of a shared goal orientation within the team. This refers to sharing a common vision and understanding about SPF use, including collective norms and objectives (Hoogland et al., 2016; Jimerson, 2014; Mandinach, 2012; Schildkamp et al., 2019). In essence, it refers to the collective frame of reference on how and why to use SPF (Schildkamp et al., 2014). Additionally, we will gauge users' experience of internal support and collaboration regarding SPF use. Support is associated with networking, brokerage and coaching (Jimerson, 2014; Schildkamp et al., 2019). Both support and collaboration foster educators' actual data use competences as well as their confidence in using data (Abrams et al., 2021; Datnow & Hubbard, 2016; Schildkamp & Kuiper, 2010; Van Gasse et al., 2017; Vanhoof et al., 2011).

2.3.2 Users' work role

Individual school team members have different data use competences (van der Kleij & Eggen, 2013), needs (Coburn & Talbert, 2006) and objectives (Schildkamp & Kuiper, 2010). Formal work roles play a part in these differences, as the nature of DDDM also differs according to an educator's position (Mandinach et al., 2011). For teachers, data use and data literacy are oriented predominantly towards instructional decision making rather than school development. Whereas school leaders are often the directors of how data use takes shape within their team by functioning as culture builders and modeling good practices (Jimerson, 2014; Schildkamp et al., 2014; Schildkamp & Teddlie, 2008; Vanhoof et al., 2012), teachers tend to need encouragement to use data (Uiterwijk-Luijk et al., 2017; Vanhoof et al., 2012) and hold the school leader responsible for setting up a data policy (Hoogland et al., 2016). Considering that school leaders are often former teachers, and teachers in turn learn from school leaders, there is a certain amount of reciprocity among these educator roles in DDDM (Jimerson, 2014).

2.3.3 Feedback sign

Positive feedback has a positive impact on feedback acceptance (Ilgen et al., 1979) while negative feedback can negatively impact recipients' perceptions (Lechermeier & Fassnacht, 2018). Data use research has indeed found that teachers are reluctant to engage with data that challenge their efficacy (Coburn & Turner, 2011; Dunn, Airola, Lo, et al., 2013a; Lockton et al., 2019). Although we know that a school's performance in an external assessment influences the way the resulting SPF is received (Verhaeghe et al., 2010; Visscher & Coe, 2003), few studies have specifically zoomed in on how the "sign" of SPF relates to usage, as we intend to do here. Corresponding to the most prevalent frames of reference employed in external standardized assessments (American Educational Research Association et al., 2014; Hellrung & Hartig, 2013), we will consider both criterion- and norm-referenced SPF results. Criterion-referenced measures are absolute and compare achievement to a standard. Norm-referenced measures are relative and compare achievement to that of a reference group.

2.3.4 Voluntariness

SPF systems are inherently focused on school improvement, since they are a self-evaluation tool, but in practice they can also serve accountability goals (Visscher & Coe, 2003). While a certain degree of accountability pressure stimulates engagement with SPF (Vanhoof et al., 2012), the self-evaluation purpose of a system needs to be explicit in order to foster engagement with the data for school improvement (Maier, 2010). Thus, a careful balance needs to be struck. As voluntariness corresponds to the degree of "free will" in adopting a certain system (Wu & Lederer, 2009), we will account for voluntariness in feedback pursuit by taking into account whether or not a school takes purposeful action to acquire SPF. In general, SPF systems tend to be successful when they address a perceived information deficiency from the recipients themselves (Hendriks et al., 2002), when they are adopted rather than imposed, and when the SPF recipients feel they have sufficient ownership over the implementation (Visscher & Coe, 2003).

3 Method

We developed an online survey that was completed by 470 Flemish educational professionals whose schools had recently been presented with an SPF report. We performed a path analysis on the survey data.

3.1 Research context

The study took place in Flanders, the Dutch-speaking part of Belgium. The Flemish educational system is largely decentralized (Organisation for Economic Co-operation and Development, 2017). Government-issued attainment targets describe minimum goals for different stages in primary and secondary education, but schools enjoy great autonomy. The inspectorate monitors whether schools comply with regulations and pay sufficient attention to internal quality, but to date, Flanders does not organize central examinations.

On system level, achievement of attainment targets is periodically measured with large-scale national assessments (NA) that cover a wide range of topics. NA take a snapshot of the performance of a population at a certain point in time by testing representative samples of students. Schools cannot volunteer for participation, nor can participation be enforced onto them, but sample schools do receive a personalized and strictly confidential SPF report. Similar SPF reports can be requested by administering parallel tests (PT) free of charge. PT are parallel versions of the tests administered in the NA, released after the national averages have been made public. NA and PT are highly standardized tests in terms of content, administration and scoring (American Educational Research Association et al., 2014), but they are low-stakes for schools and pupils. Results are not reported to the educational government nor are they made public.

The analyses used in the Flemish NA are grounded in item response theory (IRT). The standard that corresponds to achieving the attainment targets is determined by a panel of educational professionals and experts in a process based on the Bookmark procedure (Mitzel et al., 2001). The SPF reports describe the extent to which the attainment targets were reached within the school (criterion-referenced feedback) as well as the schools' performance relative to the national average (norm-referenced feedback). Note that in the PT reports, the representative NA sample constitutes the reference group. The SPF also contains value added information that corrects performance for input characteristics as a measure for “fair comparison” (Visscher & Coe, 2003).

3.2 Data collection

3.2.1 Participants

An online survey was sent out to 427 schools which had participated in an NA of French in Grade 6 or Technology in Grade 8 (148 schools), or had voluntarily taken PT on various subjects (279 schools). These schools had received SPF about five months prior to the administration of the survey. Because we aimed to illuminate SPF use from multiple perspectives, we asked for responses from the school leadership and from (the) teacher(s) involved.

On school level, a response rate of 72% was achieved. In total, 470 online surveys were completed in full by educational professionals in both primary (60%) and secondary education (40%). Overall, respondents' average age was 46, and the number of females surpasses the number of males with 69% to 31%. Some 22% percent of respondents hold a master's degree; for the majority this is a bachelor's degree (76%).

As a result of convenience sampling there is some nesting of participants within schools, but this nesting is very limited. A majority of respondents were single observations within their school, as we also need to take into account that PT respondents from one and the same school did not all focus on the same test subject report to discuss SPF use. Consequently, in the analyses, respondents were treated as not nested within schools.

3.2.2 Instrument

The online survey assessed users' perceptions regarding the (type of) SPF that their schools had been presented with, in the case of NA participation, or had actively collected, by taking PT. One scale measured the extent to which the report had actually been put to use in the school, 'SPF use' for short. Five scales measured user beliefs: 'Affective attitude', 'Cognitive attitude: Usefulness for school development', 'Cognitive attitude: Usefulness for accountability', 'Subjective norm' and 'Self-efficacy'. Finally, two scales addressed the perceived presence of 'Shared goals' and 'Support and collaboration' with regard to SPF use.

Items were selected from other studies on data use and inspired by literature on the construction of TPB-based questionnaires (Ajzen, 2002; Francis et al., 2004; Pierce et al., 2013; Van Gasse et al., 2015; Vanhoof et al., 2014; Wayman et al., 2017). They were adapted to particularly tap into perceptions about SPF from NA and PT. All items were statements to be scored on a 5-point Likert scale (1 – entirely disagree, 2 – disagree, 3 – neither agree nor disagree, 4 – agree, 5 – entirely agree) with a possibility to opt out (I don't know/This statement does not apply). In order to establish face validity, the items were submitted to peer review.

The construct validity of each scale was examined with a confirmatory factor analysis (CFA; T. A. Brown, 2006). This approach allowed us to take covariance between items into account and to optimize each model based on modification indices. Moreover, it provided an opportunity to handle missing data in an advanced manner by employing 'full information maximum likelihood' (FIML) as an estimator. The analyses were conducted in R 3.5.1 'Feather Spray' with the lavaan-package (Rosseel, 2012). In order to remodel and ultimately assess the validity of each scale, we considered the comparative fit indices (CFI), the Tucker-Lewis indices (TLI), the root mean square errors of approximation (RMSEA) and the standardized root mean square residuals (SRMSR). For the CFI and TLI, a cutoff of .95 was exceeded in all models, and good fit was confirmed by the RMSEA and SRMSR values which were all smaller than or close to the cutoff of .08 (Schreiber et al., 2006). Reviewing the factor loadings of each item on the corresponding latent concept, we found that the loading of one item pertaining to 'Subjective norm' did not meet a cutoff set at .400. Based on content validity considerations, we decided to retain the item in the scale. As shown in Table 10, all scales show adequate reliability (Nunnally & Bernstein, 1994) with Cronbach's alpha values ranging from .67 to .87.

3.3 Data analysis

3.3.1 Measures

For all scales, i.e., the dependent variable as well as user beliefs and characteristics of school culture, we moved forward with the factor scores predicted by the CFA. Respondents' work roles were coded into a dummy variable. Respondents exercising a predominantly coordinating function (52%) were assigned value 1, while those who mainly teach (48%) constituted the reference group.

The feedback sign of the SPF was retrieved from the (focal) SPF reports. We aimed to express the school's result in straightforward, continuous variables. In order to capture the school's criterion-referenced result, we calculated the average percentage of pupils that reached the attainment targets over all tests in the (focal) report. These average percentages were then standardized into Z-scores. For the norm-referenced result, we compared the proportion of pupils that reach the attainment targets within the school to the proportion that had done so in the full NA sample. We made this comparison for all tests in the (focal) report. The average differences were, again, standardized into Z-scores.

In order to assess the effect of voluntariness, we took into account whether a response pertained to a PT school, where SPF was acquired voluntarily through active participation, or to an NA school, where the SPF resulted from more passive test participation. This information was coded into a dummy variable. The full sample of 470 complete survey responses consisted of 330 responses from schools that had taken PT (70%), assigned value 1, and 140 responses from schools that had participated in an NA (30%), which constituted the reference group.

3.3.2 Path analysis

A path analysis was conducted with the R-package lavaan (Rosseel, 2012). In correspondence with the conceptual model, we started out with a model in which user beliefs mediate the effect on SPF use of school-level features. We assumed no covariance between variables. Based on the modification indices we gradually added covariance and meaningful regressions, and eliminated non-significant parameters in pursuit of a parsimonious model with optimal fit. With FIML as an estimator we were able to use 410 responses on a total of 470 responses. The final model fits the empirical data well (RMSEA = 0.029; SRMR = 0.015; CFI = 0.993; TLI = 0.961) and significantly better than the starting model ($\chi^2(7)=139.7$, p <.001).

Table 10. Overview of the survey scales

Scale ***Example item***	**Number of items**	**Cronbach's Alpha**
SPF use	4	0.80
At our school, this feedback report was thoroughly analyzed.		
Affective attitude	4	0.77
I enjoy engaging with feedback from NA/PT.		
Cognitive attitude: Usefulness for school development	4	0.83
I consider this feedback report useful for supporting vision development within the school.		
Cognitive attitude: Usefulness for accountability	4	0.77
I consider this feedback report useful for justifying our methods to outsiders.		
Subjective norm	5	0.67
People whose opinion I value, expect me to engage with feedback from NA/PT.		
Self-efficacy	5	0.82
I feel I have the necessary skills to understand the content of (the) feedback reports (from an NA).		
Shared goals	4	0.82
At our school, there is a clear vision about how to use feedback from NA/PT.		
Support and collaboration	4	0.87
At our school, people make optimal use of each other's skills in order to engage with feedback from NA/PT.		

4 Findings

In this section we will present the results of the path analysis. In order to provide some perspective to these explanatory findings, we will first briefly discuss descriptive findings for the scale variables from the survey.

4.1 Descriptive findings

As shown in Table 11, overall, Flemish educational professionals report fairly limited school-level use of SPF from NA and PT (*M* = 3.24) although responses vary considerably (*SD* = 1.10). A closer examination of the individual items revealed that in general, SPF reports are formally discussed with the team, but much less thoroughly analyzed, used as an impetus for further inquiry, or as input for formulating actions.

Table 11. Descriptive statistics of the survey scales

Scale	**n**	**M** [a]	**SD**
SPF use	426	3.24	1.10
Affective attitude	420	3.15	0.78
Cognitive attitude: Usefulness for school development	445	3.97	0.78
Cognitive attitude: Usefulness for accountability	417	3.41	0.83
Subjective norm	363	2.92	0.79
Self-efficacy	448	3.92	0.69
Shared goals	416	3.17	0.84
Support and collaboration	420	3.30	1.03

Note. SPF = school performance feedback.
[a] Mean values ranging from 1 (entirely disagree) to 5 (entirely agree).

We find that users' affective attitude towards the use of NA and PT feedback is neutral (*M* = 3.15). Cognitively, they do not take an outspokenly positive or negative stance towards its usefulness for accountability (*M* = 3.41) but do regard it as a rather useful tool for school development (*M* = 3.97). There is little indication that respondents feel pressured to make use of SPF from NA and PT (*M* = 2.92). Note, however, that the subjective norm scale comprises quite some missing data: only 363 responses out of the overall 470 could be used in the descriptive analysis of this variable. Looking at users' perception of control over SPF use, we see they gauge their self-efficacy as rather high (*M* = 3.92). So, they are quite confident they possess the necessary competences to process the results fed back to them from NA and PT. A relatively small standard deviation indicates that users are relatively united in this perception (*SD* = 0.69).

Lastly, the data indicate that users do not experience a strong school culture towards SPF use from NA and PT. On average, users neither agree nor disagree with the thesis that their school team shares a common goal-orientation regarding SPF from NA and PT (*M* = 3.17). Their perception of support and collaboration is somewhat more positive but still situated towards the center point of the scale (*M* = 3.30). Note, however, that there are relatively large differences between individual users in this respect (*SD* = 1.03).

4.2 Explanatory findings

The final path model is graphically represented in Figure 5. Full line arrows indicate regression and double-headed dashed arrows depict covariance. The standardized coefficients and the significance level of the effects are included. For user beliefs as mediating variables, and for school-level SPF use as a dependent variable, the R^2 values are mentioned in bold. For the sake of clarity, the figure only comprises those effects that are statistically significant (p < .05). A full overview of all parameters is given in Appendix C and Appendix D.

4.2.1 The (mediating) effect of user beliefs on school-level SPF use

Users' perception of the SPF's 'Usefulness for school development', and their 'Subjective norm' or the extent to which they feel it is expected of them to engage with the SPF, both bear a small positive relationship with school-level 'SPF use' (β = .17 and .13 respectively). Thus, users' cognitive attitude and normative beliefs have a statistically significant impact on school-level SPF use (cf. RQ1). These beliefs mediate the effects of SPF-oriented school culture and users' work roles to a certain extent (cf. RQ2b). Users' perception of 'Shared goals' explains about 6% of the variance in perceived 'Usefulness for school development' (β = .24) and some 13% of the variance in 'Subjective norm' (β = .36). A regression coefficient of .13 shows that only a further 2% of variance in the former is explained by whether or not the user holds a coordinating role at the school (as opposed to primarily being a teacher). Coordinators report a higher awareness of the feedback's potential for school development purposes, which is in turn associated with higher levels of reported school-level SPF use.

No other (mediating) effects of user beliefs on SPF use were identified (cf. RQ1 and RQ2b). The fact that users' affective attitude does not have a statistically significant impact on SPF use is in line with other studies exploring data use in Flanders (Van Gasse et al., 2015) but contradicts other findings that show that a favorable affective attitude can outweigh a favorable cognitive attitude (Vanhoof et al., 2014).

User beliefs do covary to varying extents. For instance, all factors on the attitudinal level covary positively, meaning that users' cognitive attitudes towards SPF (use) appear to run parallel to their affective attitude to a certain extent. 'Affective attitude' also covaries positively with 'Self-efficacy' (*cov* = .23), while negative covariance is found between 'Self-efficacy' and 'Subjective norm' (*cov* = -.11). Thus, users' level of enjoyment in engaging with SPF appears to correspond to a certain extent to the way they feel capable of this engagement, a finding we can relate to the fact that self-efficacy is often inversely related to anxiety (Dunn, Airola, Lo, et al., 2013b). On the other hand, users who experience more pressure to put SPF to use, appear to feel less capable of effectively doing so. Finally, we see that all user beliefs are impacted to varying extents by one or more school-level features we have taken into account.

4.2.2 The direct and indirect impact of school-level features on school-level SPF use

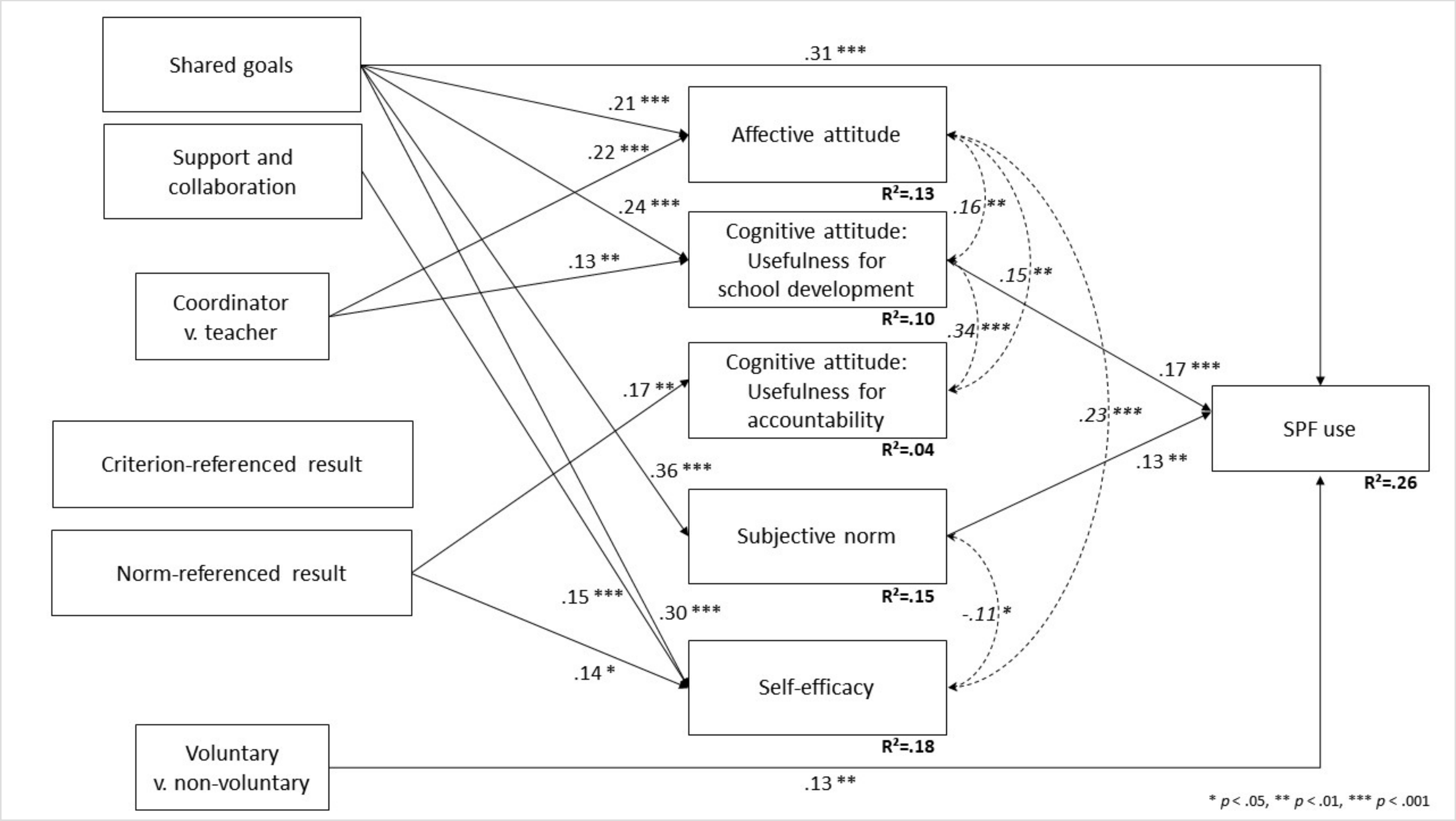
Overall, the full path model explains about 26% of the variance in school-level SPF use (R^2 = .26). An SPF-oriented school culture, and more specifically users' perception of a shared goal orientation, proves to be the most salient determinant of school-level SPF use when controlling for other factors (cf. RQ2a). A feedback user's perception of 'Shared goals' regarding the use of SPF from NA and PT affects the actual school-level use of this feedback directly (β = .31) as well as indirectly through perceived 'Usefulness for school development' and 'Subjective norm', as elaborated on above. 'Shared goals' also have a statistically significant impact on users' 'Affective attitude' (β = .21) and their 'Self-efficacy' (β = .30). The fact that this variable positively influences most of the user beliefs we measured, confirms that beliefs take shape within professional communities (Coburn & Talbert, 2006; Datnow & Hubbard, 2016; Jimerson, 2014). Users' perception of the usefulness of the SPF for accountability purposes is the only belief-driven predictor not affected by 'Shared goals'. However, accountability is not a dominant goal in this specific case and research context. Users' 'Self-efficacy', or their conviction of being capable of engaging with SPF, is further enhanced when they sense internal 'Support and collaboration' (β = .15). This is in line with our hypothesis and with other research (Abrams et al., 2021). As elaborated above, however, there is no ensuing statistically significant effect of 'Self-efficacy' on school-level SPF use.

Secondly, we find that users' formal work roles play a part in predicting their attitude towards SPF from NA and PT, which is in line with prior research (Van Gasse et al., 2015). Overall, coordinators tend to have a more positive 'Affective attitude' towards SPF use (β = .22) than teachers and they value the SPF's 'Usefulness for school development' somewhat higher (β = .13). As established above, this dimension of cognitive attitude is in turn a predictor for school-level SPF use (cf. RQ 2b). Work role does not have a statistically significant impact on users' 'Subjective norm' nor their 'Self-efficacy'. The latter is particularly counterintuitive, since we would expect coordinators to be more familiar with SPF-type data and thus more confident in handling those data. However, perhaps coordinators are more aware of the complexity of this feedback, precisely because they have more experience in dealing with it.

Thirdly, we see that feedback sign has no direct or indirect effect on SPF use (cf. RQ 2a and RQ2b). The school's criterion-referenced result bears no statistically significant relationship at all to other variables in our model: the extent to which the school reaches the attainment targets is not proportionate to the favorability of user beliefs associated with the SPF. The norm-referenced result, however, is positively related to self-efficacy beliefs (β = .17) and to users' assessment of the SPF's usefulness for accountability purposes (β = .14). Thus, a more positive comparative result is associated with a stronger sense of being able to 'make sense' of the result. This suggests that, when a school scores along the average, it can be difficult to base a conclusion on that result, but when the comparative result is more saliently positive it is easier to process. The fact that it does not correspond to higher levels of perceived usefulness for school development nor to higher levels of SPF use, however, may entail that the conclusion is that no action needs to be taken when performance is satisfactory. However, higher norm-referenced performance is deemed more useful for accounting for oneself, which is perhaps not surprising. When a school has scored markedly better than the population, the team is more inclined to use that information to account for their policy or practice to other parties.

Fourthly, we find a clear and direct effect of voluntariness on school-level SPF use (cf. RQ 2a). SPF actively requested by schools is associated with a higher level of use (β = .13) than feedback simply presented to schools. There is no mediation: whether the SPF was collected actively or not, has no statistically significant impact on user beliefs.

Figure 5. Path model



Note. SPF = school performance feedback.
*p < .05. **p < 0.1. ***p < .001.

5 Discussion and conclusion

Feeding back output indicators to schools is at the nexus of school effectiveness and school improvement (Hulpia & Valcke, 2004; Visscher & Coe, 2003). The present study approached engagement with SPF from low-stakes standardized assessments as both a belief-driven phenomenon and a situated phenomenon. It sheds new light on the relative impact and the interplay of predictors of SPF use. By conducting quantitative research on a large dataset, we explored how school-level SPF use is affected by individual users' attitudes, subjective norm and self-efficacy (RQ1). The analyses show that engagement with SPF in schools is only explained by these user beliefs to a limited extent when we control for other factors. We also examined the effects on SPF use of selected school-level features, namely SPF-oriented school culture, users' formal work roles, feedback sign and voluntariness (RQ2a). In addition, we investigated whether those effects are mediated by user beliefs (RQ2b). Our hypothesis that a user-centered outlook on SPF use needs to be situated within a school is confirmed to a certain extent. User beliefs regarding SPF are influenced by school-level features. However, not all of these relationships lead to a heightened use of SPF on school level. Thus, the mediating role of user beliefs is modest.

On the level of the individual user, our most salient findings include that SPF use will increase when users recognize the utility of SPF for school improvement. We also established that users report a higher level of engagement with SPF at their schools when they have a stronger sense that they are expected to use SPF. On a school organizational level, users' work role plays a small part in explaining perceptions about SPF use. Above all, however, our analyses demonstrate that a strong data culture is an important precondition for engagement with SPF. Users' perception of a shared goal-orientation within the team emerges as the most prominent predictor in the model. As in certain other studies, our analyses point out that it has a greater positive impact on data use than individual user characteristics (e.g., Van Gasse et al., 2015). On a contextual level, we found that data from a voluntarily adopted SPF system are used more intensively than data not actively collected. This suggests that also in an improvement-oriented context, ownership is an important driver.

Our findings attest to the fact that data themselves "do not drive" (Dowd, 2005; Lockton et al., 2019) and entail implications for practice. As with other types of educational innovations and interventions, and organizational change in general, data should meet users' improvement needs and address what they find important (Ketelaar et al., 2012; Schildkamp & Teddlie, 2008). Research suggests that a positive cognitive attitude towards DDDM can be stimulated by providing instruction to educators. Instruction serves to address concerns regarding data use, and serves to identify and challenge views that might (otherwise) lead to reluctance to engage in DDDM (Dunn et al., 2019).

In the interest of ownership and putting educational professionals in the driver's seat of data use, inquiry-based working should be stimulated in schools. Programs and interventions can also explicitly promote a stronger data orientation in school teams. The purpose of data such as SPF should be clear to the educational professionals expected to make use of them. Those expectations should be made explicit, so that individuals sense data use is an integral part of their job and that their contribution is vital to the organizational dynamic of data use.

In the spirit of fostering continuous quality monitoring and continuous improvement, the introduction of professional learning communities such as data teams (Schildkamp et al., 2019) or networked improvement communities (LeMahieu et al., 2017) can strengthen data use practices and data cultures in schools. Such initiatives foster shared goals and provide a collaborative space in which individual perspectives and expertise are appreciated. School improvement becomes continuous improvement, a form of organizational learning (Datnow & Park, 2018; Dolle et al., 2018). A data culture is a setting in which data use is done by, not to the school (S. Sutherland, 2004).

Concerning schools' performance, we found that feedback sign has no statistically significant effect on school-level SPF use in our research context. The literature suggests that, whereas the effect of feedback sign on perceptions is rather straightforward, its effect on actual behavior is much more complex (Kluger & DeNisi, 1996; Lechermeier & Fassnacht, 2018). In empirical studies on SPF use, being confronted with lower performance has indeed been found to prompt action as opposed to receiving positive feedback, but very negative feedback tends to be brushed off (Hellrung & Hartig, 2013; Verhaeghe et al., 2010). Drawing on social cognitive theory, (Visscher & Coe, 2003) addressed tensions surrounding SPF sign as follows: "Although negative feedback is necessary to motivate the need for improvement, without positive feedback, individuals are unlikely to believe themselves capable of achieving it" (p. 326). We contend that SPF systems can help educational professionals in making sense of their results by providing more sense in the message itself. In line with feedback intervention theory (Kluger & DeNisi, 1996), as discussed by (Visscher & Coe, 2003) SPF systems should provide sufficient cues in the feedback message – also when that message is positive. Positive feedback tends to be automatically processed and is therefore more easily ignored, but providing more complexity to positive feedback turns it into guidance and not "just" praise (Geddes & Linnehan, 1996).

Limitations and suggestions for further research

A TPB-approach assumes a linear perspective on behavior. In practice, however, data use is not a linear process. Therefore, further research should incorporate effects in other directions, such as the effect of school-level data use on individual beliefs (cf. Datnow & Hubbard, 2016) or the effect of individual users' attitude and self-efficacy on collaboration (cf. Dunn, Airola, Lo, et al., 2013a; Van Gasse et al., 2017). The TPB framework we applied can also be further extended. Researchers could consider including a measure for behavioral intention as a motivational mediator between psychological factors and behavior (cf. Prenger & Schildkamp, 2018), as well as a measure for actual behavioral control. Actual competences have featured in other studies about the influence of personal characteristics on data use (e.g., Vanhoof et al., 2011) and we should investigate how they complement the present model. While educators' data literacy is a fundamental prerequisite for effective DDDM, we know that educators' self-efficacy regarding data use, one of the central predictors in our conceptual model, is often not in line with their actual knowledge and skills (Dunn, Airola, & Garrison, 2013).

Additionally, the construct of perceived control in itself can be revisited. We limited this variable to a self-efficacy measure, in line with a strand of other TPB-based studies (Armitage & Conner, 2001). However, perceived control over data use can also pertain to users' sense of having straightforward and timely access to the data (Pierce et al., 2013) or to their perceived autonomy in the decision-making process (Prenger & Schildkamp, 2018). Therefore we recommend exploring self-efficacy and other operationalizations of perceived control as separate constructs (cf. Prenger & Schildkamp, 2018). Furthermore, we point out that self-efficacy, like data literacy, is a multilayered construct. We captured and combined users' confidence in interpretation and their confidence in transforming information, in one overarching 'Self efficacy' measure. Future research should distinguish between different dimensions of efficacy, as prior research has established that these are indeed separate constructs and are regarded as such by data users (Dunn, Airola, Lo, et al., 2013b).

Further research could also consider fine-tuning other factors. We attempted to explain the mechanics of SPF use and not its actual effects, but did not differentiate between different phases. It would be beneficial to look at interpretation, analysis, and translation into action separately in order to account for the individuality of each phase and the mechanisms at play in each of them. Furthermore, we regarded feedback sign as an objective attribute of the SPF. It would interesting to reconceptualize this variable into feedback valence. A message's valence pertains to its attractiveness, i.e., whether it is perceived as positive or negative by the recipient (Geddes & Linnehan, 1996). Since several other predictors in our model are based on perceptions of SPF users, it would make sense to do the same for the feedback sign variable.

Because of the specificity of the research context, it is advisable to replicate this study in order to explore the generalizability of our findings. Replication studies could consider pursuing a multilevel design in order to further unravel individual differences between users and further enrich the information on prevailing data cultures in schools. Nevertheless, setting up this study in Flanders, an educational context low in outcome accountability, has allowed us to retain a clear focus on improvement-oriented data use. Moreover, SPF from Flemish NA and PT constituted a particularly suitable case for exploring our research questions. For one, because of the inclusion of feedback sign, an understudied predictor of data use. Results on Flemish NA and PT are research-based, refer to explicit standards and offer a clear normative benchmark. Additionally, because of the inclusion of voluntariness in feedback pursuit as a potential determinant of SPF use. A comparative analysis of NA participants and PT takers captured differences in engagement between two conditions in which the instrument was the same or very similar. Overall, Flemish schools provided a fruitful context for hypothesizing differential effects as they show great variability when it comes to data use (Vanhoof et al., 2012), which was confirmed by our descriptive analysis of school-level SPF use.

General discussion

Prevalent assumptions in educational policy and test development are that data from summative assessments can serve formative purposes, and that offering high quality data to schools will successfully inform and drive school improvement endeavors. Consequently, test developers, feedback providers and policymakers strive to provide schools with robust data, in order to empower educational professionals to make informed decisions regarding school policy and instructional practice. System-level assessments, standardized tests, and other school-external tools and interventions feed a wealth of data back to schools in the form of score reports and school performance feedback. The expectation is that recipients will use those output data as a mirror to identify strengths and weaknesses, and as an impetus to take action accordingly.

From a technical-rational perspective, this is fairly straightforward. Data function as input, improvement actions are the output. In reality, however, there are stumbling blocks and deviations along the way. Educational professionals do not simply *use* data, i.e., receive a clear message and implement adjustments accordingly. From their own subjective backgrounds and within their own contexts, data users *make sense* of the data they receive or encounter.

In this dissertation, we explored how educational professionals make sense and make use of school performance feedback. By collecting theoretical insights from the data use literature, and empirical evidence from the context of the Flemish national assessments and parallel tests, we aspired to shed light on the mechanisms at play, and on factors that underlie and influence these mechanisms.

In this final chapter, we first consider the key findings from the studies that have been conducted. Subsequently, we reflect on how this dissertation contributes to the knowledge base on data-based decision making in general and school performance feedback systems in particular. We discuss strong points and vulnerabilities, and suggest pathways for further research. To conclude, we offer a number of recommendations for policy and practice.

1 Main findings

In line with previous research, we have argued extensively that educational professionals make sense of data such as school performance feedback from their own *subjective* backgrounds, within their own *contexts*. Guided by this general idea, our discussion of salient overarching findings from the four studies that were presented in this dissertation, will be structured along two dimensions. We will first reflect on processes that take place when (individual) educational professionals engage with school performance feedback. Picture the teacher or school leader who picks up a fresh report, and tries to wrap their head around the information they received. What have we learnt about what happens there? Second, we take a step (or several steps) back, and discuss how we have found educational professionals' use of school performance feedback to relate to the groups and systems they belong to. Of course, such a simple bifurcation does not do justice to the kaleidoscope that real-life sensemaking truly is, as discussed at length in Study 1. However, it is a way to anchor our insights and observations.

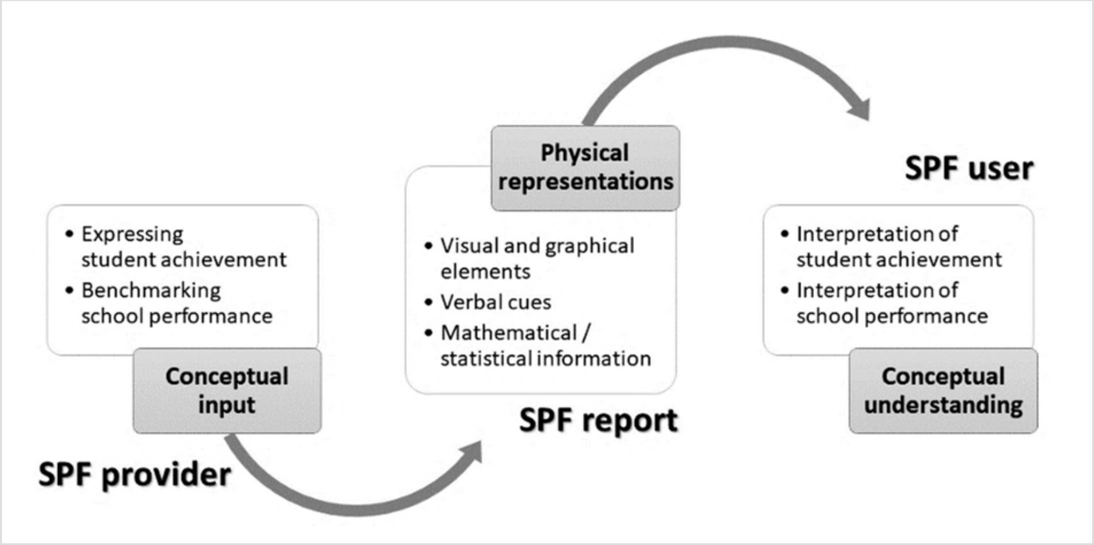
1.1 Findings relating to subjective interpretations of school performance feedback

Data acquire meaning only when someone gives meaning to those data

The realization or consideration that raw data are essentially just numbers, colors, graphs, text on a page, and that they do not convey an unequivocal message to recipients, is central to the sensemaking perspective we applied to school performance feedback use. In Study 3, we indeed found that school leaders and teachers looking at the same feedback reports, construed a different story based on the data – at least in terms of causal ascription. However, subjective meaning making already starts at a much earlier stage of course. When users make sense of data (i.e., form an understanding, construe a message, and then use this message to construct implications, Coburn & Turner, 2011) they start by individually picking up cues from raw data. They read the reports and figure out what the graphs and numbers mean. This phase of data analysis is the bud for further sensemaking to bloom from.

In Study 2, we zoomed in on this nuclear stage of sensemaking, i.e., the phase of data analysis, which precedes diagnosis and further stages of educational decision-making. Our guiding assumption was that users' early interpretations and (mis)conceptions determine the validity of their further inferences and thus the subsequent steps of the decision-making process. We came to the conclusion that there is substantial variability in users' conceptual understanding of school performance feedback elements. Most participants grasp the broad conceptual outlines of the feedback, in terms of how student achievement is expressed, and in terms of school performance is benchmarked. However, both over participants and over different components of the feedback *within* participants, there is a continuum of being able to construct a basic understanding to mastering more sophisticated conceptualizations. We established that typical building blocks of school performance feedback can become stumbling blocks, and that they do so differently for different people. So, the meaning that data providers lay into physical representations, is not a package for the user to receive and just unwrap. A feedback user starts from these physical representations in order to construct their own meaning (cf. Figure 6). Ultimately, this is the message that will live on in the construction of further narratives.

Figure 6. Simplified representation of conceptual transfer in school performance feedback (SPF) reports


Physical representations
SPF user
• Expressing student achievement
• Benchmarking school performance
• Visual and graphical elements
• Verbal cues
• Mathematical / statistical information
• Interpretation of student achievement
• Interpretation of school performance
Conceptual input
Conceptual understanding
SPF report
SPF provider

Data literacy is multidimensional

As discussed in Study 1, making sense of data is a complex problem-solving process that is not always systematic and rational, and that consists of many underlying micro-processes. Sensemaking of data such as attainment scores or school performance feedback reports comprises analyzing the raw data, making inferences, and deciding upon how to act in a next step. It entails asking oneself ‘what the data mean’, ‘what the data mean for one’s school’, and how to move forward with this knowledge, i.e., ‘what to do next’. Consequently, making sense of data such as school performance feedback calls upon users’ knowledge and skills: it requires of educational professionals that they understand the numbers and that they are able to make valid inferences based on the data (Mandinach & Schildkamp, 2021a; Schildkamp, 2019; Vanlommel & Schildkamp, 2019). This illustrates why data literacy, i.e., the capacity that is required to make sense of data in a valid and meaningful way, should not be narrowed down to *only* refer to statistical literacy. Contemporary views understand and underscore that it actually comprises a far richer spectrum of knowledge and skills (Beck & Nunnaley, 2021; Groß Ophoff et al., 2015; Groß Ophoff & Egger, 2021; Mandinach & Gummer, 2016; Vanlommel, 2022). Data literacy does not merely refer to the capacity to figure out what the numbers mean. It also entails the capacity to translate these numbers into information, relate that information to goals and issues, formulate decisions, evaluate outcomes and so on.

In the qualitative studies in this dissertation, we looked at a number of sensemaking micro-processes. In Study 2, we zoomed in on users' analysis of school performance feedback (picking at numbers and graphs in order to decipher them), and in Study 3, on their appraisal (determining whether a result is good or bad) and attribution (reflecting on the causes of a particular result) of reported performance.

As discussed above, in Study 2 we found that participants' comprehension of typical school performance concepts is clouded by misconceptions, be it to varying degrees. Also to varying degrees, such misconceptions compromise further interpretation. In line with prior research, our findings suggest that some educational professionals lack the necessary statistical literacy to accurately process school performance feedback. However, we also concur with the view that it is for feedback developers to consider the(mis)alignment between the statistical complexity of the data they provide on the one hand, and the statistical literacy of intended users on the other hand. The interpretability of feedback reports is a critical precondition for school performance feedback to realize its potential for school improvement (O'Leary et al., 2017; van der Kleij et al., 2014).

In Study 3, we observed that formulating a diagnosis based on school performance feedback is not straightforward for educational professionals. In any case, we noticed that many of the study participants mentioned a very wide range of factors when making causal ascriptions for their scores. Moreover, there is not always a clear partition between the causal models they invoke when making these ascriptions. This is not problematic in se, but if so many things play a role (which, in reality, they of course do), the question is where to start in terms of formulating improvement initiatives. Furthermore, our findings suggest that feedback users – regardless of the (perceived) favorability of their feedback – are predominantly inclined to look for clues on what is going wrong. Identifying what is going well and what needs to be fostered, appears to be a bit of a blind spot.

In the quantitative survey for Study 4, we gauged Flemish teachers' and school leaders' perceived self-efficacy in terms of making sense of school performance feedback. Perhaps surprisingly, descriptive results indicate that they overall feel fairly confident that they are able to interpret the numbers and translate them into actions or decisions. From prior research, we know, however, that educational professionals' self-efficacy regarding data use does not always correspond to their actual knowledge and skills (Dunn, Airola, & Garrison, 2013). This discrepancy could lead to a sour reality check. When users experience difficulties with initial analysis and interpretation, they sometimes give up altogether on forming a diagnosis, let alone that they formulate an action or decision in response (Verhaeghe et al., 2010).

Sensemakers seek coherence

Personal frames of reference play a very large role in sensemaking. Sensemaking is about figuring out how (and whether) you can fit 'the new' in with 'the old', in such a way that it clicks. When you receive new information, can you preserve your prior assumptions about how things work? Do you need to adjust your mental models? Would you be prepared to do so? Or, in order to restore cognitive consonance, are you going to dismiss information that does not fit and leave it at that? In Study 1, we discussed that sensemakers search for coherence between new information and information that is (or is assumed) given. We also touched upon questions of identity. Throughout their career, educational professionals develop an interpretative framework that comprises ideas about learning and instruction, but also about the self 'as an educational professional' (Kelchtermans, 2009, 2018). Sensemakers' search for coherence is shaped by their experience, their perceptions and by the roles they hold.

In all studies, we found in some way that users' (prior) knowledge, skills, experiences and subjective assumptions form the proverbial glasses through which they look at data such as school performance feedback. In Study 2 we established that, generally speaking, educational professionals look at school performance feedback from a completely different frame of reference than data providers, and that their level of (statistical) data literacy is not always in line with the level that is assumed. Additionally, different participants expressed different information needs and visualization preferences, which were not always line with the data as offered. However, the feedback reports from our research case are static and stand-alone: what you see is what you get, and that means that the snapshot provided does not always fit into the frame that was already on the proverbial wall.

We also found that users' work roles play a part. Patterns identified in Study 3, for instance, clearly showed that school leaders and teachers emphasize different factors when coming up with causal explanations for school performance, with school leaders being somewhat more focused on school-level (policy-related) factors whereas teachers tend to discuss more student-level factors. Also, the path model in Study 4 showed that a leadership role is positively associated with attitudes towards school performance feedback, compared to teachers.

1.2 Findings relating to the contextual embeddedness of school performance feedback use

Discourse and dialogue are key

In Study 1, we addressed the importance and the mechanisms of collective sensemaking of formal achievement data such as school performance feedback. Whereas our empirical studies focused more on individual than on collective sensemaking, they do provide cues as to why collective sensemaking is important. Study 2 showed that individual educational professionals do not always succeed in interpreting all elements of a school performance feedback report in a sensible way. Discussing reports among team members or even with external parties, entails meaning negotiation and making frames and interpretations explicit. This could potentially help overcome such individual difficulties. It can also offer a validity check of interpretations, reminiscent of the 'four-eye principle' from management contexts. After all, it is to be avoided that erroneous interpretations go on to assume a life of their own, and *become* the narrative: educational decision making merits more careful deliberation than that.

Aside from greater accuracy, collective sensemaking may produce fuller interpretations and richer inferences as well. In the interviews for Study 2 and Study 3, we noticed that school leaders appear somewhat more interested in norm-referenced feedback (that position their schools) while teachers seem more drawn to criterion-referenced results (that describe whether objectives were met within the school). In Study 3, we saw that school leaders and teachers from the same school often emphasize different categories when explaining (or reflecting on) their school's performance in the national assessment. School leaders apply a policy outlook, while teachers reflect more on the input from students. Combining perspectives could broaden the frame. Note also that in Study 4, perceived support and collaboration within the team was associated with a higher sense of self-efficacy in terms of interpreting the feedback.

Making sense of data does not happen in isolation

In Study 1, we discussed how the organizational context and social context of a sensemaker (in short, the various groups to which a sensemaker belongs) shape their frames of reference. The path model presented in Study 4 attests to the effect of school predictors on subjective frames and perceptions regarding school performance feedback. A perceived shared goal orientation in terms of feedback use is associated with more positive user beliefs: a higher sense of self-efficacy in terms of engaging with school performance feedback, more positive attitudes, and feeling personally 'expected' to do something with the data to a greater extent. Moreover, shared goals had a direct effect on school-level feedback use. So, it would appear that a strong, shared vision on feedback use is a tonic for engagement with school performance feedback overall. Furthermore, in Study 3, it was striking that school- and class-level attributions were more often than not formulated in the we-form. We regarded these as 'collective internal' attributions. Their formulation suggests that the feedback users we interviewed, regard school performance as a team effort and a collective responsibility.

Study 1 also discussed the importance of leadership. The role of leaders (principals, but also informal leaders) as 'sensegivers' can hardly be overstated. Principals, department heads, as well as other pacesetters create vision around data use, make room for data use, and can perhaps motivate the team to do more with data. This sensegiving role of leaders is also important in light of the broader educational context. Educational governments, umbrella organizations, boards and networks, make a lot of initiatives available to schools and express expectations about what should be achieved through assessment. A school leader translating those expectations to the team, determines to a great extent how (data) policies are implemented in practice.

Users' real-life sensemaking challenges data providers' assumptions

Approaching educational professionals' use of school performance feedback as a sensemaking process, provides cues to understand why data are sometimes not deployed as policymakers and test developers would wish, expect, intend (cf. Study 1). Data providers might like to assume that the provision of sound educational effectiveness data is, in itself, an impetus for educational professionals to thoroughly reflect on their policy and practice and think about what needs to be done. However, reality bids some complications.

First, while it is important that schools have access to high quality data such as school performance feedback, the *mere* availability of such data is not necessarily a driver. In Study 4 we found that users' perceived expectations of others for them to engage with feedback (subjective norm) have a positive effect on school-level feedback use. However, we noted the highest level of non-response on the survey questions pertaining to subjective norm, which is telling in itself. Recipients do not necessarily feel that expectations to engage in feedback use are a 'thing', or they simple do not know whether those expectations exist. Moreover, expectations could emanate more from the schools' internal data culture, rather than that educational professionals necessarily regard feedback use as a part of their job. On a contextual level, we found that data from a voluntarily adopted school performance feedback system (i.e., parallel tests) are used more intensively than data not actively collected (i.e., feedback to sample schools from national assessments). We concluded that ownership is an important driver. Data that provide an answer to an existing question, are used more intensively than data that require users to *come up with questions* which those data might answer. These findings, together with the observation that perceived usefulness for school development is a driver for feedback use, suggest that strong data cultures *in* schools ultimately have a bigger impact than the mere fact that data *come into* the school.

Second, educational professionals do not *only* look at their own practice and policy when formulating a diagnosis for school performance. Study 3 found that results are attributed to external factors to a great extent. This does not have to be problematic per se, but it defies ideals relating to data-based decision making. Moreover, we found that formulating a diagnosis is not straightforward. Teachers and school leaders make a wide range of attributions. Again, this does not have to be problematic, but it attests to the complexity of data-based decision making. In prior research on school performance feedback use, Flemish school leaders indicated that they felt particularly lost in the diagnostic phase, not only because of a perceived lack of support and guidelines, but also due to a perceived lack of concrete suggestions for improvement in the reports they received (Verhaeghe et al., 2010). This suggests that they expect (cues for) attribution in the data they are presented with. While data from summative external assessments can inform formative conclusions – and are, of course, explicitly divulged to schools for these purposes – it is left up to the recipients to contextualize the data and interpret results in light of their own goals (van der Kleij et al., 2015). In Study 2, we identified a perceived lack of normative cues in the feedback reports as one of the disconnects between what is provided, and what users expect to receive. 'Raw' reports contain extensive expositions, nice colors and pictures, but, at the end of the day, what exactly constitutes a 'good' or 'bad performance'?

Thus, we established that the availability of school performance feedback should not be assumed to be a driver in itself, and that it is not a given that such data automatically give rise to introspection. Another assumption refuted by reality is that the data transmit a clear and unambiguous message to the recipient.

For example, we looked at the influence of the feedback sign in Studies 3 and 4. One might assume that there are clear effects. For instance, that a particularly negative result would trigger avoidance or action, that a particularly positive result is associated with more favorable beliefs about the feedback, that average performance is answered with acquiescence or complacency, and so on. In line with other scholarship on feedback effects, we experienced that it is not all that straightforward. In Study 4, schools' criterion-referenced results showed no statistically significant relationship at all with other predictors in our model. So, after controlling for other variables, the extent to which the attainment targets were reached in the school – be that a low or high proportion of students – had no impact on user beliefs about the feedback, nor on the extent to which the report was put to use. Schools' norm-referenced result, on the other hand, did have an effect. Within our model, performing better than average was positively related to users' self-efficacy and perceived usefulness of the feedback for accountability purposes. In Study 3, we found that attributions about (the nature of) the assessment were (primarily) made to explain results perceived as negative.

Overall, we found that there is a language barrier (so to speak) that needs to be overcome. School performance feedback systems express achievement and benchmarks with a variety of terms and visualizations. Some of these are foreign to educational professionals (such as confidence intervals) and risk to be dismissed on that account. Others seem familiar but do not exactly designate what some feedback recipients read into them (such as 'average' or 'score'). A very specific disconnect that we identified in Study 2, is the clash in psychometric perspectives when participants use a classroom assessment paradigm to interpret external assessment results. Thus, it is essential to find some sort of common vocabulary and imagery between feedback provider and feedback user. Both parties should not assume they are speaking the same language.

2 Strengths and significance

2.1 Theoretical contribution

By examining educational professionals' use of school performance feedback from external standardized assessments, this project is at the nexus of educational effectiveness and school improvement research. A central goal of educational effectiveness research is to unpack factors that have a (differential) impact on educational quality at the system-level, on school performance, on teacher effectiveness, on student achievement. What works and what does not? Improvement research seeks to unravel how systems and schools can change for the better. Which changes and innovations are needed to boost outcomes? The provision of school performance feedback attempts to bridge a gap between both fields (i.e., a gap between identification and transformation, between evidence and elevation) (Hulpia & Valcke, 2004). However, using school performance feedback for school development is a human endeavor, not (always) a linear and predictable process that fits prescriptions and assumptions, and it does not always move forward in a desired direction. Precisely those micro-processes that give substance to data-based decision making, make up the path between ideal and actuality. This realization underlies the different studies we conducted. By purposely zooming in on and out from the 'actualities' of data use, we contributed to the existing knowledge base on data-based decision making in education in general, and school performance feedback systems in particular.

In Study 4, we explored how different predictors interplay in explaining school-level use of school performance feedback. We considered the Theory of Planned Behavior to be a suitable guiding framework, since the conceptualization of data-based or data-driven decision making as "the systematic collection, analysis, examination, and interpretation of data to inform practice and policy" (Mandinach, 2012, p. 71) implies that data use has an overt behavioral quality. In Study 3, we applied attribution theory to teachers' and leaders' causal ascriptions of school performance feedback. The aim here was to reflect on whether users seek the source of school performance within themselves, or look outward. While it is essential for teachers and school leaders to reflect on the nature of assessment and on student characteristics, the scales can also tip: a lack of introspection might encourage deficit thinking and inequity. In Study 2, we contributed to the literature on good practices for score reporting. From a user validity perspective, we argued that disconnects between users' and providers' frames of reference are an important factor in explaining users' difficulties in terms of conceptual comprehension of school performance feedback.

The core of the theoretical component of this dissertation, however, is formed by Study 1. Prior research has established that data use in general, as well as school performance feedback use in particular, is influenced by aspects of information (systems), individual users, school organizations, and the broader context (e.g., Coburn & Turner, 2011; Schildkamp et al., 2017; Schildkamp & Kuiper, 2010; Visscher, 2002). Study 1 proposes that keys to unlock these mechanisms, might be found by centering our gaze on the data user of flesh and blood – ultimately the alpha and omega of data use. Data use can be "messy" (Bertrand & Marsh, 2015) and providing data to educational professionals might not produce desirable or intended outcomes. Sensemaking perspectives complement technical-rational perspectives on data use (Datnow et al., 2012) by illuminating why that is the case. Sensemaking, as a general framework or perspective, is particularly suitable to apply to data use because it highlights how people position and explain (new) information or (unexpected) events, and how they subsequently proceed. It has a cognitive aspect: when you start trying to understand something new, you challenge it against your own prior knowledge, experience and assumptions: you look where it fits into existing frames. It also has a discursive aspect: implicitly or explicitly, you are going to try to build a story by putting into words "what is this, what is happening here?" and "what should be the next step now?".

Sensemaking perspectives have been proliferating in recent literature on teachers' and school leaders' data use. An integrated framework specifically discussing educational professionals' sensemaking of formal achievement data such as school performance feedback was, however, absent. The thorough literature review we carried out was undertaken with an open view, and included a number of studies rooted in other (educational) research traditions beyond the data use canon. We gathered lessons about interpretive mechanisms, about the role of prior knowledge and assumptions, and about what happens when people sit around the table together to discuss (formal achievement) data. Building on these lessons, a framework was constructed consisting of interrelated insights on the level of the data themselves, the data use process, the individual data user, the social context of the user, users' interactions, as well as the broader system level. The framework spotlights each of these levels in their own right, but also illuminates how and where levels connect, converse, converge. Both the individual insights and the overarching framework that were presented, have the potential to guide further research and inspire future practice and policy.

2.2 Application of multiple methods

This dissertation leans on theoretical and empirical pillars. In the empirical component, we used qualitative as well as quantitative research methods. By applying multiple methodologies, we were able to address research questions from various angles, and surpass some of the limitations of individual methods. This approach has added versatility and depth to our research and provides a more complete picture and a deeper understanding of the topic. Throughout all four studies, methods for data collection and analysis were critically chosen to align with our research aims. We took care to transparently motivate these choices, and to document the methods we applied.

In Study 4, both established and understudied predictors of feedback use were combined into a dual-focus model. A path analysis performed on a relatively large survey-based dataset, enabled us to study the relative impact of these predictors. In Studies 2 and 3, we adopted a qualitative approach because our aim was to deepen our understanding of how individuals and groups make sense of something they experience (here, receiving performance feedback for their school) from their own perspectives (Patton, 2015; Savin-Baden & Major, 2013). Interviews allowed us to tap into these perspectives and make them explicit (Patton, 2015), by probing prevailing perceptions in depth (Cohen et al., 2018; Savin-Baden & Major, 2013). In Study 2, our inquiry also incorporated a think-aloud procedure in order to examine actual user interpretations and conceptions with an open mind and (as) free (as possible) of assumptions. Framework analysis served as an overall analytical method in the qualitative studies, as it is fit to both organize and interpret the data, allows for a combination of inductive and deductive techniques, and facilitates the development of matrices to condense findings and present patterns (Gale et al., 2013; Ritchie et al., 2003). In order to identify trends, in Study 3 qualitative interview data were also 'quantitized' (Sandelowski et al., 2009). Conceptual work (Study 1) enabled us to make more insightful interpretations of the empirical results.

A final methodological note we would like to make in the margin, pertains to the qualitative studies in particular. Due to the COVID19-pandemic, the data collection for Studies 2 and 3 appealed on our creativity. As live interviews were not possible due to societal restrictions, we relied on video conferencing tools. These tools have been found to facilitate engagement (Archibald et al., 2019) and they provide screen and file sharing options, which allow for employing elicitation techniques. We experienced this as a major advantage, as it offered the opportunity to (record and) display fragments of the feedback reports under discussion during the interviews.

Moreover, we experienced that participants appreciated the convenience and time effectiveness of this method (cf. Archibald et al., 2019).

2.3 Specificity of the research context

As described in the general introduction, we know from prior research that there are substantial differences between Flemish schools and individual educators in terms of (perceptions about) data use. This provided us with fruitful research conditions for exploring usage and user perceptions relating to school performance feedback.

The Flemish educational system is characterized by low (outcome) accountability, and (thus) relatively atypical in terms of assessment practices. Overall, a lot of data use research on standardized testing focuses on assessments that hold considerable stakes for schools and/or students, more often than not situated in highly accountability-oriented contexts (Datnow & Park, 2018; Mandinach & Schildkamp, 2021a; Vanhoof & Schildkamp, 2014). In international scholarship, potential detrimental effects of control- or surveillance-oriented accountability pressures have been well-documented (e.g., Earl & Fullan, 2003; Nichols & Berliner, 2007; Nichols & Harris, 2016). Although we concur with Datnow and Park that “even when data use is framed in terms of continuous improvement, educators still may experience data use as a form of accountability” (2018, p. 137), our findings from Study 4 confirm that perceived external pressure is low in the Flemish educational context. Flemish educational professionals experience few expectations from others to do something with school performance feedback from national assessments or parallel tests. Therefore, we could say that a focus on school performance feedback use in Flemish schools is almost by default a focus on improvement-oriented data use. However, particularly in the interview studies, our findings confirmed that making sense of external assessment data is not self-evident for Flemish educational professionals. As researchers have stated before, perhaps this could precisely be attributed (to a certain extent) to a lack of accountability and to data use still being perceived as a ‘novelty’ (Schildkamp, Vanhoof, et al., 2012; Vanhoof et al., 2012). These observations provide food for thought, especially in light of the often-quoted adage from Earl and Katz (2006) that “accountability without improvement is empty rhetoric, and improvement without accountability is whimsical action without direction” (p. 12).

School performance feedback from the Flemish national assessments and parallel tests also constituted a suitable case for exploring our (empirical) research questions simply because of the information we were able to access and include in our sampling strategies and our data analyses. We were able to depart from authentic and scientifically-supported assessment data that contain criterion- and norm-referenced information, and investigate how these data are handled by the target groups of interest. In addition to teachers, who often form the primary focus in data use research, we also involved school leaders in our studies. After all, school performance feedback seeks to inform both instructional decision making and school policy.

Furthermore, we had the opportunity to examine how better or poorer school results interplay with recipients' perceptions and use of school performance feedback. In Study 4, feedback sign was used as a predictor. In Studies 2 and 3, we took school results into account in the sampling strategy, in order to stimulate variability and avoid self-selection, but also – particularly in Study 3 – to explore patterns according to perceived favorability. Finally, in Study 4, voluntariness as a potential determinant of feedback use could be studied from a comparative perspective by juxtaposing use cases of national assessment participants on the one hand and parallel test takers on the other hand. As national assessments and parallel tests offer the exact same kind of information, but the initiator differs between both contexts, the design of Study 4 bears resemblance to quasi-experimental research.

3 Limitations and suggestions for further research

3.1 Specificity of the research context (reprise)

System-level factors have an impact on sensemaking and data-based decision making in schools, as we argued in Study 1. Indeed, data use cannot be seen as independent of the educational context (Coburn & Talbert, 2006; Coburn & Turner, 2011; Malin & Brown, 2022; Mandinach & Schildkamp, 2021a; Schildkamp & Lai, 2013b; Verhaeghe et al., 2010). Cultural differences in attributions of student achievement have been reported (Matteucci & Gosling, 2004; Wang & Hall, 2018) which attests to the fact that the educational system (and its goals, norms, policies and customs) influences educational professionals' mental models and their sensemaking of achievement data (Mandinach & Schildkamp, 2021a).

The specificity of the Flemish context in which we operated, is an asset (cf. supra), but it is also a potential constraint. For instance, in Study 4 we included users' perceptions about the usefulness of the school performance feedback for accountability purposes as a predictor of feedback use. We found that a higher perceived utility of the feedback for these purposes is associated with schools who performed better compared to other schools. Additionally, it was the only predictor rooted in user beliefs that was *not* influenced by shared goals about school performance feedback use. These findings merit further interpretation, at least, or further inquiry. However, we need to acknowledge – again – that accountability is not an outspoken goal in this specific case and research context.

In order to examine to what extent results from our empirical studies are generalizable, comparative research is needed. An avenue to further explore in this regard, is the cohesion/regulation matrix used by authors such as Malin and Brown

(2022) to characterize educational systems. In this matrix, social cohesion refers to the extent to which actors in a system are prepared to engage in collaborations, while regulation refers to the degree of hierarchical control. The position of Belgium in this matrix characterizes the educational context as an individualist system (Malin & Brown, 2022; Vanlommel, 2022). It would be interesting to conduct in-depth comparative empirical research about educational professionals' perceptions and use of external assessments over different individualist systems, or over systems in multiple quadrants of the matrix.

3.2 Suggested methodological elaborations

As suggested in Study 4, the path analysis we conducted could be elaborated on or be re-imagined by involving more teachers and more schools, so that a multilevel perspective becomes possible. Provided that there are sufficient data points, i.e., schools and users within schools, this would allow for a more fine-grained analysis of how individual user beliefs relate to organizational characteristics. Our model could also be further refined to include a multigroup approach. This would allow for a comparison of models according to specific grouping variables, and could thus provide a deeper insight into how model components interact with, for instance, work role or educational level.

In the qualitative studies, we made efforts to have participants make sense of school performance feedback "as authentically as possible", for instance by asking them to mimic an imagined dialogue with a colleague during the think-aloud procedure. Nevertheless, we need to acknowledge that an interview setting with an external person creates artificial conditions. Therefore, throughout the dissertation we have echoed existing calls to conduct micro-process studies to truly unravel how educational professionals (both individually and collectively) make sense and make use of achievement data in situ (Little, 2012; Schildkamp, 2019).

Additionally, social network analysis might shed light on school teams' collective sensemaking of school performance feedback (systems) in particular. Do different team members interact with each other and with the school performance feedback data in order to bring different perspectives together, and how do they go about this (cf. Van Gasse et al., 2021)? Does school performance feedback contribute to organizational learning (cf. Supovitz, 2010), and if so, to what extent? By examining social ties between different actors, specifically with regard to school performance feedback use, social network analysis could illuminate the role of school leaders and (other) information brokers as sensegivers (cf. Daly, 2012). Are sensegiving mechanisms tied to specific persons, or to the roles they take on? Are leadership approaches and formal or informal interactions related to 'cognitive shifts' in the school organization (cf. Foldy et al., 2008)?

Furthermore, our suggestions to undertake discourse analysis to explore whether educational professionals consider reported school performance as a *personal* 'score' or not (cf. Study 3) and to explore where feedback providers' and feedback users' conceptual understandings diverge (cf. Study 1) can be expanded to delve into other relevant (identity-related) questions as well. What do individual frames of reference look like, exactly, i.e., how do educational professionals express their understanding of school performance feedback and their position on the use of this feedback (cf. Jimerson, 2014)? Revisiting the idea of sensegiving: how do school leaders and internal or external information brokers frame school performance feedback (cf. Coburn, 2006)? And where exactly is the supposed language barrier between score report language (i.e., the language of external assessment) and language used by educational professionals' in their daily practice (i.e., the language of learning and instruction) (Hattie, 2009; Horn et al., 2015; Roduta Roberts et al., 2018)?

3.3 Research questions transcending the scope of this dissertation

In the empirical component of this dissertation, we took a focused and narrow look at specific aspects of the sensemaking process that educational professionals go through when engaging with school performance feedback. Further research is needed into other aspects and phases, not in the least to examine how sensemaking relates to prior goal setting on school-level (Schildkamp, 2019). The formulation of goals (for practice, for policy, for data use itself) is an initial step in data-based decision making. One could even say, it is or can be the very first step in the sensemaking process, even before concrete data are accessed or available. Likewise, while we looked at whether reports are put to use (in Study 4 specifically) we only considered whether further steps are taken in the inquiry cycle. Further (longitudinal) research should look at the actual impact of data use (Groß Ophoff et al., 2023; Schildkamp, 2019). Does anything actually change in schools or with regard to student achievement, in the long or short term, after educational professionals have received and discussed school performance feedback?

Additionally, more research is needed with regard to sensemakers' individual frames of reference (interpretive frameworks), how these frames take shape (antecedents) and how they impact the sensemaking process (differential processes and bias). For instance, a comparison that we did not explore in our own empirical research, is that between educational professionals in primary versus secondary education. In Flanders, the majority of educators in primary education and the first stage of secondary education hold a bachelors' degree, while their colleagues in upper secondary school (particularly the general track) tend to enter the profession with a masters' degree. Flemish primary schools are operated by teams of generalists, while teachers in secondary schools have specialized in specific subjects and are organized in departments. Moreover, Flemish primary schools have more standardized instruments at their disposal compared to secondary schools. How do these differing contexts, circumstances and backgrounds relate to the way individual educational professionals make sense of school performance feedback? Also, are there perhaps differences according to the subject matter that was tested in the assessment?

Another (perhaps more profound) question to explore, is how individual educational professionals' expectations about students feed into attributional bias. Teachers have been found to attribute student achievement differently depending on student characteristics, often stable and uncontrollable characteristics such as gender or ethnicity (Wang & Hall, 2018). Continually overemphasizing such characteristics can contribute to inequity and risks reinforcing mental models about what can (and particularly: cannot) be overcome with appropriate instructional responses (Bertrand & Marsh, 2015). Attributional bias, such as repeatedly explaining student performance in a way that corresponds to prior assumptions and expectations, can anchor itself into self-fulfilling prophecies (Wang & Hall, 2018). This could be relevant if we want to understand how educational professionals interpret school performance feedback measures that take into account student population characteristics.

Finally, in Study 1 we dubbed the integrated sensemaking framework we presented, a 'pantry' for further (niche) research on educational professionals' engagement with (formal achievement) data. Indeed, we do contend that all levels that we attempted to unravel, present openings for further inquiry. However, one general note regarding the sensemaking perspective that we worked out in depth, pertains very specifically to the current Flemish educational context. As discussed in the introduction to this dissertation, Flanders is on the brink of implementing mandatory central tests. Sensemaking perspectives have extensively been used to examine how schools enact and react to assessment policies and educational reforms in general (Coburn et al., 2009; Frank et al., 2020; März & Kelchtermans, 2013; Spillane, Diamond, et al., 2002; Spillane, Reiser, et al., 2002; D. H. Sutherland, 2020). It would seem like a missed opportunity not to do this for the 'small revolution' that is unfolding in the Flemish educational landscape as we speak.

3.4 Technological advancements in school performance feedback systems

The school performance feedback system that constituted our research case, makes use of static reports. It is to be expected that some of our findings and conclusions are not readily transferable to systems that communicate digitally, such as systems offering dashboards that allow users to extract information in the formats and degrees of extensiveness they prefer. As examining the usage of feedback dashboards and associated perceptions of users will gain ever more prominence, data use researchers will not only need to rethink some of their research questions, but also need to consider which methodologies are appropriate. We can assume that digital feedback systems (will) make it possible to collect large sets of data relatively easy (albeit that ethical standards need to be considered in this regard). A mixed methods approach (Teddlie & Tashakkori, 2009) that combines datamining with focused qualitative inquiry, makes it possible to combine "high tech" and "human touch" (Schildkamp, 2019, p. 263) and explore educational professionals' sensemaking of data such as school performance feedback to entirely new depths (M. Brown, 2020; Farley-Ripple et al., 2021). For instance, which elements of school performance feedback are consulted most (e.g., criterion-referenced versus norm-referenced, or even self-referenced results), and by whom, and how do these user groups motivate their interest and preferences? Is there a differential impact on resulting decisions?

An additional endeavor could be to enrich this type of inquiry with multimodal data (Giannakos et al., 2019). Psychophysiological methods could be applied to inform feedback designers about the usability of their systems, and give cues about recipients' comprehension of the feedback elements, or even their emotions upon receiving feedback. Methods such as eye-tracking can contribute to the study of users' reading patterns and information processing in digital environments (Eger, 2018; Ha et al., 2015; Lai et al., 2013; Rayner et al., 2006). Eye-tracking, in particular, could provide information about elements that attract the attention of users. After all, a sensemaking perspective posits that mental models (internalized beliefs about causality and about "how things work") direct which data a user notices first (Coburn & Turner, 2011; Spillane & Miele, 2007).

Finally, we have discussed how educational professionals' sensemaking of school performance feedback data is directed by the frames they possess. Supporting them in feedback use, necessitates that we gain insight into individual users' frames and find ways of building these frames if the necessary knowledge or capacity is absent. The implementation of learning analytics into feedback dashboards could offer a wealth of information for feedback providers in this regard, in order to optimize dashboard design and tailored support for users (Greller & Drachsler, 2012; Papamitsiou & Economides, 2014; Verbert et al., 2013). Moving on from the traditional behaviorist, cognitivist and constructivist perspectives on learning and instruction (Greeno, 1997) – perspectives that we have all implicitly adopted to a certain extent in the different studies in this dissertation – perhaps a 'connectivist' paradigm is a next step. Connectivism has been proposed as a learning theory that acknowledges that knowledge resides not only in humans but also in digital sources (Goldie, 2016; Siemens, 2005). This could be an avenue to inspire us when investigating knowledge management and end users' sensemaking of school performance feedback in digital environments.

4 Recommendations for policy and practice

Although a number of issues were certainly raised in this dissertation, and potential pitfalls were pinpointed, uncovering obstacles is a first step towards identifying opportunities. In this final section, we offer a number of recommendations based on our research findings. These are intended as stepping stones towards answering the big questions: How can we optimally equip and support users in processing school performance feedback, and in using that feedback for school improvement? How should school performance feedback systems be designed in order for them to realize their potential? What is a productive discourse in terms of standardized testing and external assessment?

We refrain from tying individual recommendations to specific stakeholders. For school performance feedback systems to realize their 'added value', a shared effort is needed from different parties: educational policymakers who mandate assessments, test developers and feedback providers who design systems, educational professionals who get to work with the data, and (formal and informal) liaisons such as teacher trainers and counsellors. One overarching message that speaks throughout all these recommendations, and that ultimately should be top of mind for all of us, is that "data do not objectively guide decisions on their own—people do" (Spillane, 2012, p. 114).

I. Optimize the design and the delivery of school performance feedback.

Put users first, not the data.

Data use is not, and cannot be, a one-size-fits-all phenomenon. In order to turn data such as school performance feedback into a true tool for school improvement, data providers face the complex but important task to try and play into users' intricate, individual and sometimes idiosyncratic palettes of preferences and preconceptions.

Provide high-quality data and be clear about their added value.

School performance feedback can be a valuable information source, provided that the data are robust and reliable, and the interpretation of these data is solid. The importance of triangulation with other internal and external sources of information should be emphasized as well: school performance feedback can be a piece of the puzzle, not the full picture.

Ensure interpretability by examining actual user interpretations to identify disconnects.

Validity is in the eye of the beholder. Therefore, intended users' frames should be the point of departure. Be aware 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 feedback users.

Facilitate sensemaking by providing cues.

Users are the ones who appraise the favorability of school results, and who can arrive at diagnoses, but they need prompts and keys do so. Ensure that schools know which questions to start from. Ownership of data, and of the data use process, is key.

II. Foster data cultures on system-level and on school-level.

Make expectations explicit.

The purpose of data such as school performance feedback should be clear to the schools and the individual educational professionals that expected to make use of them. If data use is 'part of the work' of educational professionals, this should be communicated unambiguously, and there should be an appropriate framework.

Policy influences discourse, but policy initiatives are rarely implemented linearly in schools. Local actors, especially school leaders, can shape them into daily practices by creating vision, installing routines and implementing a system of knowledge management.

Recognize the value of collective sensemaking.

Raw data rarely mean the same thing to all users, even if they belong to the same organization. People make sense of school performance feedback from their own knowledge and beliefs. Voicing those (to themselves or to others) can cast findings in a new light. Collaboration in data use helps to broaden the frame and prevents certain perspectives from being overlooked. Support and collaboration can take shape in professional learning communities.

Information brokers should focus on developing data use capacity.

III. Rethink data literacy.

Do not stop at the numbers.

Making sense of data such as school performance feedback requires knowledge and skills. Data literacy means understanding the numbers and knowing how to translate them. Training, professional development, and support should attend to analysis, but also to further interpretation and decision making. Start with low hanging fruit, and make data users grow.

References

[*] Note: references marked with an asterisk indicate studies included in the review in Study 1.

Abrams, L. M., Varier, D., & Mehdi, T. (2021). The intersection of school context and teachers' data use practice: Implications for an integrated approach to capacity building. *Studies in Educational Evaluation*, *69*, 100868. https://doi.org/10.1016/j.stueduc.2020.100868

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

Ajzen, I. (2002). *Constructing a TpB questionnaire: Conceptual and methodological considerations*. https://pdfs.semanticscholar.org/6074/b33b529ea56c175095872fa40798f8141867.pdf

Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections. *Psychology & Health*, *26*(9), 1113–1127. https://doi.org/10.1080/08870446.2011.613995

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

Amundsen, C., & Wilson, M. (2012). Are we asking the right questions? A conceptual review of the educational development literature in higher education. *Review of Educational Research*, *82*(1), 90–126. https://doi.org/10.3102/0034654312438409

Anseel, F., & Lievens, F. (2009). The mediating role of feedback acceptance in the relationship between feedback and attitudinal and performance outcomes. *International Journal of Selection and Assessment*, *17*(4), 362–376. https://doi.org/10.1111/j.1468-2389.2009.00479.x

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom videoconferencing for qualitative data collection: Perceptions and experiences of researchers and participants. *International Journal of Qualitative Methods*, *18*, 160940691987459. https://doi.org/10.1177/1609406919874596

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Armitage, C. J., & Conner, M. (2001). Efficacy of the Theory of Planned Behaviour: A meta-analytic review. *British Journal of Social Psychology*, *40*(4), 471–499. https://doi.org/10.1348/014466601164939

Attfield, S., Fields, B., & Baber, C. (2018). A resources model for distributed sensemaking. *Cognition, Technology and Work*, *20*(4), 651–664. https://doi.org/10.1007/s10111-018-0529-4

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman.

Beck, J. S., & Nunnaley, D. (2021). A continuum of data literacy for teaching. *Studies in Educational Evaluation*, *69*, 100871. https://doi.org/10.1016/j.stueduc.2020.100871

[*] Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, *52*(5), 861–893. https://doi.org/10.3102/0002831215599251

Bolhuis, E., Schildkamp, K., & Voogt, J. (2016). Data-based decision making in teams: Enablers and barriers. *Educational Research and Evaluation*, *22*(3–4), 213–233. https://doi.org/10.1080/13803611.2016.1247728

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Breiter, A., & Light, D. (2006). Data for school improvement: Factors for designing effective information systems to support decision-making in schools. *Educational Technology and Society*, *9*(3), 206–217.

Brown, C., Schildkamp, K., & Hubers, M. D. (2017). Combining the best of two worlds: A conceptual proposal for evidence-informed school improvement. *Educational Research*, *59*(2), 154–172. https://doi.org/10.1080/00131881.2017.1304327

Brown, M. (2020). Seeing students at scale: How faculty in large lecture courses act upon learning analytics dashboard data. *Teaching in Higher Education*, *25*(4), 384–400. https://doi.org/10.1080/13562517.2019.1698540

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. Guilford Publications.

Bryk, A. S. (2010). Organizing schools for improvement. *Phi Delta Kappan*, *91*(7), 23–30. https://doi.org/10.1177/003172171009100705

Chick, H., & Pierce, R. (2013). The statistical literacy needed to interpret school assessment data. *Mathematics Teacher Education and Development*, *15*(2).

[*] 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. https://doi.org/10.1177/016146811411600203

[*] Christman, J. B., Ebby, C., & Edmunds, K. (2016). Data use practices for improved mathematics teaching and learning: The importance of productive dissonance and recurring feedback cycles. *Teachers College Record*, *118*(11), 1–32. https://doi.org/10.1177/016146811611801101

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.

Cobb, P., Wood, T., & Yackel, E. (1990). Classrooms as learning environments for teachers and researchers. In R. B. Davis, C. A. Mayer, & N. Noddings (Eds.), *Constructivist views on the teaching and learning of mathematics* (pp. 125–146). National Council of Teachers of Mathematics.

Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, *23*(2), 145–170. https://doi.org/10.3102/01623737023002145

Coburn, C. E. (2006). Framing the problem of reading instruction: Using frame analysis to uncover the microprocesses of policy implementation. *American Educational Research Journal*, *43*(3), 343–349. https://doi.org/10.3102/00028312043003343

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education*, *112*(4), 469–495. https://doi.org/10.1086/505056

[*] Coburn, C. E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *Teachers College Record*, *111*(4), 1115–1161. https://doi.org/10.1177/016146810911100403

[*] Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173–206. https://doi.org/10.1080/15366367.2011.626729

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education*, *118*(2), 99–111. https://doi.org/10.1086/663272

Coe, R., & Visscher, A. J. (2002a). Drawing up the balance sheet for School Performance Feedback Systems. In A. J. Visscher & R. Coe (Eds.), *School Improvement through Performance Feedback* (pp. 221–254). Swets & Zeitinger.

Coe, R., & Visscher, A. J. (2002b). Introduction. In A. J. Visscher & R. Coe (Eds.), *School Improvement through Performance Feedback* (pp. xi–xix). Swets & Zeitinger.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Cook, L., & Gregory, M. (2020). Making sense of sensemaking: Conceptualising how child and family social workers process assessment information. *Child Care in Practice*, *26*(2), 182–195. https://doi.org/10.1080/13575279.2019.1685458

[*] Cosner, S. (2011). Teacher learning, instructional considerations and principal communication: Lessons from a longitudinal study of collaborative data use by teachers. *Educational Management Administration and Leadership*, *39*(5), 568–589. https://doi.org/10.1177/1741143211408453

Daly, A. J. (2012). Data, Dyads, and dynamics: Exploring data use and social networks in educational improvement. *Teachers College Record*, *114*(11), 1–38. https://doi.org/10.1177/016146811211401103

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*(1), 7–28. https://doi.org/10.1007/s10833-015-9264-2

Datnow, A., & Park, V. (2018). Opening or closing doors for students? Equity and data use in schools. *Journal of Educational Change*, *19*(2), 131–152. https://doi.org/10.1007/s10833-018-9323-6

[*] Datnow, A., Park, V., & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. *Journal of Education for Students Placed at Risk (JESPAR)*, *17*(4), 247–265. https://doi.org/10.1080/10824669.2012.718944

Dervin, B. (1983). *An overview of sense-making research: Concepts, methods and results.* Paper Presented at the Annual Meeting of the International Communication Association, Dallas, TX, May.

Dervin, B. (2015). Dervin's Sense-Making Theory. In M. N. Al-Suqri & A. S. Al-Aufi (Eds.), *Information seeking behavior and technology adoption: Theories and trends* (pp. 59–80). IGI Global. https://doi.org/10.4018/978-1-4666-8156-9.ch004

Dierick, S., Laenen, I., Goffin, E., & Vanhoof, J. (2021). Hoe schoolfeedback doen renderen? Schoolfeedback gebruiken als hefboom voor schoolontwikkeling [How to make school feedback work? Using school feedback as a lever for school development]. In M. Van den Brande & W. Smets (Eds.), *Meer weten over (effectief) leren. Leraren als onderzoekers.* (pp. 123–151). Politeia.

Dolle, J., White, M. E., Evans-Santiago, B., Flushman, T., Guise, M., Hegg, S., Myhre, O., Ramirez, E., & Won, N. (2018). *Improvement science in teacher preparation at California State University. How teacher preparation partnerships are building capacity to learn to improve.* SRI International and WestEd.

Dowd, A. C. (2005). *Data don't drive: Building a practitioner-driven culture of inquiry to assess community college performance*. Lumina Foundation for Education Research Report.

Dunn, K. E., Airola, D. T., & Garrison, M. (2013). Concerns, knowledge, and efficacy: An application of the teacher change model to data driven decision-making professional development. *Creative Education*, *04*(10), 673–682. https://doi.org/10.4236/ce.2013.410096

Dunn, K. E., Airola, D. T., Lo, W.-J., & Garrison, M. (2013a). Becoming data driven: The influence of teachers' sense of efficacy on concerns related to data-driven decision making. *The Journal of Experimental Education*, *81*(2), 222–241. https://doi.org/10.1080/00220973.2012.699899

Dunn, K. E., Airola, D. T., Lo, W.-J., & Garrison, M. (2013b). What teachers think about what they can do with data: Development and validation of the data driven decision-making efficacy and anxiety inventory. *Contemporary Educational Psychology*, *38*(1), 87–98. https://doi.org/10.1016/j.cedpsych.2012.11.002

Dunn, K. E., Skutnik, A., Patti, C., & Sohn, B. (2019). Disdain to acceptance: Future teachers' conceptual change related to data-driven decision making. *Action in Teacher Education*, *41*(3), 193–211. https://doi.org/10.1080/01626620.2019.1582116

Earl, L., & Fullan, M. (2003). Using data in leadership for learning. *Cambridge Journal of Education*, *33*(3), 383–394. https://doi.org/10.1080/0305764032000122023

Earl, L., & Katz, S. (2006). *Leading schools in a data-rich world: Harnessing data for school improvement*. Corwin Press.

Education Resources Information Center. (n.d.). *ERIC Thesaurus: Standardized tests*. https://eric.ed.gov/?qt=standardized&ti=Standardized+Tests

Eger, L. (2018). How people acquire knowledge from a web page: An eye tracking study. *Knowledge Management & E-Learning: An International Journal*, *10*(3), 350–366. https://doi.org/10.34105/j.kmel.2018.10.020

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. https://doi.org/10.1111/ldrp.12123

Eurydice. (2023). *National education systems: Belgium - Flemish Community. Overview*. https://eurydice.eacea.ec.europa.eu/national-education-systems/belgium-flemish-community/overview

Evans, M., Teasdale, R. M., Gannon-Slater, N., Londe, P. G. la, Crenshaw, H. L., Greene, J. C., & Schwandt, T. A. (2019). How did that happen? Teachers' explanations for low test scores. *Teachers College Record*, *121*(2), 1–40. https://doi.org/10.1177/016146811912100202

[*] Even, R. (2005). Using assessment to inform instructional decisions: How hard can it be? *Mathematics Education Research Journal*, *17*(3), 45–61. https://doi.org/10.1007/BF03217421

[*] Falabella, A. (2020). The ethics of competition: Accountability policy enactment in Chilean schools' everyday life. *Journal of Education Policy*, *35*(1), 23–45. https://doi.org/10.1080/02680939.2019.1635272

[*] 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. https://doi.org/10.1080/09243453.2020.1777171

[*] Farrell, C. C., & Marsh, J. A. (2016). Metrics matter: How properties and perceptions of data shape teachers' instructional responses. *Educational Administration Quarterly*, *52*(3), 423–462. https://doi.org/10.1177/0013161X16638429

Fitzgerald, M. S., & Palincsar, A. S. (2019). Teaching practices that support student sensemaking across grades and disciplines: A conceptual review. *Review of Research in Education*, *43*(1), 227–248. https://doi.org/10.3102/0091732X18821115

[*] 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. https://doi.org/10.1016/j.stueduc.2020.100846

Foldy, E. G., Goldman, L., & Ospina, S. (2008). Sensegiving and the role of cognitive shifts in the work of leadership. *The Leadership Quarterly*, *19*(5), 514–529. https://doi.org/10.1016/j.leaqua.2008.07.004

Francis, J. J., Eccles, M. P., Johnston, M., Walker, A., Grimshaw, J., Foy, R., Kaner, E. F. S., Smith, L., & Bonetti, D. (2004). *Constructing questionnaires based on the theory of planned behaviour: A manual for health services researchers*. Centre for Health Services Research, University of Newcastle upon Tyne. https://openaccess.city.ac.uk/id/eprint/1735

Frank, K. A., Kim, J., Salloum, S. J., Bieda, K. N., & Youngs, P. (2020). From interpretation to instructional practice: A network study of early-career teachers' sensemaking in the era of accountability pressures and common core state standards. *American Educational Research Journal, 57*(6), 2293–2338. https://doi.org/10.3102/0002831220911065

Gale, N. K., Heath, G., Cameron, E., Rashid, S., & Redwood, S. (2013). Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology*, *13*(1), 117. https://doi.org/10.1186/1471-2288-13-117

Geddes, D., & Linnehan, F. (1996). Exploring the dimensionality of positive and negative performance feedback. *Communication Quarterly*, *44*(3), 326–344. https://doi.org/10.1080/01463379609370021

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, *48*(March), 108–119. https://doi.org/10.1016/j.ijinfomgt.2019.02.003

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. https://doi.org/10.1002/rev3.3334

Goldie, J. G. S. (2016). Connectivism: A knowledge learning theory for the digital age? *Medical Teacher*, *38*(10), 1064–1069. https://doi.org/10.3109/0142159X.2016.1173661

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. https://doi.org/10.1207/s15324818ame1702

Gotch, C. M., & French, B. F. (2013). Elementary teachers' knowledge and self-efficacy for measurement concepts. *The Teacher Educator*, *48*(1), 46–57. https://doi.org/10.1080/08878730.2012.740150

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. https://doi.org/10.1111/emip.12198

Greeno, J. G. (1997). Theories and practices of thinking and learning to think. *American Journal of Education*, *106*(1), 85–126. https://doi.org/10.1086/444177

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, *15*(3), 42–57. https://www.jstor.org/stable/jeductechsoci.15.3.42

Groß Ophoff, J., Brown, C., & Helm, C. (2023). Do pupils at research-informed schools actually perform better? Findings from a study at English schools. *Frontiers in Education, 7*(January), 1–18. https://doi.org/10.3389/feduc.2022.1011241

Groß Ophoff, J., & Egger, C. (2021). Assessment of German and Austrian students' Educational Research Literacy: Validation of a competency test based on cross-national comparisons. *Studia Paedagogica*, *26*(4), 27–45. https://doi.org/10.5817/SP2021-4-2

Groß Ophoff, J., Schladitz, S., Leuders, J., Leuders, T., & Wirtz, M. A. (2015). Assessing the development of Educational Research Literacy: The effect of courses on research methods in studies of Educational Science. *Peabody Journal of Education*, *90*(4), 560–573. https://doi.org/10.1080/0161956X.2015.1068085

Gummer, E. (2021). Complexity and then some: Theories of action and theories of learning in data-informed decision making. *Studies in Educational Evaluation*, *69*, 100960. https://doi.org/10.1016/j.stueduc.2020.100960

Gunnulfsen, A. E. (2017). School leaders' and teachers' work with national test results: Lost in translation? *Journal of Educational Change*, *18*(4), 495–519. https://doi.org/10.1007/s10833-017-9307-y

Gutwirth, G., Goffin, E., & Vanhoof, J. (2021). Sensemaking unraveled: How teachers process school performance feedback data. *Studia Paedagogica*, *26*(4), 67–97. https://doi.org/10.5817/SP2021-4-4

Ha, K., Jo, I.-H., Lim, S., & Park, Y. (2015). Tracking students' eye-movements on visual dashboard presenting their online learning behavior patterns. In G. Chen, V. Kumar, Kinshuk, R. Huang, & S. C. Kong (Eds.), *Emerging Issues in Smart Learning* (pp. 371–376). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-44188-6_51

Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* CSE Technical Report 430.

Hamilton, L. S., Halverson, R., Jackson, S. S., Mandinach, E. B., Supovitz, J. A., Wayman, J. C., Pickens, C., Martin, E. S., & Steele, J. L. (2009). *Using student achievement data to support instructional decision making*. United States Department of Education. https://repository.upenn.edu/gse_pubs/279

Hattie, J. A. C. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*, 1–15. 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. https://doi.org/10.1016/j.edurev.2012.09.001

Hendriks, M. A., Doolaard, S., & Bosker, R. J. (2002). Using school effectiveness as a knowledge base for self-evaluation in Dutch schools: The ZEBO-project. In A. J. Visscher & R. Coe (Eds.), *School Improvement through Performance Feedback* (pp. 115–142). Swets & Zeitinger.

Hoogland, I., Schildkamp, K., van der Kleij, F., Heitink, M., Kippers, W., Veldkamp, B., & Dijkstra, A. M. (2016). Prerequisites for data-based decision making in the classroom: Research evidence and practical illustrations. *Teaching and Teacher Education*, *60*, 377–386. https://doi.org/10.1016/j.tate.2016.07.012

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. https://doi.org/10.1080/0969594X.2018.1447908

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. https://doi.org/10.1016/j.stueduc.2016.11.002

Horn, I. S., Kane, B. D., & Wilson, J. (2015). Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. *American Educational Research Journal*, *52*(2), 208–242. https://doi.org/10.3102/0002831215573773

Horn, I. S., & Little, J. W. (2010). Attending to problems of practice: Routines and resources for professional learning in teachers' workplace interactions. *American Educational Research Journal*, *47*(1), 181–217. https://doi.org/10.3102/0002831209345158

Hoyle, E. (1974). Professionality, professionalism and control in teaching. *London Educational Review*, *3*(2), 13–19.

Hoyle, E. (2008). Changing conceptions of teaching as a profession: Personal reflections. In D. Johnson & R. Maclean (Eds.), *Teaching: Professionalization, Development and Leadership* (pp. 285–304). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8186-6_19

Hulpia, H., & Valcke, M. (2004). The use of performance indicators in a school improvement policy: The theoretical and empirical context. *Evaluation & Research in Education*, *18*(1–2), 102–119. https://doi.org/10.1080/09500790408668311

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: Yearbook of the National Society for the Study of Education* (Vol. 106, Issue 1, pp. 105–131). Wiley-Blackwell. https://doi.org/10.1111/j.1744-7984.2007.00099.x

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371. https://doi.org/10.1037/0021-9010.64.4.349

Janssen, R., Van Nijlen, D., De Mulder, L., & Ameel, E. (2017). *Validering van IDP en de OVSG-toets: Eindrapport*.

Janssens, F. J. G., Rekers-Mombarg, L., & Lacor, E. (2014). *Leerwinst en toegevoegde waarde in het primair onderwijs* [Learning gain and value added in primary education]. https://doi.org/10.13140/2.1.2575.4563

[*] Jennings, J. (2012). The effects of accountability system design on teachers' use of test score data. *Teachers College Record*, *114*(11), 1–23. https://doi.org/10.1177/016146811211401108

Jimerson, J. B. (2014). Thinking about data: Exploring the development of mental models for "data use" among teachers and school leaders. *Studies in Educational Evaluation*, *42*, 5–14. https://doi.org/10.1016/j.stueduc.2013.10.010

Jimerson, J. B., Garry, V., Poortman, C. L., & Schildkamp, K. (2021). Implementation of a collaborative data use model in a United States context. *Studies in Educational Evaluation*, *69*, 100866. https://doi.org/10.1016/j.stueduc.2020.100866

Jimerson, J. B., & Wayman, J. C. (2015). Professional learning for using data: Examining teacher needs & supports. *Teachers College Record*, *117*(4), 1–36. https://doi.org/10.1177/016146811511700405

Jongmans, C. T., & Beijaard, D. (1997). De professionele oriëntatie van leraren en hun betrokkenheid bij het schoolbeleid [Teachers' professional orientation and their involvement in school policy-making]. *Pedagogische Studiën*, *74*, 97–107.

Jongmans, C. T., Sleegers, P. J. C., de Jong, F. P. C. M., & Biemans, H. J. A. (1998). Teachers' professional orientation and their concerns. *Teacher Development*, *2*(3), 465–479. https://doi.org/10.1080/13664539800200060

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526. https://doi.org/10.1037/a0016755

Kane, M. T. (2013a). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Kane, M. T. (2013b). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12001

Kelchtermans, G. (2009). Who I am in how I teach is the message: Self-understanding, vulnerability and reflection. *Teachers and Teaching: Theory and Practice*, *15*(2), 257–272. https://doi.org/10.1080/13540600902875332

Kelchtermans, G. (2011). Professional responsibility. Persistent commitment, perpetual vulnerability? In C. Sugrue & T. Dyrdal Solbrekke (Eds.), *Professional responsibility: New horizons of praxis* (pp. 113–126). Routledge.

Kelchtermans, G. (2018). Professional self-understanding in practice: Narrating, navigating and negotiating. In P. A. Schutz, J. Hong, & D. Cross Francis (Eds.), *Research on teacher identity: Mapping challenges and innovations* (pp. 229–240). Springer International Publishing. https://doi.org/10.1007/978-3-319-93836-3_20

Kennedy, M. M. (2007). Defining a Literature. *Educational Researcher*, *36*(3), 139–147. https://doi.org/10.3102/0013189X07299197

Ketelaar, E., Beijaard, D., Boshuizen, H. P. A., & Den Brok, P. J. (2012). Teachers' positioning towards an educational innovation in the light of ownership, sense-making and agency. *Teaching and Teacher Education*, *28*(2), 273–282. https://doi.org/10.1016/j.tate.2011.10.004

Klein, G., Moon, B., & Hoffman, R. R. (2006a). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems, 21*(4), 70–73. https://doi.org/10.1109/MIS.2006.75

Klein, G., Moon, B. M., & Hoffman, R. R. (2006b). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, *21*(5), 88–92. https://doi.org/10.1109/MIS.2006.100

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). Lawrence Erlbaum Associates.

Klein, G., Wiggins, S., & Dominguez, C. O. (2010). Team sensemaking. *Theoretical Issues in Ergonomics Science*, *11*(4), 304–320. https://doi.org/10.1080/14639221003729177

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

[*] Knight, P., & Yorke, M. (2008). Assessment close up: The limits of exquisite descriptions of achievement. *International Journal of Educational Research*, *47*(3), 175–183. https://doi.org/10.1016/j.ijer.2008.01.005

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., Lee, M.-H., Chiou, G.-L., Liang, J.-C., & Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, *10*(88), 90–115. https://doi.org/10.1016/j.edurev.2013.10.001

[*] Lasater, K., Bengtson, E., & Albiladi, W. S. (2021). Data use for equity? How data practices incite deficit thinking in schools. *Studies in Educational Evaluation*, *69*, 100845. https://doi.org/10.1016/j.stueduc.2020.100845

Lechermeier, J., & Fassnacht, M. (2018). How do performance feedback characteristics influence recipients' reactions? A state-of-the-art review on feedback source, timing, and valence effects. *Management Review Quarterly*, *68*(2), 145–193. https://doi.org/10.1007/s11301-018-0136-8

LeMahieu, P. G., Grunow, A., Baker, L., Nordstrum, L. E., & Gomez, L. M. (2017). Networked improvement communities. *Quality Assurance in Education*, *25*(1), 5–25. https://doi.org/10.1108/QAE-12-2016-0084

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, *5*(39), 1–18. https://doi.org/10.1017/cbo9780511814563.003

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2019). Methodological issues in value-added modeling: An international review from 26 countries. *Educational Assessment, Evaluation and Accountability, 31*(3), 257–287. https://doi.org/10.1007/s11092-019-09303-w

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education*, *118*(2), 143–166. https://doi.org/10.1086/663271

Lockton, M., Weddle, H., & Datnow, A. (2020). When data don't drive: Teacher agency in data use efforts in low-performing schools. *School Effectiveness and School Improvement, 31*(2), 243–265. https://doi.org/10.1080/09243453.2019.1647442

Lo Iacono, V., Symonds, P., & Brown, D. H. K. (2016). Skype as a tool for qualitative research interviews. *Sociological Research Online*, *21*(2), 103–117. https://doi.org/10.5153/sro.3952

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. https://doi.org/10.1111/ijsa.12065

Maier, U. (2010). Accountability policies and teachers' acceptance and usage of school performance feedback - a comparative study. *School Effectiveness and School Improvement*, *21*(2), 145–165. https://doi.org/10.1080/09243450903354913

Maitlis, S., & Christianson, M. (2014). Sensemaking in organizations: Taking stock and moving forward. *The Academy of Management Annals*, *8*(1), 57–125. https://doi.org/10.1080/19416520.2014.873177

Maitlis, S., Vogus, T. J., & Lawrence, T. B. (2013). Sensemaking and emotion in organizations. *Organizational Psychology Review*, *3*(3), 222–247. https://doi.org/10.1177/2041386613489062

Malin, J. R., & Brown, C. (2022). Introduction: What can be learned from international contexts about how to foster evidence-informed practice? In C. Brown & J. R. Malin (Eds.), *The Emerald handbook of evidence-informed practice in education: Learning from international contexts* (pp. 1–13). Emerald Publishing Limited.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, *42*(1), 30–37. https://doi.org/10.3102/0013189X12459803

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. https://doi.org/10.1016/j.tate.2016.07.011

Mandinach, E. B., Gummer, E. S., & Muller, R. D. (2011). *The complexities of integrating data-driven decision making into professional preparation in schools of education: It's harder than you think.* CNA Education, Education Northwest, and WestEd.

Mandinach, E. B., Honey, M., Light, D., & Brunner, C. (2008). A conceptual framework for data-driven decision making. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 13–31). Teachers College Press.

[*] Mandinach, E. B., & Schildkamp, K. (2021a). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation*, *69*, 100842. https://doi.org/10.1016/j.stueduc.2020.100842

Mandinach, E. B., & Schildkamp, K. (2021b). The complexity of data-based decision making: An introduction to the special issue. *Studies in Educational Evaluation*, *69*, 100906. https://doi.org/10.1016/j.stueduc.2020.100906

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*(11), 1–48. https://doi.org/10.1177/016146811211401106

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education. Evidence from recent RAND research* (RAND Occasional Papers). RAND Corporation. https://www.rand.org/pubs/occasional_papers/OP170.html

Marton, F. (1981). Phenomenography — Describing conceptions of the world around us. *Instructional Science, 10*(2), 177–200. https://www.jstor.org/stable/23368358

März, V., & Kelchtermans, G. (2013). Sense-making and structure in teachers' reception of educational reform. A case study on statistics in the mathematics curriculum. *Teaching and Teacher Education*, *29*(1), 13–24. https://doi.org/10.1016/j.tate.2012.08.004

Matteucci, M. C., & Gosling, P. (2004). Italian and French teachers faced with pupil's academic failure: The "norm of effort." *European Journal of Psychology of Education*, *19*(2), 147–166. https://doi.org/10.1007/BF03173229

Matteucci, M. C., & Helker, K. (2018). Who is responsible for educational outcomes? Responsibility ascriptions for educational outcomes in a sample of Italian teachers, parents, and students. *Learning and Individual Differences*, *61*(2018), 239–249. https://doi.org/10.1016/j.lindif.2017.12.009

Mead, G. H. (1934). *Mind, self, and society*. University of Chicago Press.

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

Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). *Implementing data-informed decision making in schools: Teacher access, supports and use*. Report prepared for U.S. Department of Education, Office of Planning, Evaluation and Policy Development. SRI International.

Mesch, D. J., Farh, J.-L., & Podsakoff, P. M. (1994). Effects of feedback sign on group goal setting, strategies, and performance. *Group & Organization Management*, *19*(3), 309–333. https://doi.org/10.1177/1059601194193006

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. https://doi.org/10.1016/j.lcsi.2019.100368

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative Data Analysis. A Methods Sourcebook.* (3rd ed.). Sage Publications.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. https://doi.org/10.1037/h0076486

Mitzel, G. H., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards. Concepts, Methods and Perspectives* (pp. 249–281). Lawrence Erlbaum Associates.

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1), 1–7. https://doi.org/10.1186/s12874-018-0611-x

Nabors Oláh, L., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, *85*(2), 226–245. https://doi.org/10.1080/01619561003688688

Nichols, S. L., & Berliner, D. C. (2007). *Collateral Damage: How High-Stakes Testing Corrupts America's Schools.* Harvard Education Press.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives*, *14*(1), 1. https://doi.org/10.14507/epaa.v14n1.2006

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives*, *20*(20). https://doi.org/10.14507/epaa.v20n20.2012

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment's effects on teachers and schools. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 40–56). Routledge.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.

Nusche, D., Miron, G., Santiago, P., & Teese, R. (2015). *OECD Reviews of School Resources: Flemish Community of Belgium 2015*. OECD Publishing. https://doi.org/10.1787/9789264247598-en

Odden, T. O. B., & Russ, R. S. (2019). Defining sensemaking: Bringing clarity to a fragmented theoretical construct. *Science Education*, *103*(1), 187–205. https://doi.org/10.1002/sce.21452

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. https://doi.org/10.1111/emip.12141

Organisation for Economic Co-operation and Development. (2017). *Education policy outlook: Belgium*. Organisation for Economic Co-operation and Development. www.oecd.org/edu/profiles.htm

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, *17*(4), 49–64. https://www.jstor.org/stable/jeductechsoci.17.4.49

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Parkinson, S., Eatough, V., Holmes, J., Stapley, E., & Midgley, N. (2016). Framework analysis: A worked example of a study exploring young people's experiences of depression. *Qualitative Research in Psychology*, *13*(2), 109–129. https://doi.org/10.1080/14780887.2015.1119228

[*] Park, V., Daly, A. J., & Guerra, A. W. (2013). Strategic framing: How leaders craft the meaning of data use for equity and learning. *Educational Policy*, *27*(4), 645–675. https://doi.org/10.1177/0895904811429295

Patton, M. Q. (2015). *Qualitative Research and Evaluation Methods* (4th ed.). Sage.

Penninckx, M., Vanhoof, J., Quintelier, A., De Maeyer, S., & Van Petegem, P. (2017). *Zicht op leerwinst: scenario's voor gestandaardiseerd toetsen* [Views on learning gains: Scenarios for standardized testing]. Acco.

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 787–850). American Educational Research Association.

Pierce, R., Chick, H., & Gordon, I. (2013). Teachers' perceptions of the factors influencing their engagement with statistical reports on student achievement data. *Australian Journal of Education*, *57*(3), 237–255. https://doi.org/10.1177/0004944113496176

Podsakoff, P. M., & Farh, J. L. (1989). Effects of feedback sign and credibility on goal setting and task performance. *Organizational Behavior and Human Decision Processes*, *44*(1), 45–67. https://doi.org/10.1016/0749-5978(89)90034-4

Prenger, R., & Schildkamp, K. (2018). Data-based decision making for teacher and student learning: A psychological perspective on the role of the teacher. *Educational Psychology*, *38*(6), 734–752. https://doi.org/10.1080/01443410.2018.1426834

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. https://doi.org/10.1007/s11251-021-09535-8

Rankin, J. G. (2016). Data systems and reports as active participants in data interpretation. *Universal Journal of Educational Research*, *4*(11), 2493–2501. https://doi.org/10.13189/ujer.2016.041101

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

Ritchie, J., & Spencer, L. (1994). Qualitative data analysis for applied policy research. In A. Bryman & R. G. Burgess (Eds.), *Analyzing qualitative data* (pp. 173–194). Routledge.

Ritchie, J., Spencer, L., & O'Connor, W. (2003). Carrying out qualitative analysis. In J. Ritchie & J. Lewis (Eds.), *Qualitative Research Practice: A guide for Social Science Students and Researchers* (pp. 219–262). Sage.

Roduta Roberts, M., Gotch, C. M., & Lester, J. N. (2018). Examining score report language in accountability testing. *Frontiers in Education*, *3*(June), 1–17. https://doi.org/10.3389/feduc.2018.00042

Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A Systematic Approach.* (7th ed.). Sage Publications.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28. https://doi.org/10.1037/h0092976

Rowe, K. J., Turner, R., & Lane, K. (2002). Performance feedback to schools of students' year 12 assessments: The VCE Data project. In A. J. Visscher & R. Coe (Eds.), *School Improvement through Performance Feedback* (pp. 163–190). Swets & Zeitinger.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In *Handbook of Test Development* (pp. 677–710). Routledge. https://doi.org/10.4324/9780203874776.ch29

Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers* (2nd ed.). Sage Publications.

Sanbonmatsu, D. M., & Fazio, R. H. (1990). The role of attitudes in memory-based decision making. *Journal of Personality and Social Psychology*, *59*(4), 614–622. https://doi.org/10.1037/0022-3514.59.4.614

Sandberg, J., & Tsoukas, H. (2015). Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. *Journal of Organizational Behavior*, *60*, S6–S32. https://doi.org/10.1002/job

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On Quantitizing. *Journal of Mixed Methods Research*, *3*(3), 208–222. https://doi.org/10.1177/1558689809334210

Savin-Baden, M., & Major, C. H. (2013). *Qualitative research: The essential guide to theory and practice*. Routledge.

[*] Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Schildkamp, K., Karbautzki, L., & Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers. *Studies in Educational Evaluation*, *42*, 15–24. https://doi.org/10.1016/j.stueduc.2013.10.007

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*(3), 482–496. https://doi.org/10.1016/j.tate.2009.06.007

Schildkamp, K., & Lai, M. K. (2013a). Conclusions and a Data Use Framework. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based Decision Making in Education* (pp. 177–191). Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_10

Schildkamp, K., & Lai, M. K. (2013b). Introduction. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based Decision Making in Education* (pp. 1–7). Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_1

Schildkamp, K., & Poortman, C. (2015). Factors influencing the functioning of data teams. *Teachers College Record*, *117*(4), 1–42. https://doi.org/10.1177/016146811511700403

Schildkamp, K., Poortman, C. L., Ebbeler, J., & Pieters, J. M. (2019). How school leaders can build effective data teams: Five building blocks for a new wave of data-informed decision making. *Journal of Educational Change*, *20*(3), 283–325. https://doi.org/10.1007/s10833-019-09345-3

Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2016). Data teams for school improvement. *School Effectiveness and School Improvement*, *27*(2), 228–254. https://doi.org/10.1080/09243453.2015.1056192

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement*, *28*(2), 242–258. https://doi.org/10.1080/09243453.2016.1256901

Schildkamp, K., Rekers-Mombarg, L. T. M., & Harms, T. J. (2012). Student group differences in examination results and utilization for policy and school development. *School Effectiveness and School Improvement*, *23*(2), 229–255. https://doi.org/10.1080/09243453.2011.652123

Schildkamp, K., & Teddlie, C. (2008). School performance feedback systems in the USA and in The Netherlands: A comparison. *Educational Research and Evaluation*, *14*(3), 255–282. https://doi.org/10.1080/13803610802048874

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. https://doi.org/10.1080/01411926.2010.528556

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. https://doi.org/10.3200/JOER.99.6.323-338

[*] Sellar, S. (2015). A feel for numbers: Affect, data and education policy. *Critical Studies in Education*, *56*(1), 131–146. https://doi.org/10.1080/17508487.2015.981198

Sherman, D. K., & Cohen, G. L. (2002). Accepting threatening information: Self–affirmation and the reduction of defensive biases. *Current Directions in Psychological Science*, *11*(4), 119–123. https://doi.org/10.1111/1467-8721.00182

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). https://doi.org/10.3389/feduc.2019.00108

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*, *2*(1), 3–10. http://itdl.org/Journal/Jan_05/article01.htm

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. https://doi.org/10.1207/s15327809jls0302_1

[*] Snodgrass Rangel, V., Bell, E., & Monroy, C. (2019). Teachers' sensemaking and data use implementation in science classrooms. *Education and Urban Society*, *51*(4), 526–554. https://doi.org/10.1177/0013124517727053

[*] Spillane, J. P. (2012). Data in practice: Conceptualizing the data-based decision-making phenomena. *American Journal of Education*, *118*(2), 113–141. https://doi.org/10.1086/663283

Spillane, J. P., Diamond, J. B., Burch, P., Hallett, T., Jita, L., & Zoltners, J. (2002). Managing in the middle: School leaders and the enactment of accountability policy. *Educational Policy*, *16*(5), 731–762. https://doi.org/10.1177/089590402237311

Spillane, J. P., & Miele, D. B. (2007). Evidence in practice: A framing of the terrain. In P. A. Moss (Ed.), *Evidence and decision-making: The 106th yearbook of the National Society for the Study of Education, Part I* (pp. 46–73). Blackwell Publishing. https://doi.org/10.1111/j.1744-7984.2007.00097.x

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, *72*(3), 387–431. https://doi.org/10.3102/00346543072003387

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). JAI Press.

Steunpunt Centrale Toetsen in Onderwijs. (2022). *Toetsen voor onderwijsontwikkeling. Krachtlijnen voor de centrale toetsen in Vlaanderen* [Assessment for educational development. Key points for centralized testing in Flanders]. https://steunpunttoetsen.be/visie/

Steunpunt Toetsontwikkeling en Peilingen. (n.d.-a). *Paralleltoetsen van peilingen* [Parallel tests from national assessments]. www.paralleltoetsen.be

Steunpunt Toetsontwikkeling en Peilingen. (n.d.-b). *STEP*. www.peilingsonderzoek.be

Steunpunt Toetsontwikkeling en Peilingen, & Vlaams Ministerie van Onderwijs en Vorming. (2020). *Peiling mens en maatschappij (domeinen maatschappij, tijd, ruimte en brongebruik) in het basisonderwijs - 2019* [National assessment of People and society (domains: society, time, space and information use) in primary education - 2019]. https://einddoelen.be/peiling-mens-en-maatschappij-2019

Sullivan, J. R. (2012). Skype: An appropriate method of data collection for qualitative interviews? *The Hilltop Review*, *6*(1), 54–60. https://scholarworks.wmich.edu/hilltopreview/vol6/iss1/10

Supovitz, J. A. (2010). Knowledge-based organizational learning for instructional improvement. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *Second International Handbook of Educational Change* (pp. 707–723). Springer.

[*] Sutherland, D. H. (2020). School board sensemaking of federal and state accountability policies. *Educational Policy*, 089590482092581. https://doi.org/10.1177/0895904820925816

Sutherland, S. (2004). Creating a culture of data use for continuous improvement: A case study of an Edison Project school. *American Journal of Evaluation*, *25*(3), 277–293. https://doi.org/10.1016/j.ameval.2004.05.009

Tashakkori, A., & Teddlie, C. (2002). *Handbook of Mixed Methods in Social & Behavioral Research*. Sage Publications.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage Publications.

Uiterwijk-Luijk, L., Krüger, M., Zijlstra, B., & Volman, M. (2017). The relationship between psychological factors and inquiry-based working by primary school teachers. *Educational Studies*, *43*(2), 147–164. https://doi.org/10.1080/03055698.2016.1248901 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. https://doi.org/10.1016/j.stueduc.2013.04.002

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. https://doi.org/10.1016/j.stueduc.2014.04.004

van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 324–343. https://doi.org/10.1080/0969594X.2014.999024

Van-Dijk, D., & Kluger, A. N. (2004). Feedback sign effect on motivation: Is it moderated by regulatory focus? *Applied Psychology*, *53*(1), 113–135. https://doi.org/10.1111/j.1464-0597.2004.00163.x

Van Gasse, R., Goffin, E., Vanhoof, J., & Van Petegem, P. (2021). For squad-members only! Why some teachers are more popular to interact with than others in data use. *Studies in Educational Evaluation*, *69*, 100881. https://doi.org/10.1016/j.stueduc.2020.100881

Van Gasse, R., & Mol, M. (2021). Student guidance decisions at team meetings: Do teachers use data for rational decision making? *Studia Paedagogica*, *26*(4), 99–117. https://doi.org/10.5817/SP2021-4-5

Van Gasse, R., Vanhoof, J., Mahieu, P., & Van Petegem, P. (2015). *Informatiegebruik door schoolleiders en leerkrachten* [Information use by school leaders and teachers]. Garant.

Van Gasse, R., Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2016). Teacher collaboration on the use of pupil learning outcome data: A rich environment for professional learning? *Teaching and Teacher Education*, *60*, 387–397. https://doi.org/10.1016/j.tate.2016.07.004

Van Gasse, R., Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2017). The impact of collaboration on teachers' individual data use. *School Effectiveness and School Improvement*, *28*(3), 489–504. https://doi.org/10.1080/09243453.2017.1321555

Vanhoof, J., & Schildkamp, K. (2014). From "professional development for data use" to "data use for professional development." *Studies in Educational Evaluation*, *42*, 1–4. https://doi.org/10.1016/j.stueduc.2014.05.001

Vanhoof, J., Vanlommel, K., Thijs, S., & Vanderlocht, H. (2014). Data use by Flemish school principals: Impact of attitude, self-efficacy and external expectations. *Educational Studies*, *40*(1), 48–62. https://doi.org/10.1080/03055698.2013.830245

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. https://doi.org/10.1016/j.stueduc.2007.04.001

Vanhoof, J., Verhaeghe, G., Van Petegem, P., & Valcke, M. (2013). Improving data literacy in schools: Lessons from the School Feedback Project. In *Data-based Decision Making in Education* (pp. 113–134). Springer Netherlands. https://doi.org/10.1007/978-94-007-4816-3_7

Vanhoof, J., Verhaeghe, G., Van Petegem, P., & Valcke, M. (2012). Flemish primary teachers' use of school performance feedback and the relationship with school characteristics. *Educational Research*, *54*(4), 431–449. https://doi.org/10.1080/00131881.2012.734726

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. https://doi.org/10.1080/03055698.2010.482771

Vanlommel, K. (2022). Drivers and obstacles for evidence-informed practice in an autonomous and decentralized educational system: Belgium. In C. Brown & J. R. Malin (Eds.), *The Emerald handbook of evidence-informed practice in education: Learning from international contexts* (pp. 259–273). Emerald Publishing Limited.

[*] Vanlommel, K., & Schildkamp, K. (2019). How do teachers make sense of data in the context of high-stakes decision making? *American Educational Research Journal*, *56*(3), 792–821. https://doi.org/10.3102/0002831218803891

Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research*, *83*, 75–83. https://doi.org/10.1016/j.ijer.2017.02.013

[*] Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2021). Sorting pupils into their next educational track: How strongly do teachers rely on data-based or intuitive processes when they make the transition decision? *Studies in Educational Evaluation*, *69*, 100865. https://doi.org/10.1016/j.stueduc.2020.100865

Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2016). Data use by teachers: the impact of motivation, decision-making style, supportive relationships and reflective capacity. *Educational Studies*, *42*(1), 36–53. https://doi.org/10.1080/03055698.2016.1148582

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500–1509. https://doi.org/10.1177/0002764213479363

Verhaeghe, G. (2011). *School performance feedback systems: Design and implementation issues* (Doctoral dissertation, Ghent University). http://users.ugent.be/~mvalcke/CV/PhD Goedele Verhaeghe.pdf

Verhaeghe, G., Schildkamp, K., Luyten, H., & Valcke, M. (2015). Diversity in school performance feedback systems. *School Effectiveness and School Improvement*, *26*(4), 612–638. https://doi.org/10.1080/09243453.2015.1017506

Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: perceptions of primary school principals. *School Effectiveness and School Improvement*, *21*(2), 167–188. https://doi.org/10.1080/09243450903396005

Visscher, A. J. (2002). A framework for studying School Performance Feedback Systems. In A. J. Visscher & R. Coe (Eds.), *School Improvement through Performance Feedback* (pp. 41–72). Swets & Zeitinger.

Visscher, A. J. (2021). On the value of data-based decision making in education: The evidence from six intervention studies. *Studies in Educational Evaluation*, *69*, 100899. https://doi.org/10.1016/j.stueduc.2020.100899

Visscher, A. J., & Coe, R. (2003). School Performance Feedback Systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349. https://doi.org/10.1076/sesi.14.3.321.15842

Vlaams Ministerie van Onderwijs en Vorming. (n.d.-a). *Internationaal vergelijkend onderzoek* [International comparative research]. https://www.onderwijs.vlaanderen.be/nl/onderzoek/vlaams-en-internationaal-onderwijsonderzoek/internationaal-vergelijkend-onderzoek

Vlaams Ministerie van Onderwijs en Vorming. (n.d.-b). *Onderwijsdoelen* [Educational goals]. https://onderwijsdoelen.be/

Vlaams Ministerie van Onderwijs en Vorming. (n.d.-c). *Vlaamse toetsen* [Flemish tests]. https://www.vlaanderen.be/onderwijs-en-vorming/vlaamse-toetsen

Vlaams Ministerie van Onderwijs en Vorming. (n.d.-d). *Zo word je leraar* [This is how you become a teacher]. https://www.vlaanderen.be/lesgeven-is-alles-geven/zo-word-je-leraar

Vlaams Ministerie van Onderwijs en Vorming. (2008). *Education in Flanders. The Flemish educational landscape in a nutshell 2008*. Vlaams Ministerie van Onderwijs en Vorming.

Vlaams Ministerie van Onderwijs en Vorming. (2017). *Gebruik van gevalideerde toetsen voor interne kwaliteitszorg in het gewoon lager onderwijs. Omzendbrief BaO/2017/02* [The use of validated assessments for internal quality assurance in primary education. Circular]. https://data-onderwijs.vlaanderen.be/edulex/document.aspx?docid=15086

Vlaams Ministerie van Onderwijs en Vorming. (2023). *Ontwerp van decreet over de Vlaamse toetsen in het onderwijs* [Draft decree on Flemish tests in education] (VR 2023 1002 DOC.0133/2QUATER). Vlaamse Regering. https://beslissingenvlaamseregering.vlaanderen.be/document-view/63E623F32E929B312AB5D36C

Vlaams Ministerie van Onderwijs en Vorming, & Onderwijsinspectie. (2016a). *Referentiekader voor OnderwijsKwaliteit (OK)*. Vlaams Ministerie van Onderwijs en Vorming - Onderwijsinspectie. http://mijnschoolisok.be/

Vlaams Ministerie van Onderwijs en Vorming, & Onderwijsinspectie. (2016b). *Referentiekader voor OnderwijsKwaliteit (OK). Bronnendocument*. Vlaams Ministerie van Onderwijs en Vorming - Onderwijsinspectie. http://mijnschoolisok.be/

Vlaams Ministerie van Onderwijs en Vorming, & Onderwijsinspectie. (2019). *Onderwijsspiegel. Jaarlijks rapport van de onderwijsinspectie. Editie 2019*. https://www.onderwijsinspectie.be/nl/andere-opdrachten/andere/jaarverslag-onderwijsspiegel

Vlaams Ministerie van Onderwijs en Vorming, & Onderwijsinspectie. (2020). *Onderwijsspiegel. Jaarlijks rapport van de onderwijsinspectie. Editie 2020*. https://www.onderwijsinspectie.be/nl/andere-opdrachten/andere/jaarverslag-onderwijsspiegel

Vlaams Ministerie van Onderwijs en Vorming, & Onderwijsinspectie. (2021). *Onderwijsspiegel. Jaarlijks rapport van de onderwijsinspectie. Editie 2021*. https://www.onderwijsinspectie.be/nl/andere-opdrachten/andere/jaarverslag-onderwijsspiegel

Vlaams Ministerie van Onderwijs en Vorming, & Onderwijsinspectie. (2022). *Onderwijsspiegel. Jaarlijks rapport van de onderwijsinspectie. Editie 2022*. https://www.onderwijsinspectie.be/nl/andere-opdrachten/andere/jaarverslag-onderwijsspiegel

Walls, J. (2017). Sensemaking and school failure: Lessons from two cases. *Journal of Organizational Theory in Education*, *2*(1), 1–26.

Wang, H., & Hall, N. C. (2018). A systematic review of teachers' causal attributions: Prevalence, correlates, and consequences. *Frontiers in Psychology*, *9*(DEC), 1–22. https://doi.org/10.3389/fpsyg.2018.02305

[*] Wardrip, P. S., & Herman, P. (2018). 'We're keeping on top of the students': Making sense of test data with more informal data in a grade-level instructional team. *Teacher Development*, *22*(1), 31–50. https://doi.org/10.1080/13664530.2017.1308428

Wayman, J. C., Cho, V., Mandinach, E. B., Supovitz, J. A., & Wilkerson, S. B. (2017). *Teacher Data Use Survey : Teacher Version*. REL Appalachia and the Institute of Educational Sciences, US Department of Education.

Weick, K. E. (1995). *Sensemaking in Organizations*. Sage Publications.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

Weiner, B. (2000). Intrapersonal and Interpersonal Theories of Motivation from an Attributional Perspective. *Educational Psychology Review*, *12*(1), 1–14. https://doi.org/10.1023/A:1009017532121

Weiner, B. (2010). The development of an attribution-based theory of motivation: A history of ideas. *Educational Psychologist*, *45*(1), 28–36. https://doi.org/10.1080/00461520903433596

Wu, J., & Lederer, A. (2009). A meta-analysis of the role of environment-based voluntariness in information technology acceptance. *MIS Quarterly*, *33*(2), 419. https://doi.org/10.2307/20650298

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. https://doi.org/10.1080/0969594X.2014.936357

Zapata-Rivera, D., Vezzu, M., & VanWinkle, W. (2013). *Exploring Teachers' Understanding of Graphical Representations of Group Performance* (RM-13-04; ETS Research Memorandum).

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. https://doi.org/10.1080/10627197.2016.1202110

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. https://doi.org/10.1080/08957340903221667

Appendices

Appendix A -
Attainment targets Spatial use, Traffic and Mobility

Attainment targets tested for *Spatial use, Traffic and Mobility* in the 2019 Flemish national assessment of *People and Society* in the sixth grade of primary school:

Number	Attainment target
4.9	Pupils can recognize and distinguish rural, urban, tourist and industrial environments in reality on an appropriate map.
4.10	Pupils can situate their own region and two other regions in Belgium on a map and describe the relationship between the environment and aspects of people's daily lives.
4.11	Pupils can compare aspects of the daily life in a country from another cultural region with their own life.
4.12	Pupils can make focused observations in a landscape and they can investigate in a simple way why it looks the way it does.
4.17	Pupils know the main consequences of increasing car use and can compare the advantages and disadvantages of potential alternatives.
4.18	Pupils can plan a simple route using public transportation.

Notes.
Translated from Dutch.
For this domain, the attainment targets were reached by 71% of the students in the sample, which marked a decline compared to an earlier national assessment of the same domains (Steunpunt Toetsontwikkeling en Peilingen & Vlaams Ministerie van Onderwijs en Vorming, 2020).

Appendix B -
SPF report elements with annotation (Study 2)

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.

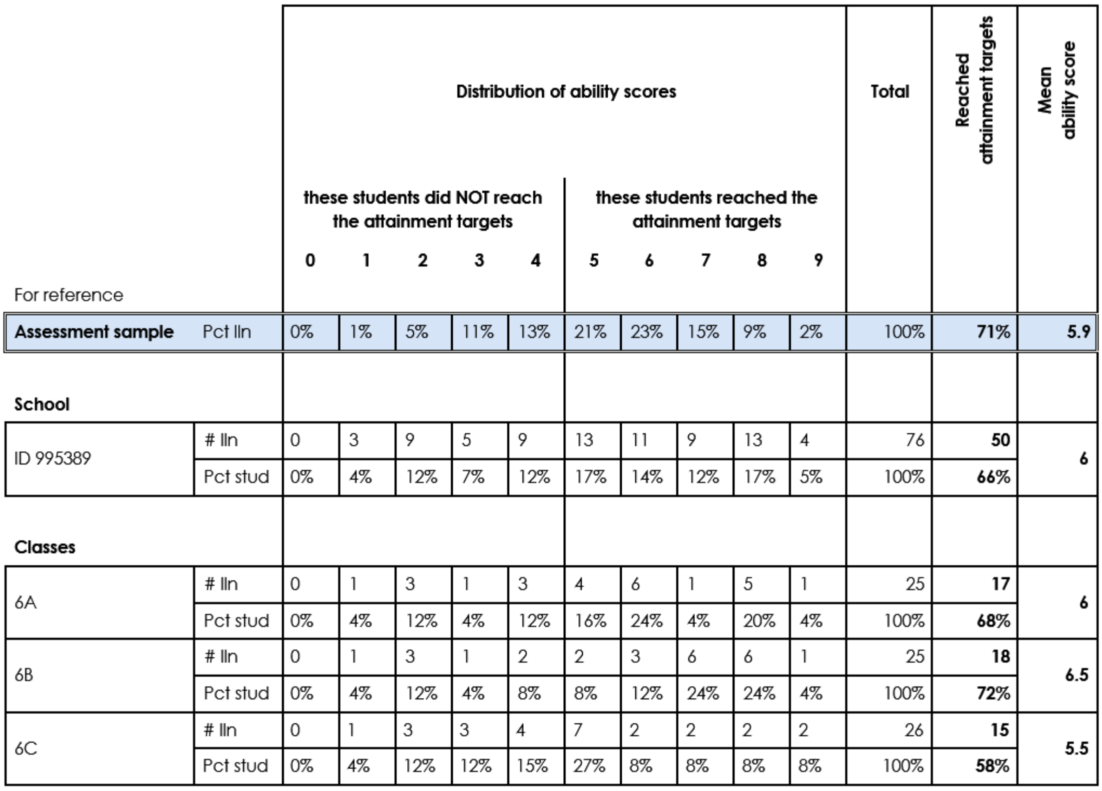
Table expressing student achievement in terms of reaching the attainment targets

The table pictured in Figure 7 gives information about the extent to which the 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".

Figure 7. Report table: Reaching the attainment targets



		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%**	

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 12, 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 12. 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

Table 12 (Continued)

Component	Dimension [a]	Examples
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

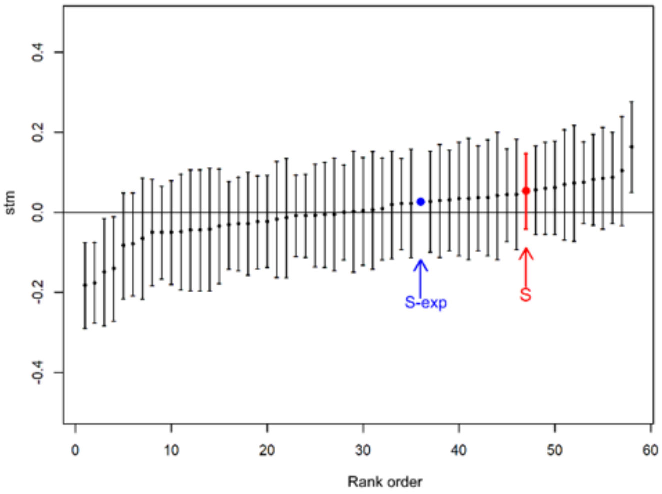
Two caterpillar plots **position the school's performance** compared to other Flemish schools' performance in the NA.

The first plot (see Figure 8) 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 9) 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.

Figure 8. Caterpillar plot positioning the school's actual and expected score

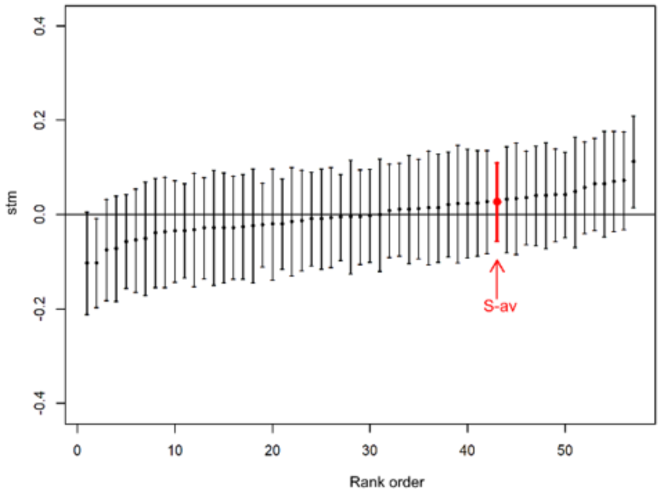


The actual average of school ID 995389 for the test Spatial use, traffic and mobility:

- 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 9. Caterpillar plot positioning the school's value-added



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

Table 13 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 12, this is not an exhaustive overview nor is the overview intended to indicate the prevalence of specific issues.

Table 13. 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

Table 13 (Continued)

Component	Dimension [a]	Examples
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.

Appendix C - Regression parameters of the path model (Study 4)

Regression	B [a]	SE [b]	Z [c]	β [d]	p [e]	Sig [f]	R² [g]
SPF use ~							.264
Cognitive attitude: Usefulness for school development	0.173	0.047	3.656	0.175	<.001	***	
Cognitive attitude: Usefulness for accountability	-0.047	0.047	-1.009	-0.046	.313	ns	
Affective attitude	0.047	0.046	1.034	0.048	.301	ns	
Subjective norm	0.136	0.048	2.839	0.132	.005	**	
Self-efficacy	0.038	0.047	0.827	0.039	.408	ns	
Shared goals	0.305	0.050	6.035	0.312	<.001	***	
Voluntary v. non-voluntary	0.253	0.087	2.906	0.127	.004	**	
Affective attitude ~							.125
Shared goals	0.211	0.052	4.077	0.210	<.001	***	
Support and collaboration	-0.005	0.049	-0.109	-0.006	.913	ns	
Coordinator v. teacher	0.411	0.088	4.699	0.221	<.001	***	
Criterion-referenced result	0.027	0.056	0.482	0.029	.630	ns	
Norm-referenced result	0.087	0.056	1.566	0.095	.117	ns	
Voluntary v. non-voluntary	0.025	0.097	0.256	0.012	.798	ns	
Cognitive attitude: Usefulness for school development ~							.100
Shared goals	0.233	0.051	4.523	0.237	<.001	***	
Support and collaboration	0.056	0.048	1.161	0.059	.245	ns	
Coordinator v. teacher	0.234	0.087	2.689	0.129	.007	**	
Criterion-referenced result	0.008	0.055	0.141	0.009	.888	ns	
Norm-referenced result	-0.007	0.055	-0.122	-0.007	.903	ns	
Voluntary v. non-voluntary	0.101	0.096	1.053	0.050	.293	ns	
Cognitive attitude: Usefulness for accountability ~							.043
Shared goals	0.095	0.051	1.846	0.100	.065	ns	
Support and collaboration	0.016	0.048	0.325	0.017	.745	ns	
Coordinator v. teacher	-0.158	0.087	-1.817	-0.090	.069	ns	
Criterion-referenced result	-0.067	0.055	-1.214	-0.076	.225	ns	
Norm-referenced result	0.150	0.055	2.717	0.172	.007	**	
Voluntary v. non-voluntary	-0.146	0.096	-1.516	-0.075	.129	ns	
Subjective norm ~							.150
Shared goals	0.344	0.048	7.107	0.360	<.001	***	
Support and collaboration	0.058	0.045	1.268	0.063	.205	ns	
Coordinator v. teacher	-0.054	0.082	-0.653	-0.030	.514	ns	
Criterion-referenced result	-0.075	0.052	-1.434	-0.085	.152	ns	
Norm-referenced result	0.039	0.053	0.748	0.045	.454	ns	
Voluntary v. non-voluntary	-0.059	0.090	-0.649	-0.030	.516	ns	
Self-efficacy ~							.175
Shared goals	0.299	0.050	5.978	0.299	<.001	***	
Support and collaboration	0.140	0.047	2.992	0.146	.003	**	
Coordinator v. teacher	-0.073	0.085	-0.862	-0.039	.389	ns	
Criterion-referenced result	-0.019	0.054	-0.357	-0.021	.721	ns	
Norm-referenced result	0.128	0.054	2.372	0.139	.018	*	
Voluntary v. non-voluntary	0.013	0.093	0.141	0.006	.888	ns	

Note. [a] unstandardized coefficient; [b] standard error; [c] z-value; [d] standardized coefficient; [e] p-value; [f] significance; [g] explained variance.

ns *p* > .05, * *p* < .05, ** *p* < .01, *** *p* < .001

Appendix D - Covariance parameters of the path model (Study 4)

Covariance	B [a]	SE [b]	Z [c]	β [d]	p [e]	Sig [f]
.Cognitive attitude: Usefulness for school development ~~ .Cognitive attitude: Usefulness for accountability	0.250	0.039	6.446	0.337	<.001	***
.Cognitive attitude: Usefulness for school development ~~ .Affective attitude	0.117	0.037	3.168	0.156	.002	**
.Cognitive attitude: Usefulness for accountability ~~ .Affective attitude	0.110	0.037	3.000	0.147	.003	**
.Affective attitude ~~ .Self-efficacy	0.168	0.036	4.613	0.231	<.001	***
.Self-efficacy ~~ .Subjective norm	-0.078	0.033	-2.326	-0.114	.020	*

Note. [a] unstandardized coefficient; [b] standard error; [c] z-value; [d] standardized coefficient; [e] p-value; [f] significance.
* p < .05, ** p < .01, *** p < .001

Author contributions

Evelyn Goffin, Jan Vanhoof and Rianne Janssen met regularly to discuss research ideas. All contributed to the design and conception of the different studies.

Data collection was conducted by **Evelyn Goffin**. Data analyses were performed by **Evelyn Goffin** in close consultation with **Jan Vanhoof and Rianne Janssen**.

All manuscripts were written and revised by **Evelyn Goffin**. **Jan Vanhoof and Rianne Janssen** reviewed and commented on successive versions of the manuscripts several times in the course of this process. **Evelyn Goffin, Jan Vanhoof and Rianne Janssen** read and approved the final versions.

Universiteit
Antwerpen

FACULTY OF SOCIAL SCIENCES
Training and Education Sciences
Supervisor: prof. dr. Jan Vanhoof

KU LEUVEN

FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES
Educational Effectiveness and Evaluation
Supervisor: prof. dr. Rianne Janssen

STEP
STEUNPUNT
TOETSONTWIKKELING
EN PEILINGEN

STEUNPUNT TOETSONTWIKKELING EN PEILINGEN (STEP)
Policy research center for test development and assessments

2023