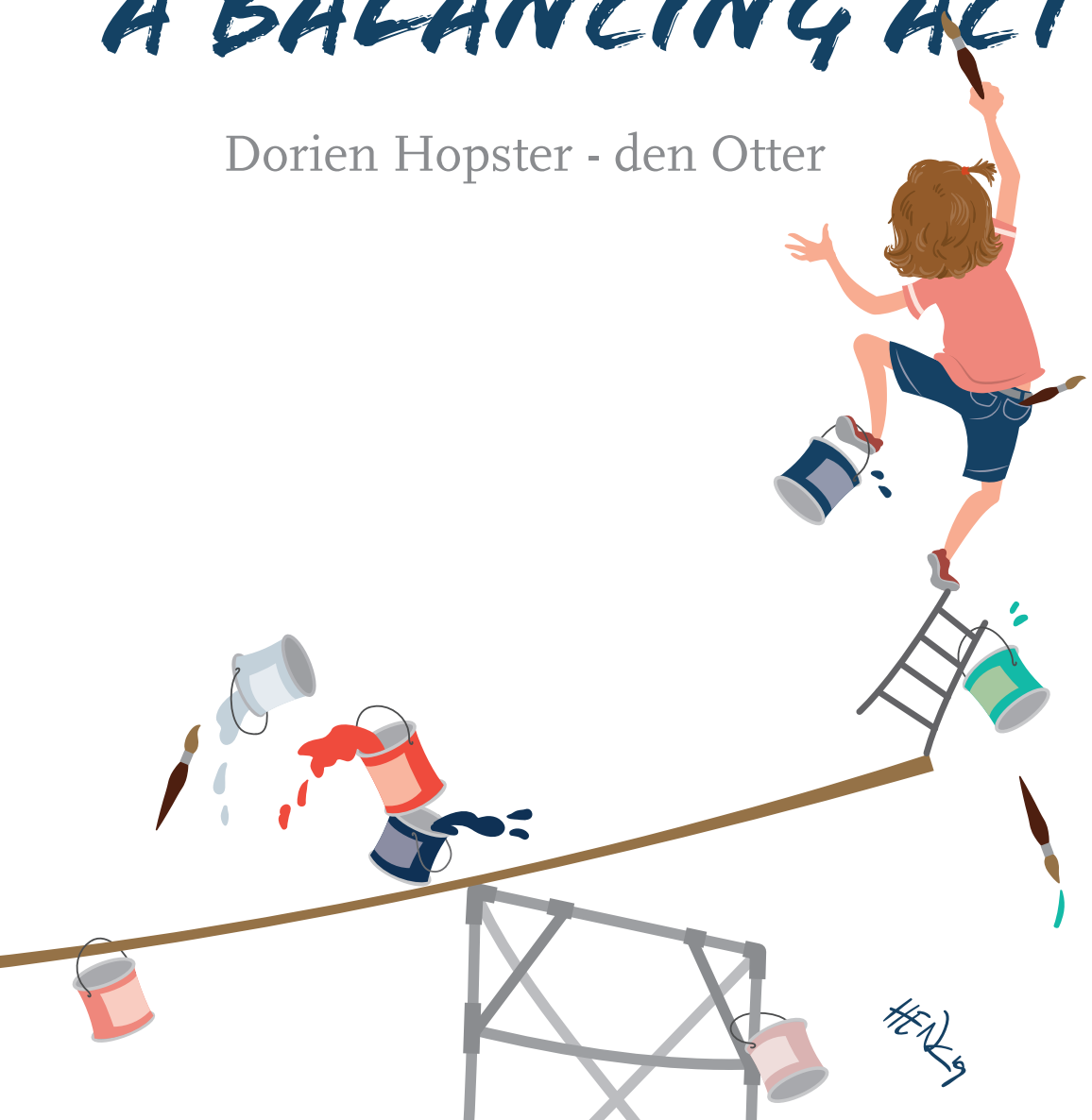


# FORMATIVE ASSESSMENT DESIGN: A BALANCING ACT

Dorien Hopster - den Otter



HENK



FORMATIVE ASSESSMENT DESIGN:  
A BALANCING ACT

Dorien Hopster - den Otter



# FORMATIVE ASSESSMENT DESIGN: A BALANCING ACT

DISSERTATION

To obtain  
the degree of doctor at the University of Twente,  
on the authority of the rector magnificus  
prof. dr. T. T. M. Palstra,  
on account of the decision of the graduation committee,  
to be publicly defended  
on Thursday, November 28, 2019 at 10.45 hours

by  
Dorothea den Otter

Born on September 19, 1991  
in Hengelo, the Netherlands

This dissertation has been approved by:

Promotor: Prof. dr. ir. T. J. H. M. Eggen  
Promotor: Prof. dr. ir. B. P. Veldkamp  
Assistant Promotor: Dr. S. Wools

**UNIVERSITY  
OF TWENTE.**

Doctoral dissertation, University of  
Twente



Supported by  
Cito, National Institute for  
Educational Measurement

**ico**

In the context of the research school  
Interuniversity Center for Educational  
Research

Design cover and chapter pages: Henk van den Heuvel, hillz.nl

Printed by: Ipskamp printing - Enschede, The Netherlands

ISBN: 978-90-365-4823-6

DOI: 10.3990/1.9789036548236

Copyright © 2019, Enschede, The Netherlands. All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

## GRADUATION COMMITTEE

Chairman: Prof. dr. T. A. J. Toonen

Promoters: Prof. dr. ir. T. J. H. M. Eggen  
Prof. dr. ir. B. P. Veldkamp

Assistant promotor: Dr. S. Wools

Members:	Prof. dr. E. J. P. G. Denessen	Leiden University
	Dr. D. Joosten - ten Brinke	Open University
	Dr. J. W. Luyten	University of Twente
	Prof. dr. S. E. McKenney	University of Twente
	Prof. dr. P. C. J. Segers	Radboud University
	Prof. dr. K. Sijtsma	Tilburg University





## TABLE OF CONTENTS

Chapter 1	
Introduction	9
Chapter 2	
A General Framework for the Validation of Embedded Formative Assessment	21
Chapter 3	
Formative Use of Test Results: A User's Perspective	47
Chapter 4	
The Visual Presentation of Measurement Error	85
Chapter 5	
The Usability of an Embedded Formative Assessment System	113
Chapter 6	
The Visual Presentation of a Learning Trajectory	141
Chapter 7	
Conclusion and Discussion	175
Summary	185
Samenvatting	193
Dankwoord	201
Publications and Presentations	207
ICO Dissertation Series	211





# CHAPTER 1



## INTRODUCTION

Educational decision-making has many implications for students' development and educational career. Decisions may pertain to the kind of instructional support a student should receive or the learning objectives a student should achieve. Decisions can also relate to whether students have met certain standards or whether they should be accepted into a certain study program.

Tests and assessments are intended to support these educational decisions, and their purpose is to collect and provide information about students' knowledge, skills, learning strategies, and/or misconceptions. The intention is that the use of this information will result in decisions that are better, or better founded, than the decisions that would have been taken intuitively in its absence (Black & Wiliam, 2009).

This potential support will benefit from an assessment instrument designed in coherence with its intended use. This means that the instrument should directly inform educators about their decisions in a understandable way so that they may act reasonably on that information (Tannenbaum, 2019; Zapata-Rivera & Katz, 2014).

Too little attention has been devoted to the issue of intended use. For example, assessment developers have mainly attended to measurement concerns, such as sampling of tasks, scoring rules, and cut scores (Katz, 2018). The development of score reports has generally been perceived as an afterthought, although it is the bridge between the information captured by the assessment and the decisions or actions of educators (Tannenbaum, 2019). The assessment literature has also focused merely on the psychometric aspects of assessment instruments.

Studies regarding score report design have been limited to guidelines about making assessment results accessible to non-technical audiences (e.g., Deng & Yoo, 2009; Hambleton & Zenisky, 2013; Wainer, Hambleton, & Meara, 1999); however, actual use may be influenced by many more user characteristics.

This limited attention has resulted in many difficulties around the understanding and use of assessment results (e.g., Hellrung & Hartig, 2013; Popham, 2009; Van der Kleij & Eggen, 2013). Research shows that most educators do not use assessment results properly or do not use these results at all (Schildkamp & Teddlie, 2008; Vanlommel, Van Gasse, Vanhoof, & Van Petegem, 2017). In particular, the concept of measurement error causes many difficulties (Zwick, Zapata-Rivera, & Hegarty, 2014).

These difficulties threaten the validity of the interpretation and use of the assessment. Validity is one of the most important quality aspects of assessments (AERA, APA, & NCME, 2014) and is often defined as the extent to which an assessment result is appropriate for its intended interpretation and use (Kane, 2013). This definition shows that understanding and use are part and parcel of the overall argument supporting assessment validity (Kane, 2016). Quite simply, if the assessment results are not understandable and useful for the intended audience, all other extensive efforts to ensure validity will be in vain (Hambleton & Zenisky, 2013; Tannenbaum, 2019).

Therefore, the intended use should be of central concern in the development and evaluation of assessments (Kane, 2013; Tannenbaum, 2019). Assessment developers have the responsibility to ensure that the assessment instrument supports the understanding and use of the intended audience (AERA et al., 2014; Zapata-Rivera & Katz, 2014). This implies that the intended use will be the starting point in the development process and that it will inform the entire design of the instrument: from assessment tasks to score reports.

The current dissertation aims to investigate the design of assessment instruments that support the crucial aspect of intended use. The focus is on formative assessment because a correct understanding and use of assessment results by the intended audience is -more so than for summative assessment- critical to its effectiveness (e.g., Bennett, 2011; Gearhart et al., 2006; Maciver, Anderson, Costa, & Evers, 2014). The concept of formative assessment is introduced in the next section.

## 1.1 A DEFINITION OF FORMATIVE ASSESSMENT

Formative assessment has been the subject of increasing amounts of attention in education, yet a uniform definition remains wanting. Without a clear understanding of what is being studied, the design, implementation, and evaluation would be difficult (Bennett, 2011; Dunn & Mulvenon, 2009). Therefore, this section begins by discussing various distinctions in the conceptualization of formative assessment that are relevant to this dissertation, with the aim of proposing a definition of the concept.

Formative assessment is often distinguished from summative assessment. It is characterized by its purpose in supporting student learning, while summative assessment is intended to provide a final decision about students' learning, for example, for selection, certification, or accountability purposes (Shavelson, 2003; Trumbull & Lash, 2013). In addition, the concept of formative assessment is used interchangeably with several other concepts in the literature, such as assessment for learning, diagnostic assessment, and data-based decision-making (Antoniou & James, 2014; Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). While these concepts reflect different learning theories or assessment paradigms (Van der Kleij et al., 2015), they all have in common that assessment results are used for steering students' learning.

Another distinction is the conceptualization of formative assessment as an instrument or process. Some authors perceive formative assessment as an instrument that provides information about students' learning. For example, Kahl (2005, p.11) defined formative assessment as "a tool that teachers use to measure student grasp of specific topics and skills they are teaching". Others emphasize the process of using this feedback, such as Clarck (2012, p. 217): "Formative assessment is not a test or a tool (a more fine-grained test) but a *process* with the potential to support learning... [italics in original]." Bennett (2011) perceives each position as an oversimplification. Even the most carefully designed instrument is unlikely to support student learning if the process surrounding its use is weakened. Similarly, the process cannot be fulfilled if the instrument is inappropriate for its intended purpose.

Formative assessment results can be used by various audiences. Traditionally, teachers have been regarded as responsible for interpreting and using assessment results to make decisions about subsequent instructional actions. In addition, students are deemed responsible for their own learning. The

belief is that they can assess themselves or their peers and suggest modifications to subsequent learning. Furthermore, parents are interested in understanding their child's achievement and what they can do to support their child to improve future performance. Moreover, school leaders can use assessment results to identify areas of need and support the teaching and learning process in the school (Black & Wiliam, 2009; Falk, 2012; Kannan, Zapata-Rivera, & Leibowitz, 2018; Schildkamp & Kuiper, 2010). According to Zapata-Rivera and Katz (2014), each audience has its own characteristics and unique type of decisions to be made. Variations among audiences point to the need to design instruments that are tailored to a target group.

In relation to the various audiences, formative assessment can be performed at different levels of education. For example, teachers would be more involved in the decision process at the level of the individual student or class, while school leaders would be more focused on the decision process at the level of the school (Brookhart & Nitko, 2008; Schildkamp & Kuiper, 2010). To distinguish between these levels, the term "formative assessment" is often used for decisions at an individual or class level, while the term "formative evaluation" refers to decisions at higher aggregated levels than the individual and class (Harlen, 2007; Van der Kleij et al., 2015).

A final conceptual distinction is that formative assessment can take multiple modes. Shavelson et al. (2008) distinguished three categories on a continuum from informal to formal: "on-the-fly," "planned-for-interaction," and "curriculum-embedded" assessments. On-the-fly formative assessments occur unexpectedly as part of a classroom activity, for example, the student or teacher seeks, reflects upon, and responds to information from dialogue that challenges the student to the next level. Planned-for-interaction assessments occur deliberately, for example, when a teacher intends to find the gap between what students know and what they need to know. Curriculum-embedded assessments are the most formal assessments and consist of predefined tasks. They are built into the educational program where an important learning objective should have been reached before students go on to the next lesson. Insights into students' current learning could be used by teachers or students for decisions about subsequent actions.

The focus of this dissertation is embedded formative assessment, as this formal category is often developed outside the school, thereby increasing the distance to educational practice. Thus, this form has the greatest challenge in terms of alignment with educators' understanding and use. Bennet's (2011)



reasoning is followed and formative assessment is defined as both an instrument and a process, whereby data from an instrument are purposefully gathered, understood, and used for decisions about actions to support student learning. In acknowledging the role of various audiences in formative assessment, it is being investigated how formative assessment instruments might serve as tools to inform teachers, as they are the key drivers in supporting or hindering students' learning. Since teachers mainly use assessment results at an individual and group level, the term "assessment" is used in this dissertation.

## 1.2 SUPPORTING INTENDED USE

A formative assessment instrument can support its intended use in two ways. First, the content of the results has to fit the information needs of teachers (William, 2011; Zapata-Rivera & Katz, 2014). This means that the assessment information guides teachers toward actions that they should take to enhance the teaching and learning process. Understanding teachers' information needs might help assessment developers in presenting the correct type of information.

Second, the assessment results has to be clearly presented to teachers (Hambleton & Zenisky, 2013; Hegarty, 2019) so that teachers can understand and use them correctly. Investigating teachers' knowledge and understanding of visual representations might support assessment developers in visualizing the information in an appropriate way.

The current dissertation aims to investigate whether a content and visual presentation of a formative assessment instrument could support its intended use. The central question is: *What characteristics of a formative assessment instrument support teachers' understanding and use?* The study is performed within the context of primary education.

### 1.3 OUTLINE

The central question of this dissertation is addressed in five studies. Chapter 2 starts with a theoretical study in which a general framework for the validation of formative assessments is provided. Moreover, it describes the concept of formative assessment and argues why a proper understanding and use of assessment results is central to the concept of validity.

Chapter 3 focuses on the content of a formative assessment instrument. The study performed a needs assessment, which investigated the type of instructional actions as well as the information needs to enable these actions. In addition, the study investigated the differences between several users. For the purpose of this study, data were gathered from questionnaires and focus groups.

Chapter 4 highlights the visual presentation of the instrument. More specifically, the study investigated the extent to which presentations of measurement error in score reports influence teachers' instructional decisions and preferences in relation to these presentations. The data were collected from a factorial survey, think-aloud protocols, and focus groups.

Chapter 5 continues with an investigation of the characteristics of a formative assessment instrument by evaluating the usability of a formative assessment platform. The platform was tried out in a natural classroom setting for three months. During this period, data were collected from log files, questionnaires, and interviews, and the findings resulted in design principles regarding the design of the formative assessment instruments.

Chapter 6 is an in-depth study on one of these design principles, indicating that teachers need a clear visualization of the learning trajectory. The study explores how to visualize a learning trajectory that reflects the underlying data structure and that can be used for the purpose of formative assessment.

The dissertation ends with a general synthesis of the findings of previous studies in relation to the central question examined herein. The characteristics of formative assessment instruments are described in relation to their visual presentation and content.

## 1.4 REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability*, 26, 153–176. doi:10.1007/s11092-013-9188-4
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. doi:10.1080/0969594X.2010.513678
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5–31. doi:10.1007/s11092-008-9068-5
- Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, NJ: Pearson Education.
- Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24(2), 205–249. doi:10.1007/s10648-011-9191-6
- Deng, N., & Yoo, H. (2009). *Resources for reporting test scores: A bibliography for the assessment community*. Prepared for the National Council on Measurement in Education. University of Massachusetts Amherst. Retrieved from [http://www.ncme.org/ncme/NCME/NCME/Resource\\_Center/LibraryItem/Score\\_Reporting.aspx](http://www.ncme.org/ncme/NCME/NCME/Resource_Center/LibraryItem/Score_Reporting.aspx)
- Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, 14(7), 1–11. Retrieved from <http://pareonline.net/getvn.asp?v=14&n=7>
- Falk, A. (2012). Teachers learning from professional development in elementary science: Reciprocal relations between formative assessment and pedagogical content knowledge. *Science Education*, 96(2), 82–99. doi:10.1002/sce.20473
- Gearhart, M., Nagashima, S., Pfothenhauer, J., Clark, S., Schwab, C., Vendliski, T., ... Bernbaum, D. J. (2006). Developing expertise with classroom assessment in K-12 science: Learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment*, 11(3-4), 237–263. doi:10.1080/10627197.2006.965290
- Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. E. Geisinger (Ed.), *Handbook of testing and assessment in psychology* (pp. 479–494). Washington, DC: APA.

- Harlen, W. (2007) *The quality of learning: Assessment alternatives for primary education* (Primary Review Research Survey 3/4). Cambridge: University of Cambridge Faculty of Education.
- Hegarty, M. (2019). Advances in cognitive science and information visualization. In D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. 19–34). New York and London: Routledge.
- 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. doi:10.1016/j.edurev.2012.09.001
- Kahl, S. (2005). Where in the world are formative tests? Right under your nose! *Education Week*, 25, 11.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. doi:10.1111/jedm.12000
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198–211. doi:10.1080/0969594X.2015.1060192
- Kannan, P., Zapata-Rivera, D., & Leibowitz, E. A. (2018). Interpretation of score reports by diverse subgroups of parents. *Educational Assessment*, 23(3), 173–194. doi:10.1080/10627197.2018.1477584
- Katz, I. R. (2018). Foreword. In D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. XIV–XV). New York and London: Routledge.
- 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. doi:10.1111/ijasa.12065
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, 48(1), 4–11. doi:10.1080/00405840802577536
- 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. doi:10.1016/j.tate.2009.06.007
- 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. doi:10.1080/13803610802048874
- Shavelson, R. (2003). *On the integration of formative assessment in teaching and learning with implications for teacher education*. Stanford, CA, and Manoa, HI: Stanford Education Assessment Laboratory and University of Hawaii Curriculum Research and Development Group.

- Shavelson, R., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., ... Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education*, 21(4), 295-314. doi:10.1080/08957340802347647
- Tannenbaum, R. J. (2019). Validity aspects of score reporting. In D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. 9–18). New York and London: Routledge.
- Trumbull, E., & Lash, A. (2013). *Understanding formative assessment: Insights from learning theory and measurement theory*. San Francisco: WestEd.
- 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. doi:10.1016/j.stueduc.2013.04.002
- 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. doi:10.1080/0969594X.2014.999024
- 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. doi:10.1016/j.ijer.2017.02.013
- Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36, 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x
- William, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.
- Zapata-Rivera, J. 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 & Practice*, 21(4), 442–463. doi:10.1080/0969594X.2014.936357
- Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, 19, 116–138. doi:10.1080/10627197.2014.903653





## CHAPTER 2





# A GENERAL FRAMEWORK FOR THE VALIDATION OF EMBEDDED FORMATIVE ASSESSMENT

In educational practice, test results are used for several purposes. However, validity research is especially focused on the validity of summative assessment. This paper aimed to provide a general framework for validating formative assessment. We applied the argument-based approach to validation to the context of formative assessment. This resulted in a proposed interpretation and use argument (IUA) consisting of a score interpretation and a score use. The former involves inferences linking specific task performance to an interpretation of a student's general performance. The latter involves inferences regarding decisions about actions and educational consequences. The validity argument should focus on critical claims regarding score interpretation and score use, since both are critical to the effectiveness of formative assessment. The proposed framework is illustrated by an operational example including a presentation of evidence that can be collected on the basis of the framework.

Keywords: formative assessment; validation; argument-based approach

This chapter was previously published as:

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement*. doi:10.1111/jedm.12234

## 2.1 INTRODUCTION

There has been increasing attention around formative assessment in education (e.g., Herman, 2013; Torrance & Pryor, 2001; Wiliam, 2011a). Formative assessment is intended to support student learning by providing evidence about this learning. This evidence needs to be used by teachers, students or their peers for decisions and actions, such as determining the next steps in learning and instruction or providing feedback to (peer)students (e.g., Falk, 2012; Schneider & Andrade, 2013).

Since poor quality formative assessment may lead to less effective and less efficient teaching and learning, good quality in formative assessment is necessary. Validity is one of the most important criteria for the evaluation of assessments (AERA, APA, & NCME, 2014) and is often defined as the extent to which an assessment result is appropriate for its intended interpretation and use (e.g., Kane, 2013). The process of purposefully collecting and evaluating evidence regarding the appropriateness of assessment results is called validation.

To validate the proposed interpretation and use of formative assessment, an explicit validation framework can be quite useful. A framework enhances the standardization of the validation process and supports validation practice (Wools, Eggen, & Sanders, 2010). However, a framework aimed at facilitating the validation of formative assessment remains wanting.

This paper aims to provide such a framework. As there are many types of formative assessment, we focus on embedded formative assessment, the most formal type. In the next section, we will explain the concept of (embedded) formative assessment and the characteristics that distinguish it from summative assessment. Subsequently, the concepts of validity and validation will be discussed, and the argument-based approach to validation will be introduced as a general validation framework. We will then present the proposed validation framework for formative assessment. To clarify the proposed framework, we will describe a formative assessment example, to which we will apply the framework. Finally, we will address some implications and recommendations.

## 2.2 DEFINITION AND CHARACTERISTICS OF FORMATIVE ASSESSMENT

Formative assessment is conceptualized in different ways and is used interchangeably with several other concepts in the literature, such as assessment for learning, diagnostic assessment, and data-based decision-making (Antoniou & James, 2014; Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). The lack of a clear definition makes it difficult to implement formative assessment and evaluate its effectiveness (Bennett, 2011). Therefore, numerous review studies have been conducted to get a better grasp of the concept (e.g., Bennett, 2011; Dunn & Mulvenon, 2009; Gulikers & Baartman, 2017; Heitink, Van der Kleij, Veldkamp, Schildkamp, & Kippers, 2016; Sluijsmans, Joosten-ten Brinke, & Van der Vleuten, 2013; Wiliam, 2011b).

In particular, some authors perceive formative assessment as an instrument that provides feedback (e.g., Dunn & Mulvenon, 2009; Kahl, 2005), while others emphasize the process of using this feedback (e.g., Clark, 2012; Popham, 2008). Bennett (2011) has perceived each position as an oversimplification. Even the most carefully designed instrument is unlikely to be effective if the process surrounding its use is flawed. Similarly, the process is unlikely to work if the instrumentation does not fit its intended purpose. This paper follows Bennett's reasoning that formative assessment should be conceptualized as a thoughtful integration of both.

Formative assessment varies on a continuum from “on-the-fly” to “planned-for-interaction” to “curriculum-embedded” assessment (e.g., Forbes, Sabel, & Biggers, 2015; Furtak, 2006; Shavelson, 2003). On-the-fly assessment is the most informal. It does not involve a planned activity and occur as part of instructional activities. Planned-for-interaction assessment occurs, for example, when a teacher deliberately interrupts a lesson to ascertain students' understanding and alters instruction as necessary. Curriculum-embedded assessment is the most formal type. It consists of predefined tasks built into the school's educational program, that provide insights into students' current learning, and that is used to adapt teaching and learning to students' problem areas.

For the purpose of this paper, we focus on this latter category of formative assessment, as it most closely relates to summative assessment for which several validation frameworks has already been developed. We define embedded formative assessment (hereafter referred to as formative assessment) as both an instrument and a process, whereby evidence is purposefully gathered, judged, and used by teachers, students or their peers for decisions about actions to support student learning. This definition excludes informal formative assessment in which evidence is elicited in an improvised and unscheduled manner (Ruiz-Primo & Furtak, 2007).

This conceptualization of formative assessment differs from that of summative assessment in several ways. Formative assessment is characterized by its purpose in supporting student learning, while summative assessment is intended to provide a final decision about students' learning, for example, for selection, certification, or accountability purposes (Shavelson, 2003; Trumbull & Lash, 2013). This difference has implications for the design and practice of formative assessment (Wiliam, 2011b). In order to make these implications clear, we will discuss the distinctive characteristics of formative assessment.

First, formative assessment is aligned directly with the teaching and learning process, since the evidence obtained is used for actions like adjusting instruction, changing learning strategies or providing feedback (Harlen & James, 1997; Schneider & Andrade, 2013; Trumbull & Lash, 2013; Wiliam, 2011b). The uses may vary from teachers adjusting their instruction to students and peers changing their learning strategies. Nevertheless, as actions are necessary to support student learning, they make the actual process a distinctive feature of formative assessment (Bennett, 2011; Black & Wiliam, 2009).

Second, alignment with the teaching and learning process implies an assessment instrument that provides fine-grained information rather than a global reflection of students' capability (Goertz, Olah, & Riggan, 2009; Timperley, 2009). This means that a simple correct or incorrect score will usually not be sufficient. Student responses needs to be scored in such a way that fine-grained information about the depth of student learning is elicited. The availability of instructionally tractable information built into the curriculum is fundamental for deciding where students are in their learning, where they need to go, and how best to get there (Broadfoot et al., 2002; Herman, 2013; Timperley, 2009; Wiliam, 2011b). Without this kind of information, it would be very difficult to use the assessment information for actions that support leaning.

To conclude, formative assessment differs from summative assessment in terms of their explicit purpose in supporting learning. This purpose results in the need for alignment with the teaching and learning process, emphasizing its use by teachers and students and the need for fine-grained information from the assessment instrument. In the next section, the concepts of validity and validation will be discussed, and the argument-based approach to validation will be introduced as a general framework. This framework has been widely adopted in the validation of several summative assessments, such as certification testing (Kane, 2004) and admission testing (Chapelle, Enright, & Jamieson, 2010). Furthermore, Nichols, Meyers, and Burling (2009) attempted to use the approach for formative assessment. They especially focused on the proposed use of assessment information, without making demands on the instrument or methodology from which the information was collected. However, we argue that there is a need for a well-designed instrument that fits the proposed use.

### 2.3 ARGUMENT-BASED APPROACH TO VALIDATION

Since the early 1950s, Cronbach and Meehl's (1955) model of construct validity has been widely accepted and has been developed into a general framework for validation. The most general version of this model is based on three basic principles for validation: (1) the need for an explicit specification of the proposed interpretation; (2) the need for conceptual and empirical evaluation of the proposed interpretation; and (3) the need to consider alternate interpretations (Kane, 2013). These principles continue to be reflected in theories on validity and approaches to validation. For example, in Messick's (1989, p. 13) definition of validity: "...an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment [italics in original]."

While construct validity as a unifying framework has been useful on a theoretical level, it has not been an effective unifying framework for validation in practice (Cronbach, 1989). For example, Messick's conceptualization of validity was translated into a validation practice with the aim of presenting as much validity evidence as possible. This resulted in an overly lengthy process that was difficult to implement. To make the validation process more pragmatic while

still being faithful to basic scientific principles of construct validity, Kane (1992, 2004, 2006, 2013) proposed an argument-based approach to validation.

The argument-based approach consists of two stages: a developmental stage and an appraisal stage. In the developmental stage, an interpretation and use argument (IUA) is developed by specifying the proposed interpretation and use of assessment results. In the appraisal stage, the IUA is evaluated by critically examining its clarity, coherence, and plausibility.

The IUA consists of inferences regarding a score interpretation and a score use (Kane, 2013, 2016). A score interpretation involves claims about test takers or other units of analysis (e.g., teachers, schools). Claims about a score use involve decisions and possible consequences about these units of analysis. During the development of the IUA, the proposed interpretation and use are made explicit by incorporating their inherent inferences and assumptions.

Figure 2.1 shows an example of an IUA for a placement testing system (Kane, 2006). The first inference, named the scoring inference, is the evaluation of the observed performance leading to an observed score. Subsequently, the observed score is generalized to a universe score on a broader test domain. Within the next inference, the universe score is extrapolated toward a claim regarding the construct of interest in the practice domain. The last inference results in a decision on a student's skill level in relation to the construct of interest and placement in a specific course. These four inferences are likely to occur in most, if not all, IUAs for summative assessment (Kane, 2013).

Upon completion of the IUA, a critical evaluation of the inferences and assumptions is made in the appraisal stage, in which a validity argument can validate the proposed interpretation and use. The validity argument examines the coherence and completeness of the IUA and the plausibility of its inferences with respect to the purpose of the test (Crooks, 2004; Crooks, Kane, & Cohen, 1996; Dorans, 2012; Kane, 2013). Although the proposed interpretation and use are evaluated together, a given validity argument is not necessarily adequate for both (Cizek, 2016; Sireci, 2016). A valid score interpretation is a prerequisite for a valid score use, but it does not automatically justify it. Similarly, the rejection of a score use does not necessarily invalidate a prior underlying score interpretation.

To sum up, the central idea of the argument-based approach is to build and evaluate an argument that helps test developers demonstrate that assessment scores are sufficiently useful for their intended purpose. To the extent that the assessment results are intended to be used for certain decisions that affect

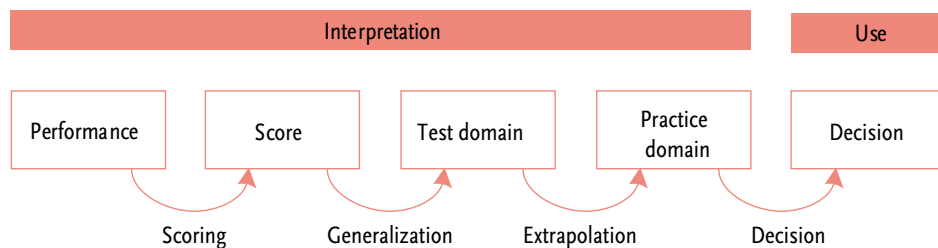


Figure 2.1 Example of an IUA

students or institutions, Kane (2013, 2016) emphasized the incorporation of inferences that are inherent in the proposed use, the evaluation of this proposed use, as well as the proposed interpretation. This also implies the inclusion of the consequences of these decisions in the validation process (Kane, 2016; Lane, 2014). If the proposed interpretation and use are supported by evidence and alternative explanations are rejected, it is appropriate to interpret and use assessment results in the proposed way (Kane, 2006). In the next section, the argument based approach is extended to a validation framework for formative assessment.

## 2.4 THE PROPOSED VALIDATION FRAMEWORK FOR FORMATIVE ASSESSMENT

The procedure of the argument-based approach would be similar for the validation of formative assessment as for the validation of summative assessment. Validation efforts would continue to be structured into a developmental stage to build the IUA as well as an appraisal stage to critically evaluate the IUA on the basis of a validity argument (Kane, 2004, 2006, 2013). We will begin the current section by describing the proposed inferences in the IUA, after which we will address the validity argument.

### 2.4.1 IUA for Formative Assessment.

The IUA for formative assessment consists of inferences regarding a score interpretation as well as inferences regarding a score use. Score-interpretation inferences cover claims about students' performance from the instrument, while score-use inferences involve decisions on this performance and possible consequences in the learning process.

With regard to the score-interpretation inferences, we propose a structure that is identical to the existing validation framework for summative assessment. This starts with 1) a *scoring inference*, whereby students' performance is converted into interpretable information about their thinking. In addition, only a limited sample of all possible items is administered to students. This then leads to 2) a *generalization inference*, in which we draw upon the scoring of a limited sample to make inferences about the generalization of this score to all possible items in a so-called test domain. Furthermore, there is 3) an *extrapolation inference*, in which the interpretation of all possible items is extrapolated to a more general claim about students' performance in a so-called practice domain. The practice domain is defined as the domain about which we would like to make a decision.

With regard to the score-use inferences, we propose a different structure from the validation framework for summative assessment. The existing 4) *decision inference* links students' performance regarding the construct in the practice domain to a decision about their performance. In addition, we propose three additional inferences, since the actual use of the decision by teachers and students is an essential part of formative assessment (Bennett, 2011; Kane, 2016). We propose 5) a *judgment inference* because inaccurate understanding of the decision could lead to inappropriate actions (Gearhart et al., 2006; Maciver, Anderson, Costa, & Evers, 2014; Moss, Brookhart, & Long, 2013). The judgment inference links the decision to a diagnosis by the teacher or student. Moreover, as teachers and students are assumed to use this diagnosis for the selection of appropriate actions (Bennett, 2011; Black & Wiliam, 2009), we propose 6) an *action inference*, which links the diagnosis to an action. Finally, the implementation of these actions is expected to support student learning. We therefore propose 7) a *consequence inference*, which links the action to student learning. The proposed IUA for formative assessment is presented in Figure 2.2. We will describe the assumptions within the inferences of the proposed IUA in the remaining part of this section.



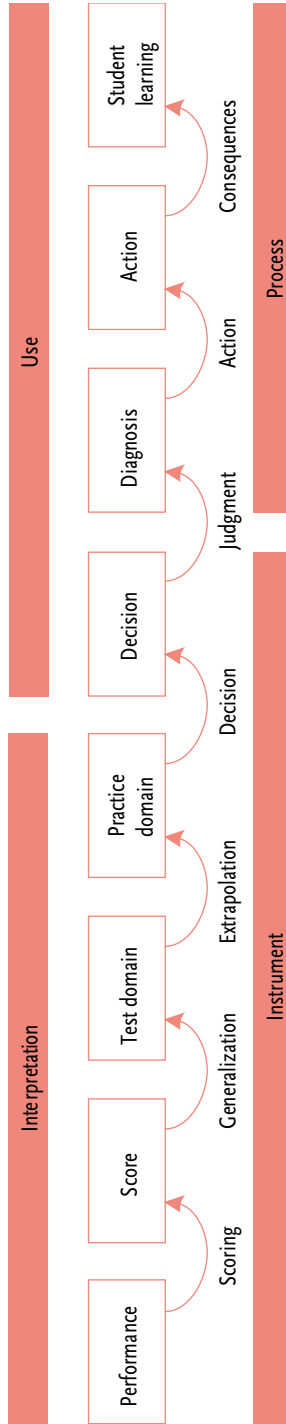


Figure 2.2 Proposed IUA for formative assessment

### *Assumptions within inferences*

*Scoring inference (performance - score).* It is proposed that students' performance on formative assessment tasks ought to be converted into interpretable information, such as a score, rubric, qualitative description, or a score profile with sub-scores. For this inference, we assume that a set of scoring rules or algorithms provides insights into student learning strategies and mistakes. For example, multiple-choice item distractors are used to score common errors in a student's understanding (Goertz et al., 2009). In the case of manual scoring, we assume that raters are able to observe students' performance and describe their thinking.

*Generalization inference (score - test domain).* To allow generalization, the tasks needs to be a representative sample of the test domain in terms of content, difficulty, and the kind of answers that provide insights into students' learning strategies and mistakes. Therefore, we assume that the sample of tasks reflect the depth of student learning. Furthermore, we assume that the sample of tasks is sufficiently large to control sampling error (Kane, 2013). A sufficiently large sample is needed to support generalization because the more confident teachers and students are about students' level, the more effectively they can adjust instruction. To illustrate, an error could be a careless mistake, a persistent misconception, or a lack of understanding caused by inadequate knowledge (Bennett, 2011). Depending on the cause, the action will range from minimal feedback to re-teaching and significant investment in eliminating misconceptions. With a representative and sufficiently large sample of items, teachers and students can select appropriate action.

*Extrapolation inference (test domain - practice domain).* For extrapolation, we assume that the tasks in the test domain reflect the particular learning objective, learning goal, or attainment goal in the practice domain. This means that the tasks include all aspects of the learning objective that are relevant for making a distinction between different student performances. None of the important aspects of the learning objective are overlooked (construct underrepresentation), and neither are other aspects confounded (construct irrelevant variance). Furthermore, it is assumed that the tasks result in the students performing the expected thinking processes we are interested in.

*Decision inference (practice domain - decision).* The decision inference is drawn from a decision rule that specifies how the decision will be made. It is

assumed that the cut-off score is in line with students' mastery of a learning objective. In addition, it is assumed that misclassifications with regard to misconceptions and learning strategies are minimized.

*Judgment inference (decision - diagnosis).* For the judgment inference, we assume that teachers and students are able to correctly understand the decision derived from the assessment instrument. This means that the presentation of the decision fits teachers' and students' level of assessment literacy (e.g., Popham, 2011). Furthermore, we assume that teachers and students are able to link the decision to students' individual circumstances, such as the amount of effort invested, progress over time, and the particular context (Bennett, 2011). This suggests that formative assessment is student-referenced (Harlen & James, 1997), with the possibility of tailoring the actions to individual students' needs and motivating them. For example, a teacher or student can conclude that a non-mastery decision was based on a careless mistake, a persistent misconception, or a lack of understanding. It is also possible that the student actually mastered the learning objective but that he or she was not focused or motivated, did not read the assignment correctly, or that the program might have been crashed.

*Action inference (diagnosis - action).* To select appropriate actions, we assume that the assessment information is tied to the curriculum and fits teachers' and students' knowledge base, including subject-matter knowledge and pedagogical content knowledge (Falk, 2012; Forbes et al., 2015; Furtak & Heredia, 2014; Goertz et al., 2009; Heritage, Kim, Vendlinski, & Herman, 2009; Herman, Osmundson, Ayala, Schneider, & Timms, 2006; Sabel, Forbes, & Zangori, 2015). This would allow a teacher or student to select a new learning objective if they diagnose that the learning objective has been mastered. If they diagnose that the learning objective has not been mastered, then the student could decide on further practice, or the teacher could choose to provide minimal feedback, reteach the learning objective, or seek to eliminate the misconception.

*Consequence inference (action - student learning).* To allow the consequences, we assume that the approach to formative assessment results in student learning. However, the impact on learning also depends on the educational context (Bennett, 2011). Even if teachers and/or students act appropriately, the educational context could minimize the effect on students' learning (Bennett, 2011; Goertz et al., 2009). Therefore, this claim also assumes that the context is sufficiently supportive, including tools for data access, school leaders stimulating the use of formative assessment, teachers sharing the learning objectives, and students

actively involved and motivated (Herman, 2013; Moss et al., 2013; Stobart, 2012; Torrance & Pryor, 2001).

#### *2.4.2 Validity Argument for Formative Assessment.*

The validity argument for formative assessment would focus on both the score interpretation and the score use, since a failure in either part can reduce its effectiveness (Bennett, 2011). If the score interpretation is wrong, the basis of the actions is weakened. Similarly, if the score interpretation is correct and is presented in an understandable and meaningful way, but the action is inappropriate, learning is also less likely to occur. Within the IUA, the underlying inferences that seem to be questionable or critical should receive the most attention because they address the weakest links in the IUA (Kane, 2006; Wools, Eggen, & Béguin, 2016).

To the extent that the inferences are supported with evidence and alternative explanations are rejected, the validity argument is concluded by stating whether it is valid to interpret and use the assessment results. It is important to note that the analytical or empirical evidence will focus on making the claims plausible for a significant number of individuals rather than for individual cases (Kane, 2016).

## 2.5 OPERATIONAL EXAMPLE OF THE VALIDATION FRAMEWORK FOR FORMATIVE ASSESSMENT

To clarify the proposed validation framework for formative assessment, Bennett (2011) argues that we need one or more operational examples that show what formative assessment built on the basis of this theory looks like. This section contains such an example, to which we will apply the framework. We used the embedded formative assessment platform Groeimeter (GM), which was developed by the Cito Institute for Educational Measurement in the Netherlands. We will start with a description of the components of GM, followed by a description of how it is used. Then, we will apply the proposed validation framework to GM and will provide some examples as a means of validating it.

### 2.5.1. *Description of GM*

GM is aimed at supporting primary school teachers and guiding students in learning arithmetic. It consists of embedded formative assessment tasks, a teacher dashboard, and a student dashboard. The formative assessment tasks are related to the learning objectives of the Dutch arithmetic curriculum. Each predefined task is supposed to measure one learning objective. There are two types of assessment tasks, depending on what best fits the learning objective to be measured. The first type is a digital test in which students answer seven predefined items online. The number of items was chosen to make the tests practical. Digital tests are used for learning objectives that can be operationalized into automatically scored items, for example: “The student is able to calculate additions and subtractions up to 20.” The items could be short-answer, multiple-choice, multiple-response, hotspot, or matching items. For example, students fill in the right answer to the short-answer item: “How many balls do John and Mike have together?” or they need to select the coins that amount to 15. For the digital test, mastery is assigned to six correct items (Béguin & Straat, 2019). The second type is an assignment, for instance, having a group discussion or making a drawing. It is used when the learning objective is not suitable for automatic scoring because it requires more cognitively complex thinking. An example of such a learning objective is: “The student can think and reason critically about length and perimeter in meaningful problem situations.” In the assignment, students were asked to come up with three different rectangles with a 16-meter perimeter and to explain their choices. In another assignment, they had to calculate the perimeter of a new fence for the parcels of land belonging to the farmer, James. For this assignment, mastery or non-mastery needed to be manually assigned after scoring the assignment.

GM contains a teacher dashboard that shows students’ performance on completed assessments as a green or orange block, indicating mastery or non-mastery, respectively, of the measured learning objective. The program allows the teacher to manually change this status. Furthermore, the dashboard displays the students’ icons, with information about their individual progress and item responses. Finally, it shows all the learning objectives of the Dutch arithmetic curriculum, including an explanation and item example of the accompanying assessment. It is possible to assign a learning objective to an individual student or to the whole group of students.

GM also contains a student dashboard that shows the learning objectives assigned to the student. In this dashboard, the student can complete the assessment, up which his or her performance is again shown as a green or orange block. It is possible to view the individual item responses on the digital test and compare them with the correct answers.

### *2.5.2 Use of GM*

Teachers and students are supposed to view the students' mastery and individual item responses on the completed digital test and compare them with the correct answers. They can also analyse the students' answers on the assignment. In this way, teachers and students can judge the results themselves. Teachers can try to explain the results by linking them to the students' individual circumstances. When teachers determine that the automatically assigned status (mastery/non mastery) does not reflect reality, they can overrule the status.

Assessment results are supposed to be used to guide follow-up action. For example, teachers are expected to provide additional instruction if they conclude that a learning objective has not been mastered due to a particular misconception. Students could undertake additional assignments to exercise a learning objective. It is assumed that the implementation of these actions supports student learning.

### *2.5.3 Designing a Validation Study for GM*

The GM example illustrates the two distinctive characteristics of embedded formative assessment. First, it consists of an instrument that provides fine-grained information about students' performance vis-à-vis the learning objectives defined in the Dutch curriculum. Second, this information is supposed to be used for actions in the teaching and learning process.

This conceptualization requires an IUA that consists of inferences regarding both a score interpretation and a score use. Table 2.1 shows the inferences and its underlying assumptions. Furthermore, it provides possible sources of analytical and empirical evidence that can be collected to evaluate validity.

Since validation is a major activity, it is important to provide most attention to the most questionable or critical inferences. In our opinion, the most questionable and critical assumption of the score interpretation would be the need for fine-grained information. It should be made plausible that the

assessment results provide enough insight into the depth of student thinking processes. In terms of the assumptions regarding score use, it should be made plausible that teachers and students are able to use the score interpretation to inform instructional actions that support learning.

Table 2.1 Inference, assumptions, and possible sources of evidence that can be collected in the validation of GM

Inferences	Assumptions	Sources of evidence
Scoring: from student performance to score.	Teachers are able to consistently mark performance on the assignments.	Interrater reliability analysis of teachers' descriptions regarding the same student undertaking an assignment.
	The scoring rules provide insights into student learning strategies and mistakes.	Analyzing whether the distractors correspond to common learning strategies and mistakes.
Generalization: from score to test domain.	Both types of tasks reflect the depth of student learning.	Evaluation of test content matrices with regard to content and difficulty.
	Both types of tasks are sufficiently large to control sampling error.	Analysis of whether (a) different (number of) items provide similar inferences about students' thinking.  Calculating a reliability coefficient.
Extrapolation: from test domain to practice domain.	The tasks result in students performing the expected thinking processes.	Think-aloud protocols with students, which investigate whether they perform at the level of the expected thinking processes while completing the items.
	The tasks include all critical aspects of the learning objective.	Study the relationship with other measures of the learning objective, for example, observations, standardized tests, etc.
Decision: from practice domain to decision.	The decision is in line with students' actual mastery of the learning objective.	Comparing students' performance on a specific learning objective to other learning objectives of the same level of difficulty.  Comparing the decision on an external criterion, such as oral exams or think-aloud studies.  Log file analysis investigating how many times the decision has been overruled by the teacher.
Judgment: from decision to diagnosis.	The assessment information supports teachers and students in correctly interpreting the decision in the teacher and student dashboards.	Think-aloud protocols that analyze how teachers and students interpret the decision.  Set-up an experiment where teachers are asked to interpret assessment information in different scenarios.

Table 2.1 (continued)

Inferences	Assumptions	Sources of evidence
Action: from diagnosis to action.	The measured learning objective is recognizably connected to teaching and learning.  The assessment information from GM supports teachers and students in selecting actions that enhance the teaching and learning process.	Interviews that investigate whether teachers were able to correctly explain the meaning of the learning objectives.  Analysis of the connection between the learning objectives in GM and the teaching methods used.  Background documents of test developers that specify the relation between teaching and learning.  Classroom observation and/or log file analysis that show what actions teachers and students perform.  Interviews or questionnaires about how teachers and students experience the usability of GM.
Consequence: from action to student learning.	The performed actions have a positive impact on student learning.  There are no obvious obstacles within the educational context.	Longitudinal study comparing schools that utilize GM and those that do not.  Evaluating the characteristics of schools in which GM works well.

## 2.6 CONCLUSION AND DISCUSSION

In this paper, we proposed an extension of the argument-based approach (Kane, 2006, 2013) to the validation of embedded formative assessment. Embedded formative assessment was defined as both an instrument and a process, whereby evidence from a purposefully designed instrument is gathered, judged, and used for decisions about actions to support student learning. This conceptualization requires an IUA consisting of inferences regarding both a score interpretation and a score use. The score interpretation connects the specific task performance from the assessment instrument with an interpretation about the student's general performance. The score use connects that interpretation to decisions about actions in the teaching and learning process that are intended to support student learning. The validity argument should focus on critical claims regarding score interpretation as well as score use, since both are critical to the effectiveness of formative assessment.



In comparing this proposed framework in Figure 2.2 to the existing validation framework exemplified in Figure 2.1, the proposed structure of the inferences regarding the score interpretation is identical. However, the content of the score interpretation regarding formative assessment differs because the alignment with the teaching and learning process requires a different level of information granularity. This would result in different kind of tasks with different formulations regarding the scoring, generalization, and extrapolation inferences. For example, the scoring inference often implies a way of scoring that provide insight into student learning strategies and mistakes, meaning that an aggregated score would usually not be sufficient. Furthermore, the generalization and extrapolation links may be less far stretching than for summative assessment due to a narrowly defined practice domain (Crooks, 2004; Crooks et al., 1996; Dorans, 2012; Stobart, 2012). Therefore, generalization and extrapolation are less problematic and pose problems that are different from those of summative assessment, which often address broad constructs such as language literacy. For broad constructs, generalization and extrapolation could be so important, that there is a need to add inferences (see, e.g., Kane, 2004; Wools et al., 2010). In addition to the score-interpretation inferences, we included three additional use inferences to make the use more visible (Bennett, 2011; Kane, 2016): a judgment inference, an action inference, and a consequence inference.

Adjustments in the IUA also changed the validity argument that evaluates the IUA; for different uses (e.g., formative vs. summative), different issues tend to become more salient. These differences demonstrate that an assessment instrument cannot be used interchangeably for both summative and formative purposes. The formative use of summative assessment and vice versa can only be applied after extensive and careful research.

Noteworthy, the GM system was used as an operational example to illustrate how the proposed framework suits the definition of curriculum-embedded formative assessment. It would be interesting to perform validation studies that provide analytical and empirical evidence with regard to the underlying assumptions.

In addition, the framework could be applied to other examples of curriculum-embedded assessment. This assumption might be investigated in a follow-up study, as an IUA needs to be developed and evaluated for each assessment in a particular context of practice (Kane, 2004). This could result in the specification of a somewhat different network of inferences and assumptions

in another case-specific IUA, with the evaluation in the accompanying validity argument.

Furthermore, we developed a framework that suits the definition of curriculum-embedded assessment, which are the most formal category of formative assessment. However, a significant number of formative assessment is informal, such as a diagnostic conversation indicating a student's strengths and weaknesses. In a follow-up study, it would be interesting to investigate whether this framework could be applied to more informal formative assessment. To do this, we would need to further specify the differences between formal and informal formative assessment and identify the consequences for validation.

The general framework could be a meaningful contribution to guide the design and evaluation of formative assessment and to enhance our reasoning on validity. For example, it emphasizes the importance of actual use by teachers and students, placing substantial demands on teachers' content and pedagogical knowledge (Herman, Osmundson, Dai, Ringstaff, & Timms, 2015). To support the judgments and actions of the user, understandable score reports could be an important tool requiring careful design. This tool could meaningfully communicate the assessment scores and reduce the demands on users' knowledge and skills (Hattie & Brown, 2008; Matuk, Linn, & Elon, 2015; Ryan, 2006; Zapata-Rivera & Katz, 2014).

Finally, this paper opens up the discussion about the scope of validity theory, which is currently under intense debate (Newton & Shaw, 2016). The perspectives surrounding this debate range from those who insist that validity should remain a technical evaluation of measurement procedures (Borsboom, Mellenbergh, & van Heerden, 2004) to those who insist that it should become a broad concept to evaluate use of assessment results in the larger system (Moss, 1998). Although it seems possible to limit the scope of "validity" to a technical evaluation of summative assessment, this is impossible for formative assessment. The actual use and educational context of formative assessment are essential aspects of the effectiveness of these assessments. Shepard (2016) thus gets to the point in her remark that "Just as test design is framed by a particular context of use, so too must validation research focus on the adequacy of tests for specific purposes" (p.273). Therefore, we felt the need to incorporate use inferences in the IUA for formative assessment, thus making the proposed use of tests an integral part of validation. The currently developed validation frameworks for summative assessment, however, do not include such use inferences. These

differences could result in confusion around the concept of validity, which is not desirable. Therefore, the necessary incorporation of use inferences for formative assessment leaves the question of whether the concept of validity should be expanded to an overall evaluation of the score interpretation as well as of the score use. Referring to all of this as validation would make it possible to strive for a uniform conceptual framework within validity theory for both summative and formative assessment.

## 2.7 REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability*, 26, 153–176. doi:10.1007/s11092-013-9188-4
- Béguin, A. A., & Straat, J. H. (2019). On the number of items in learning goal mastery testing. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 121–134). Cham, Switzerland: Springer International Publishing.
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. doi:10.1080/0969594X.2010.513678
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5–31. doi:10.1007/s11092-008-9068-5
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071. doi:10.1037/0033-295X.111.4.1061
- Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x

- Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy and Practice*, 23(2), 212–225. doi: 10.1080/0969594X.2015.1063479
- Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24(2), 205–249. doi:10.1007/s10648-011-9191-6
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Crooks, T. J. (2004). *Tensions between assessment for learning and assessment for qualifications*. Paper presented at the third conference of the Association of Commonwealth Examinations and Accreditation Bodies (ACEAB), Nadi, Fiji.
- Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, 3(3), 265–286. doi:10.1080/0969594960030302
- Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice*, 31(4), 20–37. doi:10.1111/j.1745-3992.2012.00250.x
- Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, 14(7), 1–11. Retrieved from <http://pareonline.net/getvn.asp?v=14&n=7>
- Falk, A. (2012). Teachers learning from professional development in elementary science: Reciprocal relations between formative assessment and pedagogical content knowledge. *Science Education*, 96(2), 82–99. doi:10.1002/sce.20473
- Forbes, C. T., Sabel, J. L., & Biggers, M. (2015). Elementary teachers' use of formative assessment to support students' learning about interactions between the hydrosphere and geosphere. *Journal of Geoscience Education*, 63, 210–221. doi:10.5408/14-063.1
- Furtak, E. M. (2006). *Formative assessment in K-8 science education: A conceptual review*. Commissioned paper for the Committee on Science Learning, Kindergarten through Eighth Grade, National Research Council.
- Furtak, E. M., & Heredia, S. C. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching*, 51(8), 982–1020. doi:10.1002/tea.21156
- Gearhart, M., Nagashima, S., Pfothenauer, J., Clark, S., Schwab, C., Vendliski, T., ... Bernbaum,

- D. J. (2006). Developing expertise with classroom assessment in K-12 science: Learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment*, 11(3-4), 237–263. doi:10.1080/10627197.2006.9652990
- Goertz, M. E., Olah, L. N., & Riggan, M. (2009). Can interim assessments be used for instructional change? Policy brief. RB-51. *CPRE Policy Briefs*. Retrieved from [http://repository.upenn.edu/cpre\\_policybriefs/39](http://repository.upenn.edu/cpre_policybriefs/39)
- Gulikers, J., & Baartman, L. (2017). *Doelgericht professionaliseren: Formatieve toetspraktijken met effect! Wat DOET de docent in de klas?* [Targeted professionalization: Formative assessment practices with effect! What DOES the teacher in the classroom?]. PPO-NRO 405-15-722. Den Haag: NRO
- Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3), 365–379. doi:10.1080/0969594970040304
- Hattie, J., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, 36(2), 189–201. doi:10.2190/ET.36.2.g
- Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, 17, 50–62. doi:10.1016/j.edurev.2015.12.002
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. L. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31. doi:10.1111/j.1745-3992.2009.00151.x
- Herman, J. L. (2013). *Formative assessment for next generation science standards: A proposed model* (CRESST Resource Paper No. 16). Los Angeles, CA: CRESST.
- Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timss, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Technical Report 703). Los Angeles, CA: CRESST.
- Herman, J. L., Osmundson, E., Dai, Y., Ringstaff, C., & Timss, M. (2015). Investigating the dynamics of formative assessment: Relationships between teacher knowledge, assessment practice and learning. *Assessment in Education: Principles, Policy & Practice*, 22(3), 344–367. doi:10.1080/0969594X.2015.1006521
- Kahl, S. (2005). Where in the world are formative tests? Right under your nose! *Education Week*, 25(4), 11.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535. doi:10.1037/0033-2909.112.3.527

- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135–170. doi:10.1207/s15366359mea0203\_1
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. doi:10.1111/jedm.12000
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198–211. doi:10.1080/0969594X.2015.1060192
- Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, 26(1), 127–135. doi:10.7334/psicothema2013.258
- 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. doi:10.1111/ijsa.12065
- Matuk, C. F., Linn, M. C., & Eylon, B. (2015). Technology to support teachers using evidence from student work to customize technology-enhanced inquiry units. *Instructional Science*, 43, 229–257. doi:10.1007/s11251-014-9338-1
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17, 6–12.
- Moss, C. M., Brookhart, S. M., & Long, B. A. (2013). Administrators' roles in helping teachers use formative assessment information. *Applied Measurement in Education*, 26(3), 205–218. doi:10.1080/08957347.2013.793186
- Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, 23(2), 178–197. doi:10.1080/0969594X.2015.1037241
- Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice*, 28(3), 14–23. doi:10.1111/j.1745-3992.2009.00150.x
- Popham, W. J. (2008). *Transformative assessment in action*. Alexandria, VA: ASCD.
- Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, 46(4), 265–273. doi:10.1080/08878730.2011.605048
- Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, 44(1), 57–84. doi:10.1002/tea.20163

- Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Mahwah, NJ: Lawrence Erlbaum.
- Sabel, J. L., Forbes, C. T., & Zangori, L. (2015). Promoting prospective elementary teachers' learning to use formative assessment for life science instruction. *Journal of Science Teacher Education*, 26(4), 419–445. doi:10.1007/s10972-015-9431-6
- Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action: Conclusions drawn from a special issue on formative assessment. *Applied Measurement in Education*, 26, 159–162. doi:10.1080/08957347.2013.793189
- Shavelson, R. (2003). *On the integration of formative assessment in teaching and learning with implications for teacher education*. Stanford, CA, and Manoa, HI: Stanford Education Assessment Laboratory and University of Hawaii Curriculum Research and Development Group.
- Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy and Practice*, 23(2), 268–280. doi:10.1080/0969594X.2016.1141168
- Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy and Practice*, 23(2), 226–235. doi:10.1080/0969594X.2015.1072084
- Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. (2013). *Toetsen met leerwaarde: Een reviewstudie naar de effectieve kenmerken van formatief toetsen* [Assessment with learning value: A review study into the characteristics of effective formative assessment]. NWO-PROO 411-11-697. Den Haag: NWO.
- Stobart, G. (2012). Validity in formative assessment. In J. Gardner (Ed), *Assessment and learning* (2nd ed., pp. 233–242). London: Sage.
- Timperley, H. (2009). *Using assessment data for improving teaching practice*. Paper presented at the ACER research conference on assessment and student learning, Perth, Western Australia.
- Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, 27(5), 615–631. doi:10.1080/01411920120095780
- Trumbull, E., & Lash, A. (2013). *Understanding formative assessment: Insights from learning theory and measurement theory*. San Francisco, CA: WestEd.
- 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. doi:10.1080/0969594X.2014.999024

- Wiliam, D. (2011a). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.
- Wiliam, D. (2011b). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14. doi:10.1016/j.stueduc.2011.03.001
- Wools, S., Eggen, T. J. H. M., & Béguin, A. A. (2016). Constructing validity arguments for test combinations. *Studies in Educational Evaluation*, 48, 10–18. doi:10.1016/j.stueduc.2015.11.001
- Wools, S., Eggen, T. J. H. M., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO*, 8, 63–82.
- Zapata-Rivera, J. 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 & Practice*, 21(4), 442–463. doi:10.1080/0969594X.2014.936357





# CHAPTER 3



## FORMATIVE USE OF TEST RESULTS: A USER'S PERSPECTIVE

Despite the potential of using test data to support student learning, several studies have concluded that the actual use of test data remains limited. The present study addresses this problem by examining (1) the types of actions for which teachers, internal coaches, principals and parents within primary education want to use test results and (2) the information needed to perform these actions. The results obtained from the questionnaires show that the various users want to use test results for actions that support learning, which amounts to a discrepancy relating to actual use. Furthermore, the various users perform actions on different levels, thus indicating the need for tailored reports that fit the information needs of individual users. The results of the focus group method reveal the information needs of teachers, suggesting implications for the development of new score reports.

Keywords: formative assessment; test use; test results; information needs; needs assessment

This chapter was previously published as:

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. doi:10.1016/j.stueduc.2016.11.002

### 3.1 INTRODUCTION

Research points to the potential of formative assessments as a way of supporting student learning (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014; Black & Wiliam, 2009; Popham, 2009; Schildkamp & Lai, 2013). Formative assessment provides teachers with data about student performance. This data can be used to make decisions about the next steps in instruction, which are likely to be better, or better founded, than the decisions teachers would have taken intuitively in the absence of that data (Black & Wiliam, 2009).

To be able to use test data for student learning, teachers perform several cognitive steps (Davenport & Prusak, 1998; Ebbeler, Poortman, Schildkamp, & Pieters, 2016; Marsh, 2012). First, the collected data must be interpreted by giving meaning to scores. This can be done by summarizing the data in a more concise form. Subsequently, the interpreted data has to be contextualized by, for example, comparing the interpreted data with other information. The combination of different sources of information results in usable knowledge, which serves as a basis for decisions about an action, after which the action is executed. The impact of the action on student learning can then be evaluated using new data. As such, an iterative process is created (Mandinach & Jackson, 2012).

Several studies show that teachers have difficulty completing the phases of this iterative process (e.g. Hambleton & Slater, 1997; Hellrung & Hartig, 2013; Meijer, Ledoux, & Elshof, 2011; Schildkamp & Teddlie, 2008; Van der Kleij & Eggen, 2013). They especially struggle with (1) interpreting the test results and (2) translating them into actions that support learning. There are two possible explanations for these problems. First, the presentation regarding test results does not correspond with the assessment literacy skill level of teachers, resulting in difficulty interpreting the data and thereby making inappropriate use of the test results, with all its attendant consequences (e.g. Popham, 2009; Zapata-Rivera, VanWinkle, & Zwick, 2012). Second, the content of the presented data does not fit the information needs of teachers, resulting in problems translating the data into actions that support learning (e.g. Wiliam, 2011).

A considerable number of studies address the first explanation by allowing teachers and other users to develop the required assessment literacy skills (e.g. Lukin, Bandalos, Eckhout, & Mickelson, 2004; Verhaeghe, Vanhoof, Valcke, & Van Petegem, 2011). For example, some studies show a positive effect of training in terms of developing the required knowledge and skills to analyse and interpret

data (e.g. Ebbeler et al., 2016; Van Geel, Keuning, Visscher, & Fox, 2016; Zwick et al., 2008). Other studies address the interpretation problem by adjusting the data presentation to the user's skill level (e.g. Van der Kleij, Eggen, & Engelen, 2014) since it has been suggested that the chosen method of data visualization can reduce the assessment literacy needs of users (Hattie & Brown, 2008).

The second explanation regarding the problem of using test data for student learning focuses on the content of the presented data. According to Zapata-Rivera and Katz (2014), everyone involved in the learning process of a child uses test results as presented through score reports, yet each audience has its unique types of decisions to be made on test results. If a score report designer defines the needs of the target audience, he opens up the possibility of tailoring the score report to meet the unique information needs of that audience. Within the target audience, four groups of users are distinguished: teachers, who are responsible for instruction and teaching a group of students; internal coaches, who coach teachers and support students with special needs across classes; principals, who are responsible for the school organization and parents, who support the learning of their own child.

Fitting the presented data with the information needs of users is often overlooked. According to Wiliam (2011), assessment data are made available to users under the assumption that this data are useful in some way. Too little attention has been paid to the types of actions that intended audiences want to perform on the basis of test data. The current study addresses this problem in the context of Dutch primary education. It seeks to determine the types of actions that teachers, internal coaches, principals and parents in primary education want to perform with the use of test results and the information needed to enable these actions.

### *3.1.1 Educational Decision-Making*

In education, decision-making about instructional processes is an everyday activity. These decisions are taken at individual, group and school levels and can have important consequences for student learning. For example, on an individual level, decisions may pertain to whether a student should receive additional support. On a group level, decisions can relate to categorizing students into different levels for differentiation of instruction. On a school level, decisions may pertain to selecting a new teaching method. In order to ascertain whether these

kinds of decisions are correct, it is important that decisions are informed by high-quality evidence (Brookhart & Nitko, 2008).

Test results are one source of data that can be used as evidence to support educational decision-making (Zapata-Rivera & Zwick, 2011). A test can be described as “an instrument or systematic procedure for observing and describing one or more characteristics of a student using either a numerical scale or a classification scheme” (Brookhart & Nitko, 2008, p. 5). Combined with other assessment data, such as student observations, oral questions and students’ work, an accurate picture of the student can be obtained and decisions can be informed (Brookhart & Nitko, 2008; Mandinach, 2012).

Despite the availability of test data meant to inform the didactical decisions of teachers, various studies conclude, however, that the actual use of test data for student learning is limited (Ledoux, Blok, Boogaard, & Krüger, 2009; Meijer et al., 2011; Vanhoof, Verhaeghe, Verhaeghe, Valcke, & Van Petegem, 2011; Verhaeghe et al., 2011). Instead, test data are used for other purposes, such as communication and evaluation, which do not automatically result in increased student learning. The use of data for communication has to do with informing parents about students’ ability or with informing inspectorate<sup>1</sup> for the purpose of accountability (Ebbeler et al., 2016; Van der Kleij & Eggen, 2013) while the sole purpose of the use of data for evaluation is to appraise students’ performance. The actions that could follow from these judgments are not carried out (Brookhart & Nitko, 2008).

### *3.1.2 Presentation of Test Results*

Test results are presented using score reports. Score reports are the vehicle for translating the test results into useful actions that support learning. It is a form of communication, with a sender, a message and an audience. The sender of score reports is the test developer or test agency presenting the results. The message deals with the content of the score report, and the audience consists of the people who use the test results (Hattie, 2009; Ryan, 2006).

---

<sup>1</sup> The Dutch Inspectorate assesses and stimulates the quality of primary education and reports on the quality of each school to the public

To foster the use of test results for educational decision-making, the score report content should directly inform the audience about their decisions (Aschbacher & Herman, 1991; Hattie, 2009; Zapata-Rivera & Katz, 2014). Understanding the purpose for reading the test results in a score report helps to present the right message. Questions illustrating this statement include: What are the users' goals? What do the users want to know? What decisions should the information inform, or what actions should it motivate or justify? If the score report presents content tailored to a user's desired actions or decisions, the user would always know what to do with data that have collected and presented (Aschbacher & Herman, 1991; Wiliam, 2011).

### *3.1.3 Tailoring Score Reports to Various Users*

Test results are often used by more than one intended audience, including teachers, parents, internal coaches and principals. As pointed out by Zapata-Rivera and Katz (2014) and Mandinach (2012), depending on the position of the user, each audience has its unique types of decisions to be made on the basis of test results. For example, teachers would be more involved in the decision process of an individual student or group of students while principals would be more focused on the decision process at the school level (Schildkamp & Kuiper, 2010). Internal coaches would be interested in the performance of all students while parents would be more interested in the performance of their own child (NEGP, 1998). With various intended audiences, it is likely that specially designed reports would be needed for each. The need for tailored reports will thus be reinforced depending on the variations among the decisions and information needs of the different audiences (Bradshaw & Wheeler, 2009; Hambleton & Slater, 1997).

### *3.1.4 Identifying Users' Needs*

It is the responsibility of test developers to ensure that the content of the score report fits the information needs of the user (Ryan, 2006). Because of this responsibility, various studies have called for the creation of score reports that meet the needs of different audiences (Aschbacher & Herman, 1991; Goodman & Hambleton, 2004; Hambleton & Slater, 1997; Jaeger, 1998; Wainer, Hambleton, & Meara, 1999). Hambleton and Zenisky (2013) and Zapata-Rivera, et al. (2012) present a model for score report development – a user-centred model which

starts with a needs assessment. This needs assessment should establish common ground between the test developer and the test user, bridging the gap between the information that results from an assessment and the actions the user wants to perform from the information. The results from the needs assessment will be the basis on “which all of the other steps in report design are linked” (Hambleton & Zenisky, 2013, p. 486).

The current study performed such a needs assessment. As mentioned earlier, its aim was to determine the types of actions that various users would like to perform with the use of test results as well as the information needed to enable these actions. Regarding this aim, we looked beyond the available information from existing tests and score reports. Instead, the focal point of this study was on starting from the decisions or actions that a user would ideally like to make (Wiliam, 2011) in order to support the use of test results for student learning.

### *3.1.5 Research Questions*

The main questions addressed in this study are as follows:

1. Which types of actions would users choose as desired uses of test results and how do these actions relate to actual uses?
2. What, if any, is the extent of the differences between teachers, internal coaches, principals and parents with regard to desired and actual uses and corresponding actions?
3. What information from test results is needed to perform the desired actions?

## **3.2 METHOD**

In order to answer the research questions, data were collected from different user groups within Dutch primary schools. Four different kinds of users were distinguished: teachers, internal coaches, principals and parents and guardians (hereafter, parents). A questionnaire was developed for the teachers, internal coaches and principals to identify the actions for which test results are used in the context of teaching. As the focus was on actions related to teaching and were, therefore, not applicable to parents, a separate questionnaire for parents was developed.



In addition, qualitative data from focus groups were gathered to validate the results of the questionnaire data and to further specify the information needs. Based on the results from the questionnaires, we decided to target these focus groups at teachers. In order to facilitate the readability of this article, these choices are further elaborated in the results section.

Table 3.1 shows the relation between the research instruments and the different user groups and research questions. In the next session, the instruments, procedure, data analyses and sample characteristics are discussed.

Table 3.1 Relation between instruments, respondents and research questions

Instruments	Respondents	Research questions		
		1 Actions	2 User differences	3 Information needs
Questionnaire 1	Teachers, internal coaches, principals	x	x	
Questionnaire 2	Parents	x	x	
Focus group	Teachers	x		x

### 3.2.1 Instruments and Procedure

#### *Questionnaire 1 - teachers, internal coaches and principals*

The first questionnaire was developed to investigate the actions that teachers, internal coaches and principals deemed desirable in relation to test results within the context of teaching as well as the actual use of such results. Test results were defined as results from a systematic instrument, such as written or digital tests, excluding results from other assessment methods like observations and verbal responses from students. The actual use depends on the availability of information from current tests, which in the Netherlands, are mostly standardized tests aimed at monitoring students and written or digital tests from teaching methods. In terms of desired use, respondents were asked to mention all actions independent of currently available information and tests.

Alongside the eleven items on the background of the respondents, the questionnaire consisted of three items showing a list of possible actions for which test results could be used. The first and second items consisted of multiple

response questions in which respondents were required to select actions relating to actual and desired use, respectively. As respondents could select all actions as desired use, the third item asked respondents to choose the most important desired action from their selection. This provided greater insight into the degree of interest relating to the different actions. The questionnaire is included in Table A.1 of Appendix A.

The list of possible actions resulted from the grid shown in Table 3.2. This grid consists of actions on three levels (individual level, group level and school level) and three purposes (communicating learning, supporting learning and evaluating learning). This enabled the possibility to describe some patterns in the answers. The levels were related to the precise data used for the action. For example, the placement of students into groups for differentiation is a group level action because the data from the student group is used to perform this action. Purposes refer to what data is used for. For example, determining individual students' performance compared to the national performance is meant to appraise student performance without setting new learning goals. Therefore, this action is labelled as having an evaluative purpose. Although this study is primarily aimed at actions that support learning, the purposes of communicating and evaluating learning were added in order to gain a better understanding of total actual use.

The grid containing possible actions is based on the questionnaire of Blok, Otter and Roeleveld (2001). They collected a list of actions for which data from various tests could be used. We validated the actions from the grid (Table 3.2) by asking two educational consultants, specialized in assessments, to generate as many actions as they could think of for which teachers, internal coaches and principals would like to use test results. These two educational consultants have considerable contact with all the various user groups about their desired and actual uses of tests. The mentioned actions were already included in the questionnaire. We also asked three teachers to describe what is meant by each action or to give an example of an action from their own practice. We concluded that the descriptions of the actions were clear and appropriate to the Dutch context. We pretested the questionnaire by asking five teachers to fill out the questionnaire and to indicate whether they were missing some actions or whether some questions were unclear. This resulted in a few minor adaptations such as the addition of the option "no use of test results". This option ensured a distinction between respondents who did not answer the question because they

skipped it and those who did not make use of test results. Furthermore, we added the option “other” so that the respondents could mention actions outside the list.

The teachers, internal coaches and principals were asked to complete the questionnaire through various channels. Those schools opting to participate as a focus group also received an e-mail with a link to the questionnaire, which was distributed within the school. The questionnaire was filled out electronically by the respondents.

Table 3.2 Grid consisting of three levels and three purposes of possible actions for which test results can be used

Levels	Purposes		
	Communicating learning	Supporting learning	Evaluating learning
Individual	To inform parents/guardians during individual meetings or by means of score reports.	To create individual action plans for low performing students.	To determine individual students' performance compared to the national performance.
	To inform the individual student about his/her performance.	To create individual action plans for high performing students.	To determine individual students' progress regarding learning goals or content.
	To inform other schools about an individual student by means of an educational report (school transition of the student).	To give feedback to students in order to formulate their own learning goals.	To make decisions about students' transition year.
Group	To inform parents/guardians during group meetings.	To create group action plans.	To determine the group performance compared to the national performance.
	To inform the student group about their performance.	To adapt instruction to educational needs.	To determine group progress regarding learning goals or content.
	To inform colleagues about the student group during a group discussion or transmission.	To place students into different groups for differentiation.	To compare parallel groups regarding their progress.
School	To inform people via the school prospectus or school website.	To create school or annual plans.	To determine the school's performance compared to the national performance.
	To inform the school board or participation council .	To formulate policy regarding the selection of new teaching methods.	To determine progress regarding school goals.
	To inform the inspectorate.	To create professional plans (performance appraisals, career decisions).	To compare the performance of a student group with (those of) previous years.

### *Questionnaire 2 - parents*

The questionnaire used for the parents addressed the perspective of parents on the use of test results. Alongside six items about parents' background, the questionnaire consisted of four items. This article addresses only the two items relating to the purpose of using tests and actions aimed at supporting the learning process. The first item was a multiple choice question in which parents had to select the purpose (communicating, supporting or evaluating learning) that best suited the reason for which they thought test results would be mostly used. The second item was a question investigating the supporting actions parents would like to take in order to determine the extent to which these actions differed from the actions of other users. The questionnaire is included in Table A.2 of Appendix A.

We pretested the questionnaire by asking three parents to fill out the questionnaire and to indicate whether some questions were unclear. The parents reported that all questions were clear, which was also demonstrated in the responses they provided. Therefore, the pretest did not result in adaptations to the questionnaire.

Parents were asked to complete the questionnaire using various channels. Some schools agreed to participate as a focus group and distributed the questionnaire to the parents of their school. Another example was a call on an educational website for parents. The questionnaire was filled out electronically.

### *Focus group - teachers*

Focus group meetings with the teachers of seven participating schools were held to validate the results of research question 1 and to identify the information needs for question 3. The design of the focus group method included the characteristics of a group interview as well as a group discussion (Newby, 2010).

The meeting consisted of three parts. In the first part, the results from the questionnaire were validated by identifying the purpose of teaching and the corresponding actions aimed at achieving this purpose. While the questionnaire was a reactive task whereby respondents were asked to select actions from the given list, the focus group was a generating task whereby respondents were asked to identify the actions by themselves. In the second part, the researcher discussed some conceptual aspects of formative assessments in order to achieve the same understanding of the concept. In the third part, the teachers were to select the actions from the first part for which they needed information from

formative assessments. Actions that did not require information were not selected. To illustrate, the teachers selected the action “placement of students into different instruction groups for differentiation” and not the action “using humor” because they needed some information for the first action but not for the second. Thereafter, the teachers had to think about the information needed for each action. All individual answers were recorded on paper. The teachers’ responses were then systematically grouped and validated during the focus group.

The structure of the focus group meeting and the formulation of the questions were first pretested using individual interviews with two respondents who did not participate in the focus groups. These interviews resulted in some adaptations regarding the formulation of the questions; for example, we changed the following question “which information from tests do you need in order to carry out this action?” into “which information do you need in order to carry out this action?” This is because the pretest showed that respondents only gave answers about information that known test reports are able to give, a mindset deemed too limited for this study. In addition, the first focus group was meant as a trial. Since no changes were made afterwards, the data from this focus group were included.

### 3.2.2 Data Analysis

#### *Questionnaire 1 - teachers, internal coaches and principals*

To answer the first question, frequency analyses were used to show the number of occurrences of each response chosen by the respondents. Because the questionnaire included two multiple response questions, the number of responses differed from the number of respondents. We used McNemar’s test to ascertain statistical differences between actions in terms of actual and desired use. This test evaluates the difference between two correlated proportions, which means that the two scores are not independent. In order to describe the patterns in the answers, the number of actions was then summarized into the three purposes: communicating, supporting and evaluating learning. These “purpose” subscales all had high reliabilities relating to desired use, with Cronbach’s  $\alpha$  of .80, .81 and .86, respectively. Regarding actual use, the reliabilities of these subscales were moderate, with Cronbach’s  $\alpha$  of .59, .65 and .74, respectively. For the second question, the same analyses were performed, but we divided the respondent group into teachers, internal coaches and parents. In addition, the number and

percentages of actions were summarized into three levels: individual, group and school. Since there were no additional actions mentioned under the option “other” from outside the given list, we did not analyse these answers.

#### *Questionnaire 2 - parents*

The data from the parent questionnaire were analysed both qualitatively and quantitatively. The quantitative analysis consisted of frequency analyses of those questions with a closed-answer format. The answers to the open-answer question were coded in a qualitative way. We compared the answers of the open-answer questions and then grouped related pieces of information into categories. We subsequently used these categories to classify all answers. If an answer did not fit into the existing categories, the framework was modified and the process repeated.

#### *Focus group - teachers*

The participants' responses to the questions in the focus group meetings were listed, grouped and documented during the focus group meeting. For the analyses, answers were considered irrelevant and were removed if they did not correspond with actions and needed information. For example, some teachers mentioned a method of testing (e.g. doing an observation) or some preconditions regarding teaching their students (e.g. to create an orderly group climate). Following the focus group method, all relevant data were regarded as valuable, regardless of how many teachers appointed the data. The results of the different focus groups were summarized and compared.

### 3.2.3 Sample Characteristics

#### *Questionnaire 1 - teachers, internal coaches and principals*

A total of 140 teachers, 34 internal coaches and 14 school principals filled out the questionnaire. Of these 188 respondents, 30 respondents did not complete the questionnaire, which means that the responses of 158 respondents were used for analysis. Some background characteristics of the 158 respondents are presented in Table 3.3. The sample characteristics are typical of the Dutch primary school teacher population ([www.onderwijsincijfers.nl](http://www.onderwijsincijfers.nl)).

Table 3.3 Background characteristics of respondents (N=158) from questionnaire 1

	Teachers	Internal coaches	Principals	Total
Sex				
Male	19	1	6	26
Female	100	26	6	132
Average age (SD)	40.4 (11.3)	45.9 (8.8)	49.8 (10.4)	49.8 (10.4)
Years of total experience				
0-5	17	3	0	20
5-10	26	1	0	27
10<	76	23	12	111
Total	119	27	12	158

#### *Questionnaire 2 - parents*

Altogether, 250 parents of students in primary education participated in this study. However, 33 parents did not complete the questionnaire, which means that the responses of 217 parents (48 males, 169 females) were used for analysis. The distribution relating to the educational level of the respondents was 60% completing higher education, 29% completing vocational education, and 11% had obtained a lower educational level. Overall, the sample included a relatively high proportion of female and highly educated respondents compared to the population of parents in the Netherlands.

*Focus group - teachers*

Focus groups were held at seven different schools. All teachers within a school participated in the corresponding focus group. We could therefore ensure that the data were gathered from enthusiastic teachers as well as those who were not very enthusiastic about using tests. To further enhance the representativeness, we selected schools of different sizes. The school teams varied between seven and 17 persons. In total, 84 teachers participated in the seven focus groups. We have no reason to believe that this sample does not reflect the characteristics of the school population.

### 3.3 RESULTS

#### *3.3.1 Question 1: Which Types of Actions Would Users Choose as Desired Uses of Test Results and How Do These Actions Relate to Actual Uses?*

The 158 respondents indicated 1,922 actions as desired uses of test results (Table 3.4), representing, on average, more than 12 actions per respondent. The most frequently chosen action under desired use was “to inform parents during individual meetings or by means of score reports” selected by 121 respondents (76.6%). This action accounted for 6.3% of all the desired use answers. Informing parents was also the most frequently chosen action under actual use (91.1%). However, this action was selected significantly less frequently as a desired action than as an actual use ( $\chi^2 = 14.7$ ;  $p < .001$ ). Communications to the inspectorate, the creation of group plans and some actions relating to the evaluation of test results were also selected significantly less frequently as desired use than as actual use. The creation of group plans was still the second most frequently chosen desired use action (72.2%).

Some actions were chosen significantly more often as desired uses than as actual uses. For example, the frequency of the action “to give feedback to students in order to formulate their own learning goals” doubled ( $\chi^2 = 46.2$ ;  $p < .001$ ) from 19.6% to 51.3%. Other examples included the creation of action plans for high performing students ( $\chi^2 = 5.8$ ;  $p = .02$ ) and the formulation of policy regarding the purchase of teaching methods and instruments ( $\chi^2 = 19.2$ ;  $p < .01$ ).



Table 3.4 Number of responses and respondents (N=158) choosing an action as desired use and actual use

Actions	Desired use				Actual use				
	Responses n	%	Respondents %	Responses n	Responses n	%	Respondents %	Responses n	%
Communicating learning									
* To inform parents/guardians during individual meetings or by means of score reports	121	6.3	76.6	144	7.2	91.1			
* To inform the individual student about his/her performance	83	4.3	52.5	68	3.4	43.0			
To inform other schools about an individual student by means of an educational report (school transition of the student)	89	4.6	56.3	102	5.1	64.6			
To inform parents/guardians during group meetings	29	1.5	18.4	27	1.4	17.1			
* To inform the student group about their performance	39	2.0	24.7	22	1.1	13.9			
To inform colleagues about the student group during a group discussion or transmission	100	5.2	63.3	113	5.7	71.5			
To inform people via the school prospectus or school website	15	0.8	9.5	13	0.7	8.2			
To inform the school board or participation council	32	1.7	20.3	42	2.1	26.6			
* To inform the inspectorate	46	2.4	29.1	88	4.4	55.7			
Supporting learning									
To create individual action plans for low performing students	111	5.8	70.3	115	5.8	72.8			
* To create individual action plans for high performing students	104	5.4	65.8	86	4.3	54.4			
* To give feedback to students in order to formulate their own learning goals	81	4.2	51.3	31	1.6	19.6			
* To create group action plans	114	5.9	72.2	137	6.9	86.7			
To adapt instruction to educational needs	101	5.3	63.9	91	4.6	57.6			

Table 3.4 (continued)

Actions	Desired use						Actual use					
	Responses			Respondents			Responses			Respondents		
	n	%	%	n	%	%	n	%	%	n	%	%
To place students into different groups for differentiation	108	5.6	68.4	118	5.9	74.7						
To create school or annual plans	39	2.0	24.7	32	1.6	20.3						
* To formulate policy regarding the selection of a new teaching method	63	3.3	39.9	32	1.6	20.3						
To create professional plans (performance appraisals, career decisions)	20	1.0	12.7	14	0.7	8.9						
Evaluating learning												
* To determine the individual students' performance compared to the national performance	67	3.5	42.4	97	4.9	61.4						
To determine individual students' progress regarding learning goals or content	107	5.6	67.7	118	5.9	74.7						
* To make decisions about students' transition year	64	3.3	40.5	82	4.1	51.9						
* To determine the group performance compared to the national performance	71	3.7	44.9	89	4.5	56.3						
To determine group progress regarding learning goals or content	92	4.8	58.2	90	4.5	57.0						
To compare parallel groups regarding their progress	33	1.7	20.9	28	1.4	17.7						
To determine the school's performance compared to the national performance	58	3.0	36.7	55	2.8	34.8						
To determine progress regarding school goals	62	3.2	39.2	66	3.3	41.8						
* To compare the performance of a student group with (those of) previous years.	65	3.4	41.1	89	4.5	56.3						
No use of test results	8	0.4	5.1	2	1.0	1.3						
Total	1,922	100		1,991	100							

Note: \*p&lt;.05

Other frequently mentioned actions as desired uses, although they were not chosen significantly more frequently, were the creation of individual action plans for low performing students (70.3%) and the placement of students into groups for differentiation (68.4%). These actions, including the creation of group action plans (72.2%), were all examples of actions relating to the category of supporting learning.

Table 3.5 presents a summary of the number and percentages of actions into the three purposes: communicating, supporting and evaluating learning. The action “no use of test results” was a separate category that did not belong under any of the other three purposes. Notwithstanding the fact that fewer actions were selected as desired use ( $n = 1,922$ ) in comparison with actual use ( $n = 1,991$ ), the number of actions relating to supporting learning was higher for desired use ( $n = 741$ ) than for actual use ( $n = 656$ ). The opposite was true for the purposes of communicating and evaluating learning. Regarding the relative distribution of desired use, respondents mostly chose actions relating to supporting learning (38.6%). This result differed from actual uses whereby actions relating to the purpose of evaluating learning were most commonly chosen (35.9%), and actions relating to the purpose of supporting learning were chosen less frequently (32.9%).

Table 3.5 Number and percentage of actions relating to desired and actual use per purpose

Purpose	Desired use		Actual use	
	Count	%	Count	%
Communicating learning	554	28.8	619	31.1
Supporting learning	741	38.6	656	32.9
Evaluating learning	619	32.2	714	35.9
No use of test results	8	0.4	2	0.1
Total	1,922	100	1,991	100

The results shown in Table 3.5 were confirmed by the answers on the third questionnaire item, which required respondents to choose the most important action as a desired use of tests. In total, 53.9% of the respondents chose an action relating to the purpose of supporting learning. The most frequently chosen action in this category was “to adapt instruction to educational needs” ( $n = 25$ ), followed by “to give feedback to students in order to formulate their own learning goals” ( $n = 18$ ) and “to create group action plans” ( $n = 16$ ). Actions relating to the purpose of evaluating learning were chosen by 27.5% of the respondents. This result was mainly due to the action “to determine individual students’ progress regarding learning goals or content” ( $n = 29$ , 17.4%).

The view of parents corresponded with this result; 45.2% of them indicated that test results were mainly used to support their child’s learning. This was followed by 40.1% of parents, who thought that student-level evaluation was the main goal, and 14.7% who said that communicating results to the parents, principal or inspectorate was the central goal.

Based on these results, we conclude that respondents mostly chose actions relating to the purpose of supporting learning, which amounts to a discrepancy relating to actual use. In order to create useful score reports of test results, we investigated whether these actions were the same or different for the various audiences.

### *3.3.2 Question 2: What, if Any, Is the Extent of the Differences Between Teachers, Internal Coaches, Principals and Parents With Regard to Desired and Actual Uses and Corresponding Actions?*

Table B.1 in Appendix B presents the percentages of teachers, internal coaches and principals choosing an action for desired and actual use. No major differences were found with regard to the three purposes of communicating, supporting and evaluating learning (Table 3.6). Teachers and principals mostly chose actions as desired use under the purpose of supporting learning while internal coaches chose almost as many actions for the purpose of supporting learning as for the purpose of evaluating learning. All user groups indicated that current test results were primarily used to evaluate learning.

Table 3.6 Response percentages of actions chosen by users in relation to the different purposes

Purpose	Desired use			Actual use		
	Teachers (n=119)	Internal coaches (n=27)	Principals (n=12)	Teachers (n=119)	Internal coaches (n=27)	Principals (n=12)
Communicating learning	28.6	28.5	31.5	31.5	28.6	33.7
Supporting learning	39.8	35.3	37.6	33.8	31.0	31.1
Evaluating learning	31.1	36.0	30.4	34.6	40.4	35.2
No use of test results	0.5	0.2	0.6	0.1	0.0	0.0
Total	100	100	100	100	100	100

There were, however, differences between the user groups with regard to the different levels of actions (Table 3.7). The teachers especially selected actions relating to the individual level (45.5%) and subsequently chose many actions relating to the group level (37.1%). Only a small number of teachers' responses represented actions relating to the school level (16.9%). The answers furnished by the internal coaches showed a similar pattern although they had a greater preference than teachers to perform some actions at the school level (25.9%). The principals' answers showed the opposite, with most selected actions relating to the school level (35.9%). To illustrate this difference, at the school level, the development of school plans was selected far more frequently by principals (91.7%) than by teachers (13.4%) and internal coaches (44.4%).

Table 3.7 Response percentages of actions chosen by users in relation to the different levels

Level	Desired use			Actual use		
	Teachers (n=119)	Internal coaches (n=27)	Principals (n=12)	Teachers (n=119)	Internal coaches (n=27)	Principals (n=12)
Individual	45.5	39.9	33.1	44.5	40.7	31.2
Group	37.1	34.0	30.4	38.1	32.2	28.1
School	16.9	25.9	35.9	17.3	27.1	40.7
No use of test results	0.5	0.2	0.6	0.1	0	0
Total	100	100	100	100	100	100

Parents ( $N = 217$ ) mentioned also actions relating to supporting the learning process. Helping their child with homework was, for example, the most mentioned action (19.0%). Furthermore, 17.2% of the parents would like to practice the learning material with their child at home. Some parents would give their child additional support by providing learning material to remedy weaknesses (14.5%). Other examples of actions mentioned included reading books (9.3%), testing their child on his/her knowledge for a test (7.9%), learning in a playful way (4.5%), helping to develop learning skills like planning school work (4.1%), giving some educational games (3.4%) and visiting cultural organizations like museums (3.4%). All such actions were in relation to their own individual child.

These differences in actions between the various users indicated that there is a need for score reports to be tailored to the specific user groups, corresponding to the actions that these kinds of users would like to undertake (Zapata-Rivera & Katz, 2014). This means that we should investigate the information needs of each user group separately. Based on the previous finding that test results would rather be used to support learning, we decided to limit our focus on teachers for the third question. Teachers' primary task was to support the learning process of students. They were also the users who actually communicated these results to other users such as students and their parents.

### *3.3.3 Question 3: What Information From Test Results Is Needed to Perform the Desired Actions?*

The results of research question 1 were validated during the seven focus group meetings. The teachers in the focus groups underlined the general principle that they would support student learning by developing the cognitive and social knowledge and skills of their students.

Subsequently, the actions for achieving this purpose were generated. Nine actions in the questionnaire were related to the purpose of supporting learning. The most frequently chosen actions in the questionnaire were also generated by the teachers in the focus groups. Starting with the most frequently mentioned action, these actions were: (1) alignment of learning material and learning objectives with the initial level of students, (2) placement of students into different instruction groups for differentiation, (3) student-teacher conversations about well-being and learning, (4) development of group and individual action plans and (5) alignment of learning materials to learning objectives and preferences, with action (4) from

the focus group covering three actions from the questionnaire. Two actions from the questionnaire were not mentioned in the focus group. However, these actions were also chosen less frequently by teachers but more often by principals and internal coaches. As the actions were formulated in the focus groups, the conceptualizations of these actions were slightly different from the description in the questionnaire. For example, student-teacher conversations about well-being and learning was related to the questionnaire action “to give feedback to students in order to formulate their own learning goals”.

Table 3.8 presents the five actions and the corresponding information needs of each action. The action mentioned by all seven focus groups was the alignment of learning materials and learning objectives with the initial level of students, which corresponds with the questionnaire item “to adapt instruction to educational needs”. In order to perform this action, teachers need information about the learning objectives for each year and subject as well as information about students’ mastery of these learning objectives. Furthermore, information is needed with regard to the sequence of acquiring learning objectives, realistic expectations for the next learning objective and learning material suggestions of how to achieve this objective.

Most of the information needs mentioned for student-teacher conversations were about well-being and learning. For this action, teachers need information about students’ learning in order to give students feedback, for example, information about students’ strategy to solve assignments. Furthermore, teachers need information about students’ personal aspects, like students’ well-being, working attitude and self-efficacy.

The overall results presented in Table 3.8 show that teachers have different information requirements for performing actions: on one hand, information about general teaching aspects like the learning objectives for each year and subject and, on the other hand, information about students such as their learning progress. Furthermore, teachers need information about both the personal aspects of students, like their interest in subjects, and their cognitive aspects, such as their mastery of a learning objective.

These results also indicate that teachers need the same information for different actions. For example, students’ mastery of a learning objective was needed to perform four actions. Realistic expectations regarding subsequent learning objectives were mentioned for three actions. Other kinds of information were only selected for one action, like students’ working attitude.

Table 3.8 Information needs mentioned by focus groups (N=7) to perform actions to support learning

Information needs	Actions				
	1	2	3	4	5
Learning material suggestions					
Learning objectives for each year and subject					
Realistic expectations for next learning objective					
Sequence of acquiring learning objectives					
Starting level of students' knowledge					
Students' interest in subjects					
Students' learning preferences					
Students' learning progress					
Students' mastery of a learning objective					
Students' motivation					
Students' self-efficacy					
Students' strategy to solve assignments					
Students' strong and weak points					
Students' well-being					
Students' working attitude					
Suggestions for placement of students into three groups					

1 = Alignment of learning material/objectives to initial level – questionnaire item: to adapt instruction to educational needs (n = 7); 2 = Differentiation – questionnaire item: to place students into different groups for differentiation (n = 6); 3 = Student-teacher conversations – questionnaires item: to give feedback to students in order to formulate their own learning goals (n = 5); 4 = Group and individual action plans – questionnaire item: to create group/individual action plans for low/high performing students (n = 5); 5 = Alignment of materials to learning objectives/preferences – questionnaire item: to formulate policy regarding the selection of new teaching methods (n = 3).

Because we started the focus group by discussing the actions for the purpose of teaching rather than the more specific actions relating to test results, and because we asked the focus groups for actions requiring information from all possible sources, including tests, teachers also mentioned three actions that were not directly related to the actions listed in the questionnaire: (1) placement of students into different groups for cooperative learning and collaboration, (2) connection to students' perception of the world and (3) creation of ownership. Cooperative learning and collaboration means that a group of students have to cooperate equally on assignments in order to achieve learning goals. For this action, teachers need information about the learning objectives for each



year and subject and about students' mastery of these learning objectives. This is the same information as that mentioned earlier. However, teachers also need information about students' behavior, their social and communicative skills and their willingness to collaborate. The second action concerned the connection to students' perception of the world, which contained the adoption of the chosen examples, thematic topics and the introductions of lessons relating to students' experiences and interest. For this action, teachers also need additional information, such as the dynamic of a student group, the proportion of boys and girls and students' home situation. The third action concerned giving students responsibility to support their own learning. The information requirements included students' persistence and ability to work independently.

### 3.4 CONCLUSIONS AND DISCUSSION

This study investigated the types of actions users want to perform with the use of test results and the information needed to enable these actions. By administering two questionnaires and conducting seven focus group meetings, both qualitative and quantitative data were gathered. In the analyses, distinctions were made among various users, including teachers, internal coaches, principals and parents.

The results of this study suggest that in relation to desired uses, respondents mostly chose actions relating to the purpose of supporting learning. The study also showed that this desired use of test results was not the same as the actual use; test results were primarily used to evaluate the learning process by determining the student's ability. These results corroborate the results of previous studies, suggesting the limited use of test results for formative purposes (Ledoux et al., 2009; Meijer et al., 2011; Vanhoof et al., 2011; Verhaeghe et al., 2011).

Furthermore, we conclude that the various users want to perform actions on different levels and in different contexts. Teachers and parents reported that they want to perform actions at the level of the individual student whereby teachers act in an educational setting and parents perform in a more informal situation. Internal coaches and principals selected more actions relating to the school level. This result is in accordance with the expectation regarding the unique decisions of each user group (Zapata-Rivera & Katz, 2014).

Based on the results of the first and second questions, we decided to limit our third question to teachers. The results from the first question were validated, and we gathered insights about the information needs of teachers to perform each action mentioned. The results show the need for different kinds of information, for instance, relating to students' strategy to solve an assignment, students' motivation and their working attitude. This result confirms Brookhart and Nitko's (2008) and Mandinach's (2012) argument that test data are only one source of information in supporting educational decision-making and that an accurate picture from the student could be obtained with the use of other assessment data. The results also indicate that teachers sometimes need the same information for different actions; for example, information about students' mastery of a learning objective was mentioned for the performance of four actions.

Finally, the formulation of the question "what information do you need in order to carry out this action?" expanded the mindset of respondents but also resulted in information needs which might not arise from tests. For example, the information need "sequence of acquiring learning objectives" likely formed a greater part of the content knowledge of the teacher. This illustrates the view of Gummer and Mandinach (2015) that the process of using test data is complex and that for instructional decision-making, teachers need to combine an understanding of data with "standards, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge, and an understanding of how children learn" (p. 2).

#### *3.4.1 Limitations of the Study*

This study was limited in several ways. First, the sample size was limited (especially the number of principals and internal coaches), so the results can only be generalized to a limited degree. However, the fact that most of the focus groups mentioned the same type of actions, which were also in the questionnaire, suggests that we have identified the most important actions for teachers.

The three additional actions mentioned in the focus groups suggest that the actions list in the questionnaire might have been incomplete. This is because we asked for actions from two different perspectives. We started the focus groups by discussing the actions for the purpose of teaching rather than the more specific information needs relating to test results. Furthermore, we asked the focus groups for actions requiring information from all possible sources, including tests.

Thus, the answers from the focus group included actions from a wider perspective. Moreover, no additional actions were mentioned during the pretesting of the questionnaire; the “other” option in the questionnaire was not used; and two of the additional actions were mentioned by just two focus groups. For this reason, we considered this difference of actions to be of minimal importance.

Finally, the users chose the actions for which they want to use test results, but this choice was made within an existing frame of reference consistent with existing tests in the current national system of the Netherlands. This imagination seems to be difficult and contextualized, which limits generalization.

### *3.4.2 Implications for Practice*

This study provided insights into the actions and corresponding information needs of teachers. The results show that teachers and others would like to use test results for uses for which current measurement instruments are not validated. This may result in misuse or limited use of current test results. The results are informative to teachers and others, in terms of the use of different instruments, for their educational decisions about actions as the information needs of teachers cannot be obtained from one test. Furthermore, test developers could use the insights herein for the development of tests and score reports aimed at teachers. If the score report presents content that is tailored to the actions that teachers would like to undertake, then teachers would always be able to perform the action once the data is collected and presented. For example, teachers can use data from test results to make up different groups for differentiation, to align the learning objective with the initial student level or to develop group action plans. Compared to the available information from current tests, test developers should develop tests that offer more detailed information like student strategies to solve assignments. In this way, test results will be used to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken intuitively in the absence of such data (Black & Wiliam, 2009). The development of tailored score reports can contribute to the potential of formative assessment as a way of supporting student learning.

### *3.4.3 Implications for Future Research*

The results and limitations imply a future research agenda. First, it seems worthwhile to examine the actions and information needs of students as a user group. Since teachers indicated that they would formulate learning objectives together with students more often than they actually did, it would be useful to also examine the actions and information needs of students. This result is in accordance with the trend towards activating students as owners of their own learning as a key strategy of formative assessment (William & Thompson, 2007).

Second, despite the growing body of research on effective score reporting (Zenisky & Hambleton, 2012), there has been little effort on users' actual use of developed score reports. Future research is needed on how to design effective score reports for teachers that visualize test results that are appropriate to the identified actions and information needs reported in this study. Moreover, it would be useful to study the extent to which the presentation of the identified information needs result in more data use for student learning.

Third, the results showed that teachers need detailed information from tests, such as the extent to which each student has mastered a certain learning objective. This implies that score reports should visualize smaller levels of information (e.g. from total test scores to subscores and items). Accuracy is however related to the level of reporting. When reports are more detailed, the accuracy of test scores is often negatively impacted; e.g. accuracy is lower, and scores are more uncertain (Monaghan, 2006; Ryan, 2006). The total score is often a more accurate measure of an individual's knowledge or skills in a subdomain of interest than a subscore derived only from those items that purport to measure the subdomain directly (Monaghan, 2006). Assessment organizations have a duty to provide teachers with sufficient information about these accuracies to allow them to make valid inferences based on test results (e.g. AERA, APA, & NCME, 2014; Newton, 2005). Future research is needed to investigate how to best communicate this accuracy information.

### 3.5 REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for effective score reporting* (CSE Technical Report 326). Los Angeles, CA: UCLA Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Baird, J. A., Hopfenbeck, T. N., Newton, P., Stobart, G., & Steen-Utheim, A. T. (2014). *State of the field review assessment and learning* (case number 13/4697). Oslo: Knowledge Center for Education.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31. doi:10.1007/s11092-008-9068-5
- Blok, H., Otter, M. E., & Roeleveld, J. (2001). *Het gebruik van leerlingvolgsystemen anno 2000* [The use of student monitoring systems in the year 2000]. Amsterdam, the Netherlands: SCO-Kohnstam Instituut.
- Bradshaw, J., & Wheeler, R. (2009). *National foundation for educational research: International survey of results reporting* (OFQUAL 10/4705). London: Office of Qualifications and Examinations.
- Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, NJ: Pearson Education.
- Centraal Bureau voor de Statistiek (CBS), Dienst Uitvoering Onderwijs (DUO) en het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) (n.d.). *Onderwijs in cijfers* (Education in numbers). Retrieved from <http://www.onderwijsincijfers.nl>
- Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston: Harvard Business School Press.
- Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (2016). Effects of a data use intervention on educators' use of knowledge and skills. *Studies in Educational Evaluation*, 48, 19–31. doi:10.1016/j.stueduc.2015.11.002
- 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*, 17(2), 145–220. doi:10.1207/s15324818ame1702
- Gummer, E., & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record*, 117(4), 1–22. Retrieved from <https://www.tcrecord.org>

- Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: UCLA Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. E. Geisinger (Ed.), *Handbook of testing and assessment in psychology* (pp. 479-494). Washington, DC: APA.
- Hattie, J. (2009). *Visibly learning from reports: The validity of score reports*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.
- Hattie, J. A. C., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, 36(2), 189–201. doi:10.2190/ET.36.2.g
- 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. doi:10.1016/j.edurev.2012.09.001
- Jaeger, R. M. (1998). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress* (Working paper 2003-11). Palo Alto, CA: American Institutes for Research.
- Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). *Opbrengstgericht werken: Over de waarde van meetgestuurd onderwijs* [Data-driven decision making: About the value of measurement oriented education]. Amsterdam, the Netherlands: SCO-Kohnstam Instituut.
- Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, 23(2), 26–32. doi:10.1111/j.1745-3992.2004.tb00156.x
- Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, 47(2), 71–85. doi:10.1080/00461520.2012.667064
- Mandinach, E. B., & Jackson, S. S. (2012). *Transforming teaching and learning through data-driven decision making*. Thousand Oaks, CA: Corwin.
- Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, 114(11), 1–48. Retrieved from <http://www.rdc.udel.edu>
- Meijer, J., Ledoux, G., & Elshof, D. (2011). *Gebruikersvriendelijke leerlingvolgsystemen in het primair onderwijs* [User-friendly pupil monitoring systems in primary education] (Report 849). Amsterdam: Kohnstamm Instituut.
- Monaghan, W. (2006). *The facts about subscores* (Report No. RDC-04). Princeton, NJ: Educational Testing Service.

- NEGP. (1998). *Talking about tests: An idea book for state leaders*. Washington, DC: U.S. Government Printing Office.
- Newby, P. (2010). *Research methods for education*. Harlow, UK: Longman.
- Newton, P. E. (2005). The public understanding measurement inaccuracy. *British Education Research Journal*, 31(4), 419–442. doi:10.1080/01411920500148648
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, 48(1), 4–11. doi:10.1080/00405840802577536
- Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Mahwah, NJ: Lawrence Erlbaum.
- 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. doi:10.1016/j.tate.2009.06.007
- Schildkamp, K., & Lai, M. K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 9–21). Dordrecht: Springer. doi:10.1007/978-94-007-4816-3\_1
- 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. doi:10.1080/13803610802048874
- 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. doi: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. doi:10.1016/j.stueduc.2014.04.004
- Van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. (2016). Changes in educators' data literacy during a data-based decision making intervention. In Keuning, T. & Van Geel, M. J. M. (red.), *Implementation and effects of a schoolwide data-based decision making intervention: A large skill study* (pp. 73–96). Enschede, The Netherlands: Universiteit Twente.
- 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. doi:10.1080/03055698.2010.482771
- Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2011). Effecten van ondersteuning bij schoolfeedbackgebruik [Effects of support in school feedback use]. *Pedagogische Studien*, 88(2), 90–106.

- Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36(4), 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x
- William, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.
- William, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Erlbaum.
- Zapata-Rivera, J. 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 & Practice*, 21(4), 442–463. doi:10.1080/0969594X.2014.936357
- Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (ETS RM-12-20). Princeton, NJ.
- Zapata-Rivera, D., & Zwick, R. (2011). *Test score reporting: Perspectives from the ETS score reporting conference* (ETS RR-11-45). Princeton, NJ.
- Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, 31(2), 21–26. doi:10.1111/j.1745-3992.2012.00231.x
- Zwick, R., Sklar, J. C., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice*, 27(2), 14–27. doi:10.1111/j.1745-3992.2008.00119.x



## APPENDIX A: QUESTIONNAIRES

Table A.1 Questionnaire 1

---

Respondents' background

1. What is your position within the school?
  - a) Teacher; b) Internal coach; c) Principal
2. What is your sex?
  - a) Male; b) Female
3. What is your age?
4. In which district do you work?
  - a) Friesland; b) Groningen; c) Drenthe; d) Overijssel; e) Gelderland; f) Limburg; g) Noord-Brabant; h) Zeeland; i) Utrecht; j) Noord-Holland; k) Zuid-Holland; l) Flevoland
5. How many years' experience do you have in primary education? (For internal coaches/principals only)
  - a) Less than 5 years; b) 5 to 10 years; c) More than 10 years
6. How many years have you worked as a teacher/internal coach/principal?
  - a) Less than 5 years; b) 5 to 10 years; c) More than 10 years
7. Which grade do you teach the most? (Note: Answer the remaining questions for the students in this grade) (For teachers only)
  - a) Lowest grades (Group 1,2); b) Middle grades (Group 3,4,5); c) Upper grades (Group 6,7,8)
8. In which grade do you function as internal coach? (For internal coaches only)
  - a) Lowest grades (Group 1,2); b) Middle grades (Group 3,4,5); c) Highest grades (Group 6,7,8); d) Whole school (Group 1-8). Other:
9. What, if any, other functions do you fulfill in school? (For internal coaches/principals only)
  - a) No other functions; b) Principal; c) Teacher; d) ICT coordinator; e) Language or math specialist; f) Otherwise, namely...
10. Does your school use any of the following principles?
  - a) Anthroposophy; b) Dalton; c) Freinet; d) Jenaplan; e) Montessori; f) O4NT; g) none of these
11. Which of the following student monitoring systems do you use?
  - a) Cito-LVS; b) Parnassys; c) Esis; d) Dot.com; e) Other:

---

Actual use of test results

12. Test results may be used for different actions. We presented a number of actions below. Check the purposes for which you have used test results in recent school years (2014-2015, 2015-2016). Note: we mean the use of test results for actions with the majority of students, not for exceptional circumstances. It is possible to give more than one answer. In the last school years, I have used test results for the following actions...

---

---

 Desired use of test results

13. Suppose you were allowed to design test score reports yourself so that it presents you with all the information you need, check for which actions you would like to use test results. Note: we mean the use of test results for actions with the majority of students, not for exceptional circumstances. It is possible to give more than one answer. In an ideal situation, I would like to use test results for the following actions ...

---

 Most important purpose of desired use

14. We presented your chosen actions regarding test results below. Which action do you find most important in the use of test results? Chose the most important one.

---

 For question 12, 13 and 14, we presented the following action list:

No use of test results	To adapt instruction to educational needs
To inform parents/guardians during individual meetings or by means of score reports	To determine group progress regarding learning goals or content
To create individual action plans for low performing students	To inform the school board or participation council
To determine individual students' performance compared to the national performance	To formulate policy regarding the selection of new teaching methods
To inform parents/guardians during group meetings	To determine progress regarding school goals
To create group action plans	To inform other schools about an individual student by means of an educational report (school transition of the student)
To determine the group performance compared to the national performance	To give feedback to students in order to formulate their own learning goals
To inform people via the school prospectus or school website	To make decisions about students' transition year
To create school or annual plans	To inform colleagues about the student group during a group discussion or transmission
To determine the school's performance compared to the national performance	To place students into different groups for differentiation
To inform the individual student about his/her performance	To compare parallel groups regarding their progress
To create individual action plans for high performing students	To inform the inspectorate
To determine individual students' progress regarding learning goals or content	To create professional plans (performance appraisals, career decisions)
To inform the student group about their performance	To compare the performance of a student group with (those of) previous years.

---

Table A.2 Questionnaire 2

---

 Respondents' background

1. What is your sex?  
a) Male; b) Female
  2. What is your age?
  3. In which district do your children attend school?  
a) Friesland; b) Groningen; c) Drenthe; d) Overijssel; e) Gelderland; f) Limburg; g) Noord-Brabant; h) Zeeland; i) Utrecht; j) Noord-Holland; k) Zuid-Holland; l) Flevoland
  4. What is your highest level of education?  
a) No education/primary education; b) preparatory secondary vocational education; c) general secondary education; d) vocational education; e) senior general secondary education/university preparatory education; f) university of applied sciences; g) Master of Arts/Science/PhD
  5. In which grade is your oldest child?  
a) Lowest grades (Group 1,2); b) Middle grades (Group 3,4,5); c) Upper grades (Group 6,7,8); d) My oldest child has left primary school
  6. Does your school use any of the following principles?  
a) Anthroposophy; b) Dalton; c) Freinet; d) Jenaplan; e) Montessori; f) O4NT; g) none of these
- 

## Central questions

7. Which of the following purposes do you think best suits the reason to test your child at school?  
a) Determining the level of your child; b) Adapting instruction to the educational needs of your child; c) Reporting and communicating the results to parents, the school board or inspectorate
  8. Do you receive the score reports of your child from the student monitoring system?  
a) No; b) Yes, in the score report of my child; c) Yes, during individual meetings with the teacher; d) Yes, during group meetings with parents; e) Other:
  9. Would you like to support the learning process of your child?  
a) No, in my opinion, this task belongs to the school; b) Yes, by means of the following actions....
  10. This research looks at how test results are presented. What information from test results would you like to receive about your child? It is possible to give more than one answer.  
a) The test scores of my child, focusing especially on the different subjects; b) The progress of my child with regard to the different subjects; c) The level at which my child is, compared to that of other children, with regard to the different subjects; d) The level at which my child is with regard to the different parts of a subject; e) Learning material suggestions with regard to the different subjects in order to help my child in his/her learning; f) Other:
-

## APPENDIX B: ACTUAL AND DESIRED USE

Actions	Desired use			Actual use		
	Teachers n=119	Internal coaches n=27	Principals n=12	Teachers n=119	Internal coaches n=27	Principals n=12
Communicating learning						
To inform parents/guardians during individual meetings or by means of score reports	77.3	77.8	66.7	92.4	92.6	75.0
To inform the individual student about his/her performance	47.9	74.1	50.0	40.3	48.1	58.3
To inform other schools about an individual student by means of an educational report (school transition of the student)	52.1	77.8	50.0	60.5	77.8	75.0
To inform parents/guardians during group meetings	16.0	22.2	33.3	18.5	7.4	25.0
To inform the student group about their performance	20.2	37.0	41.7	14.3	7.4	25.0
To inform colleagues about the student group during a group discussion or transmission	61.3	77.8	50.0	73.9	70.4	50.0
To inform people via the school prospectus or school website	5.0	14.8	41.7	4.2	7.4	50.0
To inform the school board or participation council	8.4	44.4	83.3	14.3	48.1	100.0
To inform the inspectorate	20.2	55.6	58.3	46.2	77.8	100.0
Supporting learning						
To create individual action plans for low performing students	66.4	85.2	75.0	73.9	77.8	50.0
To create individual action plans for high performing students	63.0	81.5	58.3	53.8	63.0	41.7
To give feedback to students in order to formulate their own learning goals	45.4	77.8	50.0	16.0	29.6	33.3

Table B.1 (continued)

Actions	Desired use			Actual use		
	Teachers n=119	Internal coaches n=27	Principals n=12	Teachers n=119	Internal coaches n=27	Principals n=12
To create group action plans	71.4	77.8	66.7	89.1	85.2	66.7
To adapt instruction to educational needs	57.1	85.2	83.3	50.4	77.8	83.3
To place students into different groups for differentiation	68.1	74.1	58.3	75.6	77.8	58.3
To create school or annual plans	13.4	44.4	91.7	11.8	25.9	91.7
To formulate policy regarding the selection of a new teaching method	36.1	51.9	50.0	15.1	33.3	41.7
To create professional plans (performance appraisals, career decisions)	9.2	18.5	33.3	5.9	3.7	50.0
Evaluating learning						
To determine the individual students' performance compared to the national performance	35.3	66.7	58.3	55.5	85.2	66.7
To determine individual students' progress regarding learning goals or content	66.4	77.8	58.3	73.9	85.2	58.3
To make decisions about students' transition year	37.8	55.6	33.3	48.7	63.0	58.3
To determine the group performance compared to the national performance	37.0	77.8	50.0	48.7	77.8	83.3
To determine group progress regarding learning goals or content	54.6	77.8	50.0	57.1	59.3	50.0
To compare parallel groups regarding their progress	15.1	44.4	25.0	14.3	29.6	25.0
No use of test results	5.0	8.3	5.1	1.7	0	0





# CHAPTER 4





## THE VISUAL PRESENTATION OF MEASUREMENT ERROR

This study investigated (1) the extent to which presentations of measurement error in score reports influence teachers' decisions and (2) teachers' preferences in relation to these presentations. Three presentation formats of measurement error (blur, colour value, and error bar) were compared to a presentation format that omitted measurement error. The results from a factorial survey analysis showed that the position of a score in relation to a cut-off score impacted most significantly on decisions. Moreover, the teachers ( $N = 337$ ) indicated the need for additional information significantly more often when the score reports included an error bar compared to when they omitted measurement error. The error bar was also the most preferred presentation format. The results were supported in think-aloud protocols and focus groups, although several interpretation problems and misconceptions of measurement error were identified.

Keywords: assessment; test score reports; measurement error; educational decision making; preferences

This chapter was previously published as:

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. doi:10.1080/0969594X.2018.1447908

## 4.1 INTRODUCTION

In education, decision-making is an everyday activity. For example, teachers make decisions about the next steps in instruction, the placement of students into different instruction groups or the need to provide a student with additional support. Since these decisions may have serious consequences for teaching and learning, they need to be informed by high-quality evidence (Brookhart & Nitko, 2008).

Several data sources can be used to inform decisions, such as student observations, oral questions, students' work, parental reports and test scores (Brookhart & Nitko, 2008; Mandinach, 2012). Due to a careful construction process, test scores are often considered a valuable source. In general, these scores are regarded as very reliable and non-biased (Shepard, 2006); however, they are also subject to a certain amount of measurement error (Gardner, 2013).

Measurement error (ME) can be conceptualised as the difference between a student's actual or obtained score and the theoretical true score counterpart (Gardner, 2013). Feldt and Brennan (1989) list four categories of ME: (a) inherent variation in human performance, (b) variations in the environment within which the measurements are obtained, (c) variations in the evaluation of responses and (d) variation arising from the selection of the test items asked. In practice, different measures are used to quantify ME, including the standard error of measurement, the standard error of estimation and the test information function, depending on the measurement model being used.

As some degree of ME is common to all tests, corroboration between the test score and additional data sources is often recommended (AERA, APA, & NCME, 2014). This recommendation is even more important if the test score contains a relatively large ME or if it, along with its ME, is positioned around the cut-off score of high-stakes decisions that cannot easily be reversed (AERA et al., 2014). High-stakes decisions may trigger major consequences for students, for example, students might not be assigned to an appropriate instruction group and, thus, might not get the instruction they need (e.g. Goodman & Hambleton, 2004; Newton, 2005; Phelps, Zenisky, Hambleton, & Sireci, 2010). When combined with other data sources that are potentially more authentic, such as student observations or other test scores, a more accurate picture of the student can be obtained, and decisions can be better informed (Brookhart & Nitko, 2008; Mandinach, 2012).

The extent to which ME around test scores influences teachers' educational decisions is hitherto unknown. This influence, however, determines the usefulness of displaying ME. On one hand, confusion around the concept of ME could result in misinformed decisions with adverse consequences for students. Several studies indicate some misunderstanding by teachers around the interpretation of ME visualisations (e.g. Impara, Divine, Bruce, Liverman, & Gay, 1991; Zwick, Zapata-Rivera, & Hegarty, 2014). Considering the possible consequences of misinformed decisions, test designers would avoid the presentation of ME in score reports (Bradshaw & Wheeler, 2009; Epp & Bull, 2015) or would place this information in the technical manuals (Phelps et al., 2010). On the other hand, the lack of ME reporting could be a serious problem as teachers would interpret test scores more accurately than they might be. Therefore, test publishers would have a duty to provide teachers with error information that would allow them to make valid inferences based on test results (AERA et al., 2014). Although teachers may not have a full understanding of the nature of ME, the presentation of ME could lead to greater awareness about the imprecision around test scores compared to a score report that omits ME. This awareness could stimulate teachers to gather additional information about a student's ability, resulting in more informed decisions.

This study investigated the extent to which various presentation formats of error information influence teachers' decisions within the context of primary education. Specifically, we examined the extent to which the ME presentation formats result in the need to gather additional information to enable decision-making regarding students, for example, from other information sources. The need for additional information was defined as an indication for awareness of ME. Furthermore, we investigated teachers' own perspectives on the presentation of ME. We asked teachers about their preference levels for each presentation format since several studies have suggested that user preference and performance did not always coincide (e.g. Wainer, Hambleton, & Meara, 1999; Zwick et al., 2014). Teachers' decisions and preferences were examined in the context of a familiar type of action: the assignment of students into instruction groups. Two research questions were formulated:

1. To what extent do various ME presentation formats result in teachers' need for additional information compared to a presentation format that omits ME?
2. Which of the various presentation formats do teachers prefer?

## 4.2 THE PRESENTATION OF MEASUREMENT ERROR

The presentation of error information has received growing attention across a range of disciplines outside the field of education (e.g. Brodlie, Osorio, & Lopes, 2012; Kinkeldey, MacEachren, Riveiro, & Schiewe, 2015), resulting in the development of many potential visual tools for presenting ME. To help designers choose a presentation format, various taxonomies have been proposed (e.g. Gershon, 1998; Pang, Wittenbrink, & Lodha, 1997), and several review studies have been conducted (e.g. Epp & Bull, 2015; Kinkeldey et al., 2015; Kinkeldey, MacEachren, & Schiewe, 2014; MacEachren et al., 2005). These studies conclude that the presentation format could make a difference for user decision-making and understanding of the concept. Based on these studies, three promising formats presenting ME will be further explored: blur, colour value and error bar. Figure 4.1 presents these formats as well as a presentation format that omits ME.

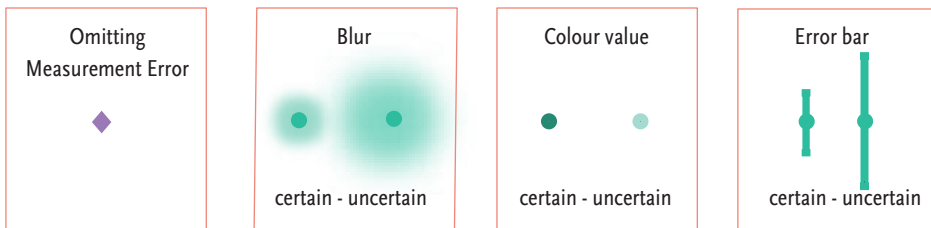


Figure 4.1 ME Presentation formats compared to the presentation format omitting ME

Blur can be defined as changes in the clarity or fuzziness of objects (Epp & Bull, 2015). The technique provides a general overview of uncertainty without quantifying exact values. It seems to be a promising and widely used tool for presenting error information because users intuitively associate blur with uncertainty (e.g. Johnson & Sanderson, 2003; MacEachren et al., 2012).

Colour value is a naturally orderable presentation format that can be defined according to changes from light to dark (Epp & Bull, 2015). Lighter values are associated with higher uncertainty, while darker values correspond to lower uncertainty (e.g. Kinkeldey et al., 2014; Leitner & Buttenfield, 2000). Colour value is used as a categorical presentation format containing a number of discrete value levels.

Error bars are additional graphic objects in the visualisation (Gershon, 1998). Because of the numerical and continuous representation, it is a suitable technique for presenting quantitative data (Brodie et al., 2012; Wainer, 1995). Several studies have however concluded that error bars dominate the certainty scores because the greatest visual emphasis is on the long bars, that present the most uncertainty (e.g. Sanyal, Zhang, Bhattacharya, Amburn, & Moorhead, 2009). In addition to the amount of uncertainty, the length of the bar is influenced by the type of confidence interval (e.g. 68%, 90% or 95%) that is represented. A sufficient level of statistical literacy is required to accurately interpret the length of the bar (Hullman, Rhodes, Rodriguez, & Shah, 2011; Zwick et al., 2014). Nevertheless, it is a commonly used technique for visualising ME within educational contexts (e.g. Phelps et al., 2010).

### 4.3 METHOD

A mixed-methods design was used to examine teachers' decisions and preferences regarding the various presentation formats (blur, colour value, error bar and omitting ME). Quantitative data were collected by means of a factorial survey. Qualitative data were collected by means of think-aloud protocols and focus groups to verify our findings, and to obtain a deeper analysis of the quantitative results.

#### *4.3.1 Design of the Visualisations*

Real student data from a standardised test were used to develop the test score reports. This test is used at 85% of Dutch primary schools and covers various domains of mathematics (e.g. counting and comparing numbers and addition and subtraction sums). The data are usually gathered every six months to monitor student performance and to develop a group action plan for the next six months.

For this test, ME is commonly determined by calculating the standard error of the ability estimate using the one-parameter logistic model of item response theory. To simulate a real decision-making process, this calculation was also used in the current study. This resulted in a 68% confidence interval consisting of one standard error above and one standard error below the ability estimate or score. Due to the use of actual data, the confidence interval for the higher score points

was smaller than for lower score points. However, since this occurred due to the use of actual data, it was not altered.

Because blur, colour value and error bar were considered promising presentation formats in the literature regarding the presentation of ME, we incorporated these formats into this study. This resulted into the comparison of two categorical (blur and colour value) and a numerical presentation format (error bar) with a presentation format omitting ME. Each of these presentation formats is associated with a certain amount of ME information. For example, the error bar is an exact and continuous presentation of the ME values, while blur and colour value provide only a global indication of the amount of ME. We investigated how these characteristics influence teachers' decisions and preferences.

In order to obtain a valid representation of the influence of the ME presentation formats, other essential ME characteristics that could influence teachers' decisions were investigated. Six educational measurement specialists were interviewed to indicate other essential ME characteristics that could influence teachers' decisions. Based on their input, two other characteristics were added: the position of the error in relation to the cut-off score (i.e. the cut-off score is outside, within or exactly in the middle of the confidence interval) and the size of the error (i.e. large or small). This resulted in four presentations x three positions x two sizes = 24 visualisations for each respondent (see Table 4.1).

#### 4.3.2 Respondents

Data on 487 pre-service and in-service teachers of Dutch primary education were collected, after contacting pre-service teachers from all 44 Dutch colleges as well as in-service teachers by email, Facebook and LinkedIn. From these 487 teachers, 150 did not complete the survey, which means that the responses of 337 teachers ( $N_{\text{male}} = 40$ ;  $N_{\text{female}} = 297$ ) were used for analysis. The male to female ratio is typical for the Dutch primary school teacher population ([onderwijsincijfers.nl](http://onderwijsincijfers.nl)).

The teaching experience of the teachers varied: 77.7% of them were pre-service teachers in the last year, 6.8% taught less than five years, 5.0% taught five to ten years, and 10.4% taught more than ten years. Furthermore, 82.2% of the teachers did not take a course on testing during their study, and 86.9% of them indicated that they had little or no statistical experience. Think-aloud protocols were conducted with a typical selection of 14 teachers, and eight focus groups were held ( $N = 35$ ) with an average of four teachers per focus group.

Table 4.1 Test score report visualisations

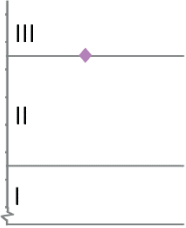
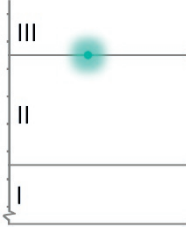
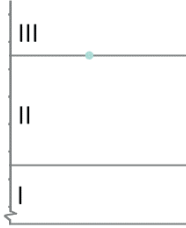

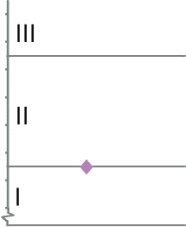


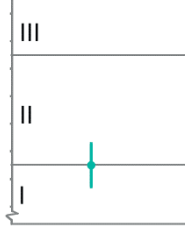
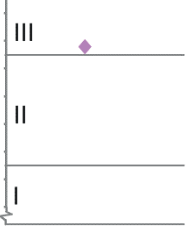
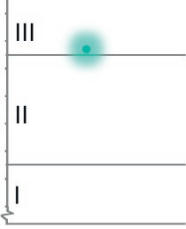
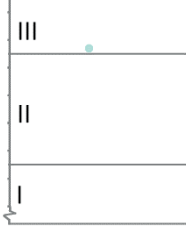
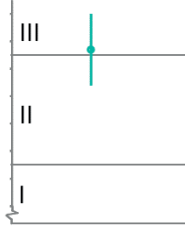


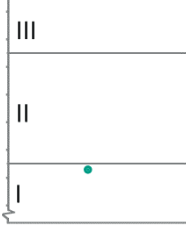
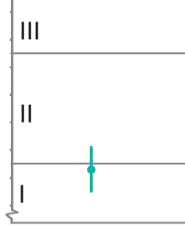
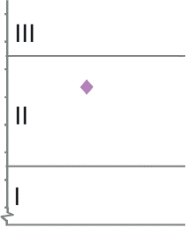
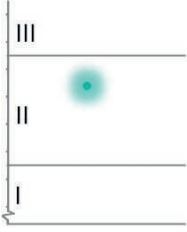
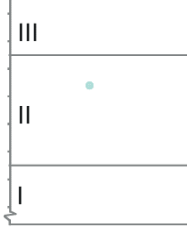
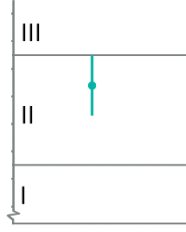
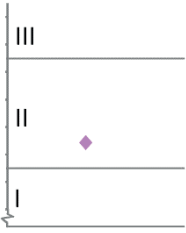
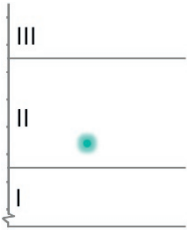
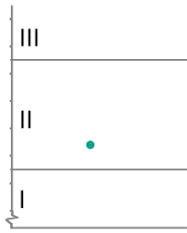
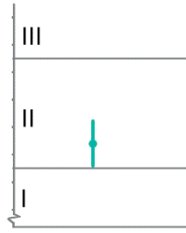
Position and size	Presentation format			
	Omitting ME	Blur	Colour value	Error bar
Exactly in the middle				
Large				
Small				
Within				
Large				
Small				

Table 4.1 (Continued)

Position and size	Presentation format			
	Omitting ME	Blur	Colour value	Error bar
Outside				
Large				
Small				

### 4.3.3 Instruments and Procedure

#### Survey

To study teachers’ decision-making processes, a factorial survey with true-to-life cases was developed (Auspurg & Hinz, 2014). In this survey, all the respondents were presented with all 24 visualisations (Table 4.1). We started with the six visualisations omitting ME, which is the usual format presented to Dutch teachers. As we started with these formats, the respondents’ answers were not influenced by the ME visualisations. Following this, the 18 visualisations containing blur, colour value and error bar were shown in random order as set by the online survey program. Teachers were shown a single visualisation on the screen, which showed a score report of one student. For every visualisation, the respondents were asked to judge a familiar type of educational action: the assignment of students to an instruction group for tailored instruction.



In the Netherlands, the assignment of students is often done by dividing them into three instruction groups: (I) an extended instruction group, (II) a basic instruction group and (III) a shortened instruction group. The 25% lowest scoring students are usually assigned to the extended instruction group for which teachers provide additional instruction using concrete learning materials. The 25% highest scoring students are commonly assigned to the shortened instruction group, in which they receive brief instruction and in-depth exercises. The remaining 50% of the students are assigned to the basic instruction group, in which teachers provide regular instruction.

In this study, respondents were asked to specify which instruction group (I, II, III) they would assign the student to or to indicate that they needed additional information about the student to make this decision. The need for additional information was defined as an indication for awareness of ME. It suggested the desire to gather multiple sources of information before making a decision, since the single test score contains some uncertainty.

Alongside the investigation of respondents' decisions regarding the 24 visualisations, the survey consisted of six items on the respondents' background and three questions about respondents' preferences (Appendix A). With regard to their background, the respondents were asked questions about their gender, their level of educational attainment, their years of experience teaching primary education, the name of the high school of teacher training, their courses on testing, and what they considered to be their own ability in statistics. With regard to the preference questions, respondents were asked to rank the four presentation formats from most preferred to least preferred. Furthermore, they were asked to indicate the extent to which the various presentation formats influenced their decisions as well as the extent to which the error presentations influenced their confidence in their decisions.

The survey was pretested, with 22 test experts completing the survey and indicating whether some questions were unclear. Subsequently, we pretested the survey with two teachers. Both pre-tests resulted in some minor adaptations, like changing the score point of no measurement error into a purple rhombus for a clear distinction with the colour value presentation. During the pre-test and data collection, the survey was completed online by the respondents.

*Think-aloud protocols and focus groups*

Think-aloud protocols were used to obtain insight into the cognitive processes underlying the teachers' decision-making processes (Bannert & Mengelkamp, 2008). The respondents were asked to verbalise their thoughts (i.e. think-aloud) while responding to items in the survey. The researcher was not allowed to request explanations because this could interfere with the respondents' cognitive processes.

After filling out the survey, focus groups were held to verify and clarify the findings of the survey. The respondents were asked to indicate their decision for a varying selection of four visualisations and to explain their choice. Furthermore, we investigated their interpretation of the score report and their comprehension of the ME concept, an explanation of their preferences and their perspective regarding the usefulness of visualising ME. The design of the focus group method included the characteristics of a group interview as well as a group discussion (Newby, 2010). The researcher fulfilled the role of moderator.

The think-aloud protocol and focus group were pretested with three teachers, resulting in some points of attention. The verbalisation of the think-aloud protocols and the focus group discussion were tape recorded.

*4.3.4 Data Analysis**Teachers' decisions*

To test whether the presentation formats resulted in a significantly greater need for additional information, the respondents' answers to the 24 score reports were recoded into dichotomous variables. Score 'o' indicated the assignment of a student to group I, II or III. Score 'i' indicated the need for additional information.

After performing frequency analyses, we conducted a Generalized Linear Mixed Model (GLMM) using the lme4 package for R (Bates et al., 2017). This model provides a method for analysing a dichotomous dependent variable in hierarchically structured data, which means a dependent variable containing precisely two distinct values and a dataset that is organised at more than one level. In this study, the teachers' decisions were defined as the dichotomous dependent variable, containing a score 'o' or 'i'. The data was hierarchically structured, given the 24 cases of data nested within each respondent. This data structure resulted in a random intercept for persons. The independent variables were the teachers'

background variables and the visualisation characteristics' position, size and format.

We started with a simple random intercept model containing a fixed intercept and a random intercept for persons. The independent variables were then added successively, and the fit of the new model was compared to the previous one. As the previous model was nested in the new one, a likelihood ratio (LR) test was used to test the improvement in goodness of fit. The resulting test statistic is chi-squared distributed, with the number of free parameters of the alternative model minus the number of free parameters in the null model as the degrees of freedom. Furthermore, the AIC, BIC and -LL indices were used, with lower values indicating a better fit.

In addition to the survey data, think-aloud transcriptions were divided into 24 units belonging to the 24 visualisations. For each unit, we identified factors that the teachers kept in mind during the decision-making and the categories of misconceptions emerging for certain presentation formats. The final coding scheme was used to double-code 10% of the transcriptions. An inter-rater reliability analysis was subsequently performed to determine the consistency between the two raters, which was found to be substantial (Cohen's  $\kappa = .738$ ). The transcriptions of the focus groups regarding the respondents' decisions and their interpretation of the score report and ME concept were classified and summarised.

### *Teachers' preferences*

To analyse the respondents' preferences regarding the presentation formats, the respondents' answers to the second part of the survey were analysed using frequency analysis. Furthermore, the discussion in the focus group around the preferences and usefulness of ME were summarised. In this article, the results are illustrated by examples translated from Dutch.

## 4.4 RESULTS

### *4.4.1 Teachers' Decisions*

Frequency analysis showed that the error bar format most often resulted in the need for additional information (see Table 4.2). The blur and colour value formats

both resulted in less need for additional information compared to the omitting ME format. According to the respondents' think-aloud protocols, additional information for all instruction groups comprised information about previous test scores, scores of peers, sub-scores of the corresponding test, working attitude, student age and student anxiety.

Table 4.2 Percentage of respondents (N = 337) needing additional information for each visualisation

Position and size	Presentation format			
	Omitting ME	Blur	Colour value	Error bar
Exactly in the middle				
Large	69.4	69.4	67.4	66.8
Small	63.8	67.1	60.5	71.5
Within				
Large				
Small	46.0	47.2	51.0	53.4
	51.6	43.9	32.3	54.6
Outside				
Large	13.1	16.3	21.4	18.7
Small	18.1	14.5	9.8	20.8
Total	43.7	43.1	40.4	47.6

Table 4.3 presents an overview of the comparisons between the estimated models of the GLMM analysis. Model 1a included all background characteristics. Only statistical knowledge had a significant effect ( $F(3, 8075) = 4.84, p = .002$ ) on the decisions. Respondents who assessed themselves as having a great deal of statistical experience requested additional information more often than respondents who assessed themselves as having no ( $B = -1.53, SE = .58, p = .009$ ), little ( $B = -1.64, SE = .57, p = .004$ ) or quite a lot of ( $B = -2.17, SE = .59, p < .001$ ) statistical experience. No additional statistical differences were found between respondents with no, little or quite a lot statistical experience. Because of the significant improvement of Model 0 (see Model 1b), we decided to retain this background variable in subsequent analyses.

In Model 2, the role of the presentation format, position and size on the respondents' decisions was examined by adding these as fixed effects. The model improved significantly as all effects were reported as significant at  $p \leq .001$ .

Table 4.3 Overview of the estimated models

Model	Nested model	Effects	Random over persons					LR test	
			Fixed	AIC	BIC	-LL	#p	df	$\chi^2$
0		Intercept	Intercept	9543.1	9557.1	-4769.6	2		
1a	0	+ sex, level of education, teacher experience, education in tests, and statistical knowledge	"	9547.8	9645.7	-4759.9	14	12	19.37
1b	0	+ statistical knowledge	"	9533.8	9568.8	-4761.9	5	3	15.32*
2	1b	+ format, position, size	"	7331.8	7408.8	-3654.9	11	6	2214*
3a	2	+size x format, size x position, position x format	"	7287.0	7441.0	-3621.5	22	11	66.76*
3b	2	+ size x format	"	7281.2	7379.2	-3626.6	14	3	56.55*

\* p &lt; .01

Based on the results of Model 2 and the visualisation of the presentation formats, we investigated the interaction effects between the presentation format, position and size in Model 3a. We hypothesised at least an interaction between presentation format and size because the colour value, blur and error bar formats differ in size from each other. The model improved significantly as the interaction between presentation format and size was significant at  $p < .001$ . Based on this result, we removed the other interactions and maintained only the interaction between format and size in Model 3b. Model 3b resulted in the best fitting model, showing significant fixed effects for statistical experience, presentation format, position, size and size\**presentation format* interaction. We found random intercepts for persons ( $Variance = 4.30$ ;  $SD = 2.07$ ;  $p < .001$ ). The results of Model 3b are presented in Table 4.4 and discussed in detail below.

Table 4.4 Estimates of unstandardised (B) and standardised ( $\beta$ ) effects on teachers' decisions in model 3b

	B	SE	$\beta$	p
Intercept	0.26	.75	0.00	.725
Background characteristics				
Statistical experience <sup>a</sup> (reference: A great deal (work activities))				
No	-2.70	.77	-2.68	<.001
Little (one course)	-2.81	.77	-2.82	<.001
Quite a lot (more courses)	-3.32	.83	-2.04	<.001
Visualisation characteristics				
Format <sup>b</sup> (reference: Omitting ME)				
Blur	-0.21	.12	-0.18	.091
Colour value	-0.82	.13	-0.72	<.001
Error bar	0.35	.12	0.30	.005
Position <sup>c</sup> (reference: Outside)				
Exactly in the middle	3.70	.10	3.51	<.001
Within	2.42	.09	2.30	<.001
Size <sup>d</sup> (reference: Small)				
Large	-0.13	.12	-0.13	.287
Interaction				
Size x Format <sup>e</sup>				
Large x Blur	0.33	.17	0.22	.063
Large x Colour value	1.11	.18	0.74	<.001
Large x Error bar	-0.08	.17	-0.05	.661

<sup>a</sup>F(3, 8075) = 2.09,  $p = .098$ ; <sup>b</sup>F(3, 8075) = 13.628,  $p < .001$ ; <sup>c</sup>F(2, 8075) = 707.304,  $p < .001$ ; <sup>d</sup>F(1, 8075) = 10.134,  $p = .001$ ; <sup>e</sup>F(3, 8075) = 19.023,  $p < .001$

### *Presentation format*

Model 3b showed a significant main effect for presentation format ( $F(3, 8075) = 13.628, p < .001$ ). The error bar presentation format resulted in the most need for additional information. In order to yield interpretable odds ratios, the fixed effect must first be exponentiated. Thus, the estimated odds that additional information will be chosen for the error bar format above the omitting ME format is  $\exp(0.35) = 1.42$  times.

No significant difference was found between the blur and omitting ME formats ( $B = -.21, SE = .12, p = .091$ ). According to the think-aloud protocols, it seems that respondents' interpretation of blur was too literal, as opposed to the actual meaning, as blur reflected a categorical value. For example, respondent 3 interpreted the outline of blur as a 68% confidence interval: *"Okay, I see a dot and a blur around it, indicating that this student scored around 37. However, (...) it could also be a score of 35 or even 40."* As a small blur presentation was smaller than the real 68% confidence interval, the blur presentation suggested a smaller ME size than it actually was.

Colour value resulted in significantly less need for additional information ( $B = -.82, SE = .13, p < .001$ ). The estimated odds that additional information would be chosen for colour value above omitting ME were  $\exp(-0.82) = 0.44$  times. The think-aloud protocols illustrated that the lighter and darker values were not as associative as they should have been. Respondent 11 thought: *"This is certain, right? No, this is uncertain?"* Respondent 13 illustrated that even a change in the meaning, such as a light colour value format, was associated with certainty: *"This student has a certain score at the border of group III. Therefore, I would assign her to group III purely because it is a certain score."*

To sum up, different ME presentation formats did not always result in a need for additional information. The respondents chose to seek additional information significantly more often only for the error bar. The blur format did not change the decisions, probably because of its literal interpretation. Colour value resulted in significantly less need for additional information, probably due to the confusing association of the values.

### *Position*

Model 3b yielded a significant main effect for position,  $F(2, 8075) = 707.304, p < .001$ . Given the standardised effects in Table 4.4, the decisions were most impacted by the position of a score in relation to a cut-off score. The estimated

odds that additional information would be chosen for a cut-off score exactly in the middle of the error above a cut-off score outside the error was  $\exp(3.70) = 40.45$  times. The odds for a cut-off score within the error above a cut-off score outside the error was  $\exp(2.42) = 11.25$  times. This meant that the respondents would more often request additional information in the event of a cut-off score exactly in the middle of the error, followed by a cut-off score within and outside the error. This corresponds to the idea that it would be more difficult to assign a student to a group when the test score approaches the cut-off score.

The think-aloud protocols corroborated this result. For example, respondent 7 argued about a cut-off score exactly in the middle of the error: *“This student scored exactly between groups I and II, so I would like to have additional information about the extent to which she is able to perform in group II”*. This was in contrast with a cut-off position outside the error: *“This student is in the second group, so I would assign her to the second group.”*

### Size

Model 3b showed a significant main effect for size on decision ( $F(1, 8075) = 10.134, p = .001$ ). However, the size estimates were no longer significant in Model 3b ( $B = -0.13, SE = .12, p = .287$ ) as a result of the significant interaction effect between size and presentation format ( $F(3, 8075) = 19.023, p < .001$ ). This indicated that size had different effects on respondents' decisions, depending on which presentation format was shown. Decomposition of the interaction revealed that a dark colour value format – which is a small error size – resulted in more assignments of students to groups ( $B = 1.11, SE = .18, p < .001$ ).

The think-aloud protocols showed that respondents considered a dark colour value as a very certain format. Respondent 14, for example, argued: *“This is a certain score in group I. Therefore, I would assign this student to group I”*. By contrast, the other presentation formats were seen as less certain because the small error size was even larger than the colour value point: *“This [error bar] is a little uncertain; group II, I, III? Let me think. Perhaps additional information because it is a little vague”* (respondent 12). Furthermore, some respondents confused the colour value format by regarding it as the smallest error size in terms of blur: *“I would assign him to the second group since there is no variation around the score”* (respondent 13).

To sum up, there was an interaction effect between size and presentation format. Colour value resulted in significantly less need for additional information



because a dark colour value was interpreted as very certain. As a result, the respondents tended to assign many students when a dark colour value was presented; however, this format also included a small error size.

### *Conceptions and misconceptions of ME.*

Although the error bar resulted in an increased need for additional information, the respondents varied in their understanding of ME. Several reasons explaining the cause of ME were given. All focus groups indicated the cause of ME as a variation in human performance and environment, such as influences relating to the well-being of the student or the location. In three out of eight focus groups, a respondent indicated the cause of ME as a variation caused by the selection of test items: *“I think it is about the selection of items. The more items you select, the more precise the results are – like the more research you do, the more solid your research is. I think that’s what it means.”*

However, there were also several misconceptions around the ME concept (see Table 4.5). Respondents in three focus groups attributed the cause of ME, for example, to the difference between the test score and the perspective of the teacher. During the think-aloud, respondent 6 thought that the error indicated the uncertainty of the students themselves, like their test anxiety. Furthermore, respondents from three focus groups had no idea about the causes of ME.

Table 4.5 Misconceptions about (the cause of) ME among focus groups (N=8)

Misconception	Number of focus groups	Illustrative example
Difference between test score and other test scores	5	‘Perhaps contradicting scores on tests from teaching methods compared to standardised tests – so the standardised test has a very high score, and the score on the teaching method test is very low’.
Difference between test score and teacher’s perspective	3	‘Maybe a teacher adds the uncertainty. He or she is not sure about the test score, or the score is not in line with his or her perspective on the student’.
Difference between process and outcome	2	‘The student performed the calculations well but gave the wrong answer. So you do not know what exactly went wrong’.
Response pattern of the student	1	‘Is it about duration: how much time does a child need to answer the question? (...) Or is it about changing the filled-in response often?’
An invalid item	1	‘You have a contextual item in a math test, but the student cannot read, which resulted in a wrong answer, even though he is a good math student. (...) Then the question is not certain’.

#### 4.4.2 Teachers' Preferences

Frequency analyses showed that the error bar was the most preferred presentation format for ME ( $M = 3.06$ ,  $SD = 1.09$ ), followed by blur ( $M = 2.63$ ,  $SD = 0.93$ ), colour value ( $M = 2.58$ ,  $SD = 0.97$ ) and omitting ME ( $M = 1.72$ ,  $SD = 1.05$ ). The respondents in the focus groups preferred the error bar due to the exact presentation of the numerical ME values, while the blur and colour values were categorical presentations. The clear borders of the error bar were most highly appreciated, although the vague borders of the blur format led the respondents to consider ME more often. Furthermore, both the higher values of the error bar and blur were seen as associated with greater uncertainty, while the association of a lighter colour value with greater uncertainty was lacking. One disadvantage of the error bar according to the respondents was the extensive length of some bars, resulting in less confidence in the decision. Blur was less preferred because the width of the blur format had no meaning. A disadvantage of colour value was the limited possibility to convert the presented coloured format in a black and white score report when printed at home or at school, as it will become poorly readable. The reasons for omitting ME were the prevention of confusion and the availability of sufficient test information, including other test scores. A disadvantage of omitting ME was the lack of insight into the reliability of a test.

The results of the survey showed that the respondents believed that the different ME presentations affected their educational decisions. On a 5-point Likert scale ranging from never to always, error bars influenced sometimes ( $M = 3.11$ ), followed by blur ( $M = 3.02$ ) and colour value ( $M = 2.72$ ). During a think-aloud protocol, a respondent said: *"I notice that I've often indicated the need for additional information. This is quite logical for me because I now realise that the test scores are not as exact as they seem, and you still want to make good decisions for your students"* (Respondent 3).

The presentation of ME did not affect confidence in educational decisions for 51.4% of the respondents, and 30.8% of them indicated that the presentation of ME had a positive impact on their confidence. For example: *"Anyhow, a lot more confidence: because now you do justice to the students. If you did not use this information, if you did not have the representation of uncertainty, you would make your decision based on an exact presented score. And now, you have included the influencing factors on the test score"* (Respondent 3).

The remaining 17.8% indicated less confidence in their decisions, as indicated by respondent 10: *“Yes, you will still doubt because you see that there may be uncertainties. I think this gives less confidence. I would like to indicate the need for additional information more often”*.

## 4.5 CONCLUSION AND DISCUSSION

This study set out to determine teachers’ decisions and preferences regarding various ME presentations. Quantitative and qualitative data were collected by means of a factorial survey, think-aloud protocols and focus groups.

The results showed that ME presentations influence teachers’ educational decisions compared to presentations that omit ME. The error bar format resulted in significantly greater need for additional information about a student. The colour value format resulted in significantly less need for additional information, while the blur format did not differ significantly from the omitting ME format. Furthermore, the results showed that the position of a score in relation to a cut-off score had the most impact on the decisions. The size was influenced by the format and had no independent effect in this study.

Moreover, the error bar was found to be the most preferred format because of its exact presentation of the numerical ME values. The desirability of this advantage can be questioned because ME is not exact. It is an estimated value and can be visualised by a 68% confidence interval as well as by 95% and 99% confidence intervals. By contrast, the vague borders of blur ensured that there would be no exact interpretation, resulting in further thoughts about the ME concept. However, the respondents interpreted blur as a numerical variable, while this study presented blur as a categorical variable. Colour value and omitting ME were the least preferred presentation formats.

As every study is accompanied by some measurement error, we should draw the conclusions of this study carefully. The first limitation is that 150 respondents did not complete the survey. Since we do not know whether they differed (e.g. regarding their statistical experience) from those who completed the experiment, we urge caution in the interpretation of the results of the study. Secondly, the frequency of the decision regarding the need for additional information can be underestimated as the think-aloud protocols showed that the respondents assigned students to one of the groups but in fact wanted to gather additional

information. For example, respondent 9 said: *“I would assign this student to group III and observe the progress. It is difficult to know that with one test score”*. Thus, the assignment of students to an instruction group was less a matter-of-fact decision for teachers than we assumed. In addition, the think-aloud protocols provided insight into the teachers’ cognitive processes; however, it seems that the teachers did not repeat their reasoning for each visualisation. Although the investigator encouraged the respondents to continue thinking aloud, and visualisations were shown in random order, the results might be an underestimation of the number of times the teachers really looked at the format, position and size when taking a decision. Finally, the context of the current study may have influenced the results obtained. Although we chose a common type of educational decision and a commonly used test, other types of decisions and tests may result in teachers wanting more or less additional information. For example, the presentation of real test results resulted into visualisations in which a small error size is always accompanied with lower scores compared to a large error size with higher scores. The results of the focus groups and think aloud protocol, however, did not give reason to think that the height of student’s score is confounded with the error size.

The results and limitations point to some suggestions for future research. First, this study indicates the fruitfulness and necessity of evaluating score reports with the intended audience so that they can be interpreted and used in a valid way. Therefore, based on this study’s results, we suggest an investigation into whether a combination of blur and error bar functions can be a suitable presentation of ME in test score reports. Advantages relating to the error bar include the numerical presentation, the positive influence on decisions and teachers’ preferences. Those relating to blur are the natural association with uncertainty and the avoidance of exact interpretation. A combination of blur and error bar is known as a gradient plot (see Correll & Gleicher, 2014) and consists of an error bar with blurred ends. As Correll and Gleicher recommend this gradient plot for indicating uncertainty among general audiences, it is interesting to examine the extent to which this presentation is deemed suitable for presenting ME in test score reports. Second, it seems worthwhile to examine the influence of other design factors on teachers’ understanding and use, which were less relevant for the currently used context of test scores. For example, the width of the Y-axis, the number of cut-off scores and the visualisation of previous scores can change the way in which teachers make their decisions. Moreover, it would be interesting to study the influence

of ME on other kinds of educational decisions such as planning regarding the next steps in instruction. Third, despite the potential impact of ME presentation, the teachers demonstrated several new misconceptions about the concept itself. Future research is needed into teachers' understanding and misconceptions of ME and effective ways to reduce misconceptions. The study of Zapata-Rivera, Zwick and Vezzu (2016) is a useful contribution to this area. It developed an ME tutorial to help teachers understand score report results. Future research should investigate the long-term effects of such tutorials on teachers' interpretation and use of test scores.

The findings of this study enhance our understanding of the usefulness of displaying ME. The results can be used in the design of new test score reports. Practical implications would include the use of ME in order to make teachers more aware of the imprecision around scores as well as fostering the use of multiple sources for taking educational decisions, such as other test scores, observations and students' work. In deciding on the use of alternative sources, it is important to consider the psychometric characteristics that are inherent to specific data sources. The findings also imply the need for a clear explanation of the ME-concept as several misconceptions of teachers were identified. This way, carefully designed test score reports could lead to a better understanding by teachers, thereby improving the quality of their educational decisions.

## 4.6 REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Auspurg, K., & Hinz, T. (2014). *Factorial survey experiments. Applications for the Social Sciences* (Vol. 175). Thousand Oaks, CA: Sage Publications.
- Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning*, 3, 39–58. doi:10.1007/s11409-007-9009-6
- Bates, D., Maechler, M., Bolkers, B., Walker, S., Christensen, R. H. B., Singman, H., ... Green, P. (2017). *The lme4 package*. Retrieved from <http://r-forge.r-project.org/projects/lme4/>

- Bradshaw, J., & Wheater, R. (2009). *National foundation for educational research: International survey of results reporting* (OFQUAL 10/4705). London: Office of Qualifications and Examinations.
- Brodie, K. W., Osoria, R. A., & Lopes, A. (2012). A review of uncertainty in data visualization. In J. Dill, R. Earnshaw, D. Kasik, J. Vince, & P. C. Wong (Eds.), *Expanding the frontiers of visual analytics and visualization* (pp. 81–110). London: Springer.
- Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, NJ: Pearson Education.
- Centraal Bureau voor de Statistiek (CBS), Dienst Uitvoering Onderwijs (DUO) en het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) (n.d.). *Onderwijs in cijfers* (Education in numbers). Retrieved from <http://www.onderwijsincijfers.nl>
- Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20, 2142–2151. doi:10.1109/TVCG.2014.2346298
- Epp, C. D., & Bull, S. (2015). Uncertainty representation in visualizations of learning analytics for learners: Current approaches and opportunities. *IEEE Transactions on Learning Technologies*, 8, 242–260.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.
- Gardner, J. (2013). The public understanding of error in educational assessment. *Oxford Review of Education*, 39, 72–92. doi:10.1080/03054985.2012.760290
- Gershon, N. (1998). Visualization of an imperfect world. *IEEE Computer Graphics and Applications*, 18, 43–45. doi:10.1109/38.689662
- 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*, 17, 145–220. doi:10.1207/s15324818ame1702
- Hullman, J., Rhodes, R., Rodriguez, F., & Shah, P. (2011). *Research on graph comprehension and data interpretation: Implications for score reporting* (ETS RR-11-45). Paper presented at the ETS Score Reporting conference, Princeton, NJ.
- Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice*, 10(4), 16–18.
- Johnson, S., & Sanderson, A. R. (2003). A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5), 6–10.

- Kinkeldey, C., MacEachren, A. M., Riveiro, M., & Schiewe, J. (2015). Evaluating the effect of visually represented geodata uncertainty on decision-making: Systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science*, 44, 1–21. doi:10.1080/15230406.2015.1089792
- Kinkeldey, C., MacEachren, A. M., & Schiewe, J. (2014). How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, 51, 372–386. doi:10.1179/1743277414Y.0000000099
- Leitner, M., & Battenfield, B. P. (2000). Guidelines for the display of attribute certainty guidelines for the display of attribute certainty. *Cartography and Geographic Information Science*, 27, 3–14. doi:10.1559/152304000783548037
- MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32, 139–160. doi:10.1559/1523040054738936
- MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization & Computer Graphics*, 18, 2496–2505.
- Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, 47(2), 71–85. doi:10.1080/00461520.2012.667064
- Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Education Research Journal*, 31, 419–442. doi:10.1080/01411920500148648
- Pang, A. T., Wittenbrink, C. M., & Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer*, 13, 370–390. doi:10.1007/s003710050111
- Phelps, R. P., Zenisky, A., Hambleton, R. K., & Sireci, S. G. (2010). *On the reporting of measurement uncertainty and reliability for U.S. educational and licensure tests* (OFQUAL 10/4759). London: Office of Qualifications and Examinations.
- Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., & Moorhead, R. J. (2009). A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15, 1209–1218. doi:10.1109/TVCG.2009.114
- Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623-646). Westport: American Council on Education and Praeger Publishers.
- Wainer, H. (1995). *Depicting error* (Technical Report No. 95-2). Princeton, NJ: Educational Testing Service.

- Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36, 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x
- 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, 215–229. doi:10.1080/10627197.2016.1202110
- Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, 19, 116–138. doi:10.1080/10627197.2014.903653



## APPENDIX A: SURVEY

Table A.1 Survey

---

### Respondents background

1. What is your sex?

a) Male; b) Female

2. What is your highest level of educational attainment?

a) Higher general secondary education; b) Pre-university education; c) Vocational education; d) Higher education; e) University

3. At which high school do you attend the teacher training? (Only for final-year pre-service teachers)

4. Did you take a course on testing during the teacher training?

a) No; b) Yes, namely....

5. How much experience do you have with statistics? Chose the most appropriate answer.

a) I have no experience with statistics; b) I have little experience with statistics (e.g. one course during secondary education); c) I have quite a lot of experience with statistics (e.g. more courses); d) I have a great deal of experience with statistics (e.g. more courses and own work activities).

6. How many years' experience do you have in primary education? (only for in-service teachers)

a) Less than 5 years; b) 5 to 10 years; c) More than 10 years

---

### Score reports omitting ME

On the next page, you will see the test score of a group of students on the national mathematics test. The score reports will be used to create a group action plan for the next semester. The group action plan will consist of three groups:

Group I: extended instruction [consisting of students who get additional instruction]

Group II: basic instruction [consisting of students who require the regular amount of instruction]

Group III: shortened instruction [consisting of students who only need brief instruction]

Please assign each student to a group (I, II or III) or indicate that additional information (from other tests, method assignments, etc.) would be needed to make this decision.

NB. The number of students per group may not be the same. For example, you may also assign all students to Group III. Base your choice on the corresponding score report, and do not look at the reality of the group action plan.




---

Choose: a) group III; b) group II; c) group I; d) I need additional information about the student to make this decision.

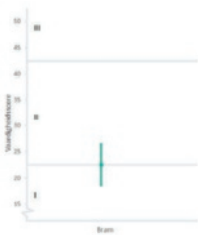
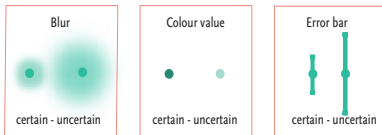
---

---

### Score reports with ME

On the next page, you will see the score reports of the other students for the national mathematics test. For each score report, we added information about the certainty regarding the test score as a good estimation of the students' mathematics skills, other factors remaining equal. The figure on this page shows three examples of test score reports.

Indicate the group (I, II or III) to which you would assign each student, or indicate that additional information would be needed (from other tests, method assignments, et cetera) in order to make this decision.



Choose: a) group III; b) group II; c) group I; d) I need additional information about the student to make this decision.

---

### Preference

Finally, we look forward to your experience and preferences for these presentations.

1. Which of the presentations do you prefer? Order them according to: 1 = most preferred; 4 = least preferred.

a) Presentation A. Error bar; b) Presentation B. Blur; c) Presentation C. Colour value; d) Presentation D. Omitting ME

Omitting ME

2. To what extent did the presentation of uncertainty affect your decision compared to the presentation omitting uncertainty?

a) Error bar: never-rarely-sometimes-very often- always; b) Blur: never-rarely-sometimes-very often- always; c) Colour value: never-rarely-sometimes-very often- always

3. To what extent did the presentation of uncertainty affect your confidence regarding your decision compared to the presentation omitting uncertainty?

a) Less confidence 1-2-3-4-5 More confidence

4. Comments section

---



# CHAPTER 5



# THE USABILITY OF AN EMBEDDED FORMATIVE ASSESSMENT SYSTEM

Formative assessment has been considered a promising way to support student learning. However, there is a need for validity evidence that supports the underlying assumptions of formative assessment. In this study, validity evidence was collected to support assumptions regarding the intended use of formative assessment. The evidence focused on the question of whether assessment results are usable for teachers' formative assessment practices. A prototype of an embedded formative assessment for math in primary education was used to collect evidence. The prototype was used by 29 teachers in a natural classroom setting for three months, during which time, data were collected from log files, questionnaires, and interviews. The results show that the prototype was largely usable in terms of establishing where the students were in their learning and where they needed to go and that it was somewhat usable in relation to how best to get there. Moreover, some improvements were needed in order to use the assessment results in the intended way. These improvements were described as design principles regarding the development of formative assessment instruments.

Keywords: formative assessment; usability; validity evidence

This chapter has been submitted as:

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). Collecting validity evidence about the usability of an embedded formative assessment system.

## 5.1 INTRODUCTION

Teachers continuously make decisions about the instructional process. For example, they decide on learning objectives for students, analyze what students have learned, and decide how to guide them toward planned learning objectives. This guidance may involve giving feedback or providing additional instruction.

The implementation of formative assessment is one way in which to support teachers' decision-making (Black & Wiliam, 2009; Brookhart, 2007). It can be conceptualized as a thoughtful integration of both an instrument and a process (Bennett, 2011). A well-designed instrument provides data about a student's learning. During the process, these data are judged and used for instructional actions that support that learning.

Since poor formative assessment can result in less effective teaching and learning, good quality in formative assessment is paramount (Brookhart & Nitko, 2008). Validity is one of the most important criteria for the quality of assessment (AERA, APA, & NCME, 2014) and is often defined as the extent to which an assessment result is appropriate for its intended interpretation and use (Kane, 2013; Sireci, 2013). The intended interpretation is about the meaning of the assessment result. For example, it is assumed that the assessment tasks reflect the construct of interest. The intended use involves the decisions, instructional actions, and possible consequences. For example, it is assumed that assessment results enable teachers to select appropriate instructional actions.

The validation of formative assessment requires a critical evaluation of the intended interpretation and use. For this critical evaluation, evidence that supports the assumptions and ensures that no alternative explanations are found must be collected (Kane, 2006). This evidence can be collected during and after development. Based on the evidence, a conclusion can be drawn about the extent to which the assessment results can be interpreted and used for the purpose of formative assessment.

In order to collect evidence for validation, there is a need to provide examples of formative assessment (Bennett, 2011) so as to demonstrate what formative assessment, from a theoretical perspective, might look like and how it might work in a natural setting. An example has been provided by Brown, O'Leary, and Hattie (2019), who describe the development and validation of an online teaching and learning system—the Assessment Tools for Teaching and Learning system (asTTle)—which has been deployed in New Zealand's schools to support effective

formative assessment practices by teachers. Since few assessment systems have been thoroughly evaluated and documented in the same manner as the asTTle system, validity evidence for formative assessment remains limited. As such, more examples are needed to support its underlying assumptions and to advance this promising concept (Bennett, 2011; Brown et al., 2019) in the assessment literature.

One of the assumptions underlying formative assessment is that the corresponding results are usable for teachers. This means that teachers are able to use the results for establishing where the students need to be, where they are, and how to close the gap (e.g., Furtak, 2006). Research shows that assessment instruments can provide a significant amount of data without paying attention to the types of decisions that intended users want to perform (Hattie, 2005; Wiliam, 2011). This has resulted, for example, in data providing a global reflection rather than fine-grained information about a student's capability (Goertz, Olah, & Riggan, 2009; Timperley, 2009).

The current study aims to collect validity evidence about the usability of formative assessment instruments by taking an example from the Dutch context. The main question of the study is as follows: *To what extent are assessment results usable for teachers' formative assessment practices?* An embedded formative assessment system prototype, called Groeimeter (GM), was used to answer this question. GM aims to support primary school teachers and students in their efforts to teach and learn math. The prototype is being developed through an educational design research approach (McKenney & Reeves, 2012; Plomp, 2013), which means that design principles derived from a conceptual framework underlie the development of GM and serve as criteria for evaluation. In the next section, we will present the conceptualization of formative assessment and the accompanying design principles.

## 5.2 CONCEPTUAL FRAMEWORK

Formative assessment comprises three major questions (Broadfoot et al., 2002; Furtak, 2006; NRC, 2001; Wiliam & Thompson, 2007): (1) Where do students need to go? (2) Where are they now? and (3) How best to get there? The first question is about establishing the desired state of learning, while the second relates to identifying students' current state of learning. The third question uses

this information to advance students' learning from the current to the desired state. We will elaborate on these questions in the remainder of this section. The accompanying design principles are presented in Table 5.1.

### *5.2.1 Establishing Where Students Need to Go*

In order to establish the desired state of learning, teachers need to determine what students should know and be able to do at the end of some period of instruction (Alonzo, 2011). Stiggins (2001) wrote:

The quality of any assessment depends first and foremost on the clarity and appropriateness of our definitions of the achievement target to be assessed...We cannot assess academic achievement effectively if we do not know and understand what that valued target is. (p. 19)

As Stiggins stated, it is important that teachers have clear learning objectives before they assess students' work and responses. These learning objectives articulate which aspects of students' knowledge and skills might be particularly salient. To determine the specific learning objective that students need to undertake, learning trajectories might help teachers understand how knowledge and skills develop within a domain as well as the connections between learning objectives (Alonzo, 2011; Furtak & Heredia, 2014; Heritage, 2008).

### *5.2.2 Establishing Where Students Are*

Once the learning objective has been determined, teachers have to ascertain where students currently stand in relation to that learning objective. They need to elicit evidence about students' performance, using tasks that reflect the learning objective. To illustrate, some learning objectives demand that students have understanding of some knowledge, while others require that students use some knowledge to reason and solve problems (NRC, 2001). The evidence is examined from the perspective of what it shows about students' conceptions, misconceptions, knowledge, and skills (Heritage, Kim, Vendlinski, & Herman, 2009). As there are always factors that can cause misrepresentation of students' current performance, teachers have linked evidence to other information about the student, such as the progress over time, the amount of effort invested, and the particular context (Bennett, 2011; Harlen & James, 1997). The combination of



different sources of information results in a diagnosis, which serves as a basis for instructional decisions (Mandinach, 2012).

### 5.2.3 *Establishing How Best to Get There*

In order to advance students' learning from a current to a desired state, teachers make decisions about instructional actions that they think can support students' learning. If a teacher diagnoses a gap between a current and desired state, then the teacher could provide feedback, reteach the learning objective, or seek to eliminate a misconception. If the gap is bridged, then the teacher could plan a new learning objective on the way to a final destination. To select appropriate instructional actions, the assessment information has to be tied to the curriculum and fit the teacher's knowledge base, including knowledge about instructional strategies that support students' progress (Falk, 2012; Forbes, Sabel, & Biggers, 2015; Goertz et al., 2009; Heritage et al., 2009; Herman, Osmundson, Ayala, Schneider, & Timms, 2006).

Table 5.1 Design principles for the development of a formative assessment system

If the formative assessment system is designed for the purpose of formative assessment in primary education, then the system...
Establishing where students need to go
...provides clear descriptions of the learning objectives.
...shows a clear visualization of the learning trajectory.
Establishing where students are in their learning
... consists of assessment tasks that reflect the learning objective.
... visualizes the assessment results and students' progress.
... provides the possibility to overrule the assigned status.
Establishing how best to get there
...is tied to the curriculum.
...provides the possibility to plan learning objectives.

### 5.3 RESEARCH QUESTIONS

Given the conceptualization of formative assessment, the main research question (To what extent are the assessment results usable for teachers' formative assessment practices) can be subdivided into several sub-questions:

1. To what extent are the assessment result usable for teachers in establishing where students need to go?
2. To what extent are the assessment results usable for teachers in establishing where students are?
3. To what extent are the assessment results usable for teachers in establishing how best to get there?

GM is used as an operational example to answer these sub-questions. The next section will present a description of GM.

### 5.4 ABOUT GM

GM is a formative assessment system, which is being developed through an educational design research approach (McKenney & Reeves, 2012; Plomp, 2013). This approach has the dual aim of generating research-based solutions for complex problems in educational practice and of advancing our theoretical understanding about the characteristics of these interventions. Three main phases can be distinguished: (1) a needs and context analysis; (2) the design and formative evaluation of the prototype tools; and (3) a semi-summative evaluation of the final product. The development of GM is in the second phase, which features the design, development, and formative evaluation of several prototypes (Nieveen & Folmer, 2013).

GM has been designed according to the design principles presented in Table 5.1 (see also the Dutch description: <https://www.cito.nl/kennis-en-innovatie/citolab/citolab-projecten-po/citolab-po-groeimeter>) and is being developed for use in the Dutch context, which is characterized by local curricular decision-making at the school and class levels (Van Zanten & Van den Heuvel-Panhuizen, 2018). This means that individual schools may fill in specific details regarding the learning content within the legal framework of the core objectives and reference framework. There is much room for interpretation regarding what mathematics

students should learn in primary school. The GM system is aimed at supporting teachers in these curricular decisions.

GM provides teachers with a dashboard (Figure 5.1), which shows all the learning objectives for grades 2, 3, and 4 (7–9 year olds) in the arithmetic curriculum, as developed by the Netherlands Institute for Curriculum Development (Noteboom, Aarsten, & Lit, 2017). The meaning of each learning objective is clarified on a separate screen on the dashboard (Figure 5.2), which shows an explanation, an example of a test item, and sometimes an instructional video about the learning objective. Teachers can use filter options (Figure 5.1) showing only learning objectives for a certain grade or for a certain learning domain. They can also use filter options for related learning objectives in the learning trajectory or search for a certain concept in the learning objective description. This way, they can determine students' learning objective.

To measure students' current performance, the learning objectives are operationalized in pre-defined formative assessments. There are two types of assessments, which depend on what best fits the learning objective to be measured. The first type is a digital test, in which students answer seven items online. This type is used for learning objectives that can be operationalized into automatically scored items (e.g.: "The student is able to calculate additions and subtractions up to 20"). The items could be short-answer, multiple choice, multiple response, hotspot, or matching items (e.g.: students fill in the right answer to the short-answer item: "How many balls do John and Mike have together?" or they need to select coins that, when summed, amount to 15). Mastery of the learning objective is automatically assigned when six items are answered correctly (Béguin & Straat, 2019). This cut-off score minimizes the chance of wrongly assigning mastery status to a student who does not master the learning objective. The second type is an assignment, for instance, having a group discussion or designing a drawing. It is used when the learning objective is not suitable for automatic scoring because it requires more cognitively complex thinking (e.g.: "The student can think and reason critically about length and perimeter in meaningful problem situations"). In the assignment, the students were asked to come up with three different rectangles with a 16-meter perimeter and to explain their answer. In another assignment, they had to calculate the perimeter of a new fence for the parcels of land belonging to the farmer, James. Mastery of the assignment was manually assigned by the teacher.

In the dashboard, teachers can review students' performance on the learning objectives. The mastery or non-mastery of the learning objective is represented as green or orange, respectively (Figure 5.1). Teachers can view students' individual item responses on the digital test and compare them with the correct answers (Figure 5.3). They can try to explain the responses by linking them to the students' individual circumstances. When teachers determine that the automatically assigned status (mastery/non-mastery) does not reflect reality, they can overrule the status. In addition, they can view students' progress within a learning domain through a green bar that expands as additional learning objectives are mastered. This way, teachers can establish where students currently are (Figure 5.4).

The assessment results are supposed to be used for follow-up actions. For example, teachers are considered to provide additional instructions if they conclude that a learning objective was not mastered due to a certain misconception. It is also possible to plan new learning objectives for an individual student as well as for the whole grade of students (Figure 5.1). When a learning objective is assigned to a student, the student can start the accompanying assessment if he or she feels capable.



Figure 5.1 Dashboard teacher. Left: learning objectives; top: filter options; right: for each student, it is shown whether a learning goal has been mastered (green), non-mastered (orange), planned (white), or not planned (gray).

Figure 5.2 Learning objective screen. Left: the learning objective and the explanation; right: the test item example and the instructional video.

Figure 5.3 Student response screen. Left: student's answer; right: the correct answer. The orange boxes show wrong answers by students on the assessment; the white boxes show correct answers.



Figure 5.4 Student progress screen. Top: the most recent assessments. Bottom: As more learning objectives are mastered, the green bars fill up for each learning domain.

## 5.5 METHOD

To collect validity evidence about the usability of GM, teachers and students used GM in a natural classroom setting for three months. A convergent, parallel mixed-methods design (Schoonenboom & Johnson, 2017) was used to examine its usability. We collected and analyzed data from log files, questionnaires, and interviews. The data triangulation provided a better understanding than would have been possible with the use of a single data source. The log files provided quantitative data on how the teachers actually used GM. The questionnaires provided primarily quantitative information on usability from a teacher's perspective. The qualitative data from the interviews with the teachers were used to further refine and interpret how the teachers used and experienced GM.

### 5.5.1 Respondents

In total, 29 teachers (teaching 582 students) from 20 primary schools voluntarily enrolled and participated in this study. Twenty-five of these participants completed the questionnaire, and interviews were conducted with all 29. The majority of the 29 teachers were female (75.8%), which is typical of the Dutch primary school teacher population ([www.onderwijsincijfers.nl](http://www.onderwijsincijfers.nl)). The teachers were from schools spread all over the country: six schools were located in the northern region, six in the middle region, and eight in the southern region. Teaching methods for arithmetic were used by 22 teachers, such as the method Wereld in Getallen (13 teachers). The other seven teachers put together (digital) materials themselves. From the 29 teachers, 24 had homogeneous classes, while five had heterogeneous classes (combining two or three grades). Four teachers also participated in the evaluation of prototype 2.

### 5.5.2 Instruments

#### *Log files*

We collected digital recordings of the teachers' actions. These recordings were unobtrusive, making it suitable for collecting non-reactive data about the current use of GM. The system logged every click that the teachers made for navigational purposes. From all the clicks made, we used the recording of the following actions: 1) filtering within the learning trajectory, 2) reviewing students' answers, 3) overruling an assigned status, and 4) planning learning objectives. There were 4,435 records in total.

#### *Questionnaire*

We developed a questionnaire to measure the teachers' perspective on the usability of GM. The questionnaire was divided into five parts: 1) overall impression of usability, 2) usability regarding "where the students need to go," 3) usability regarding "where the students are," 4) usability regarding "how best to get there," and 5) background items. It consisted of multiple-choice items, multiple-response items, open-answer items, and items on a 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5). Each part of the questionnaire ended with a box for additional comments (if any). Table 5.2 provides the number of items and an example item for each part of the questionnaire.

Table 5.2 Detailed information on questionnaire subjects

Subject	n	Example items
Overall impression	5	GM is usable for establishing where the students are in their learning.
Where do the students need to go?		
Learning objectives	2	The learning objectives are clearly formulated.
Visualization of the learning trajectory	4	I understand the symbols for the learning domains.
Where are the students in their learning?		
Assessment tasks	5	The digital tests enable me to establish students' learning strategies and misconceptions.
Visualization of students' progress	5	The student screen provides good understanding of students' progress.
Overruling the assigned status	1	GM allows the possibility to change the assigned status, for example, from orange (non-mastery) to green (mastery). This possibility is useful for me.
How best to get there?		
Instruction	1	GM needs to offer more instructional support if a learning objective has not been mastered.
Planning	2	It is easy to plan a learning goal for one or a few student(s).
Background	14	Did you experience technical problems during the use of GM?

### *In-depth interviews*

To obtain a richer understanding of the usability of GM, we conducted in-depth interviews in which the teachers' interpretations and uses of GM were investigated. The teachers were asked to analyze a student's performance from their own data and to indicate possible follow-up actions. Furthermore, they were asked to explain their experiences of using GM. An example item was: "What do you conclude on the basis of this assessment result regarding mastery/non-mastery and students' learning strategies and misconceptions? Does this information meet your expectations?"



### 5.5.3 Procedure

Through a national open call, the schools were invited to voluntarily participate in the study. After registration, the schools received log-in codes and the assignment tasks on paper. They also received a concise manual explaining the technical use of GM. The teachers were asked to implement GM themselves in their classroom. Interviews were held with all 29 teachers during their use of GM, which took place either physically at the school or by telephone. The questionnaire was distributed online at the end of the pilot and was completed by 25 teachers.

### 5.5.4 Data Analysis

#### *Log files*

From the 4,435 records in the log files, we analyzed the frequency with which each teacher used the filter options, reviewed the students' answers, overruled an assigned status, and planned learning objectives. In terms of filtering options, we examined filtering by learning domain; by grade level, which meant selecting successive learning objectives on the learning trajectory path; and by search field. In relation to reviewing the students' answers, we distinguished between reviewing right and wrong answers. In terms of overruling, we distinguished overruling a non-mastery to a mastery status or vice versa. For the planning of learning objectives, we differentiated between planning for individual students or for a group of students.

#### *Questionnaire*

The data from the questionnaires were analyzed both qualitatively and quantitatively. The quantitative analysis consisted of frequency analyses of the multiple-choice items and multiple-response items. We calculated the median (*Mdn*) and interquartile range (*IQR*) to describe the distribution of the ordinal data from the Likert scales. The answers to the open-answer questions were coded in a qualitative way, together with the responses to the interview questions. We compared the answers and then grouped related pieces of information into categories. We subsequently used these categories to classify all the answers. If an answer did not fit into the existing categories, the framework was modified and the process repeated.

### *In-depth interviews*

The transcribed responses to the interview questions were coded in a qualitative way, as described above. The suggestions for improving GM were categorized according to three layers of necessity. Suggestions labeled “must have” were critical for the successful implementation of formative assessment. Suggestions labeled “should have” were also important for implementation but were not specific to formative assessment. These concerned suggestions that were also relevant for other assessment contexts, for example, the suggestion to provide a read-aloud function for dyslexic children. The label “could have” contained desirable improvements that could improve user-friendliness. However, these improvements were not conditional upon adequate interpretation and use of the assessment results.

## 5.6 RESULTS

Overall, the teachers perceived GM as usable for establishing where students were in their learning ( $Mdn = 4$ ,  $IQR = 4 - 4$ ) and where they needed to go ( $Mdn = 4$ ,  $IQR = 3 - 4$ ). They were neutral about the usability regarding how best to get there ( $Mdn = 3$ ,  $IQR = 2 - 3$ ). Some positive elements of GM were its clear design, its direct feedback about students’ mastery of learning objectives, the possibility to adapt it to individual students, its attractiveness for students, and the autonomy for both teachers and students. To illustrate, teacher 12 declared: *I am really very happy about this. This is such a nice way to really see what you need to do with the students, where they are, and what problems they encounter. I think it is a really nice measuring tool.*

General suggestions for improvement were related to the user-friendliness, design clarity, the required language skills for reading the items, the program manual, and the technical implementation. For example, 84% of the teachers experienced digital problems during the try-out. The problems mentioned included that the assessment results were not saved, the teacher could not review the assessment results, or that the program crashed or became very slow. For example, teacher 4 mentioned: *“The program crashed regularly (...), the students closed the program and then nothing was saved. Or it was saved, but then the students could not see the results by themselves.”*

### 5.6.1 Question 1: To What Extent Are the Assessment Results Usable for Teachers in Establishing Where Students Need to Go?

#### *Learning objectives*

The teachers felt that the learning objectives were clearly formulated ( $Mdn = 4$ ,  $IQR = 4 - 5$ ). In addition, they used the explanatory description of the learning objective (92%), the item example (84%), and the instructional video (64%) to further understand the learning objective. This result was confirmed in the interview data. While teacher 23 would have liked to see all the test items, the other teachers were satisfied with the current item example.

#### *Visualization of the learning trajectory*

According to the questionnaires, the teachers agreed that GM enabled them to understand the learning trajectories ( $Mdn = 4$ ,  $IQR = 4 - 4.5$ ), as it clearly indicated the grade and learning domain to which a learning objective belonged ( $Mdn = 4$ ,  $IQR = 4 - 5$ ). However, the symbols indicating the various learning domains were not clear to all the teachers ( $Mdn = 3$ ,  $IQR = 2 - 4$ ). In the interviews, the teachers spoke of a lack of clarity regarding the learning trajectory, which was caused by the presentation of a long list characterized by a lack of a hierarchical structure, an illogical sequence of learning objectives, and unclear symbols for the learning domains. For example, teacher 11 noted: *“It is only that there is a lot on one screen. During planning, I thought ‘there is no end on the screen.’ You can go down and down. That is why I prefer tabs. If you put one learning domain on one tab, then it becomes more organized.”*

The different options for filtering within the learning trajectory line were met with appreciation. Altogether, 96.4% of the teachers liked filtering by learning domain; 89.3% liked filtering by grade level and by related learning objectives in the learning trajectory; and 52.6% liked filtering by search field. The log files show that filtering by grade level (38.9%) and learning domain (26.7%) were the most frequently used filters of all the filtering recorded. Two teachers suggested the possibility of adding multiple domains and groups to the filter at the same time. Table 5.3 presents suggestions for improving GM in terms of the visualization of the learning trajectory.

Table 5.3 Suggestions for improving GM regarding “where the students need to go”

Suggestions	Must have	Should have	Could have
Visualization of the learning trajectory			
- Provide more structure in the list of learning objectives through sub-headings and overarching goals	x		
- Use more self-explanatory symbols and figures to indicate learning domains		x	
- Enable the possibility of filtering multiple domains and groups at the same time.			x

### 5.6.2 Question 2: To What Extent Are the Assessment Results Usable for Teachers in Establishing Where Students Are?

#### Assessment tasks

The teachers indicated that the digital tests ( $Mdn = 4$ ,  $IQR = 4 - 4$ ) and assignment tasks ( $Mdn = 4$ ,  $IQR = 2.5 - 4$ ) could be used to establish the students' mastery of a learning objective. The items represented the learning objective, and both tests were highly appealing for students in terms of demonstrating their performance.

The log file data showed that 89.7% of the teachers examined their students' answers on the digital tests. They only looked at wrong answers, which were usually unexpectedly wrong, according to the interview results. For example, teacher 25 said: “I do not have enough time to examine all answers, but I especially reviewed answers from students from whom I think: ‘Why does he perform so poorly?’”

The examination of the students' answers was insufficient to determine their learning strategies and misconceptions ( $Mdn = 2$ ,  $IQR = 2 - 3$ ). The teachers indicated that the digital tests should provide more in-depth feedback to teachers and students. For example, teacher 1 suggested: “Students got feedback about right or wrong. However, it would also be useful to provide a calculation of the right answer. How do you calculate it?” Teacher 12 suggested the use of multiple-choice distractors to provide more information about the students' thinking.

In contrast to the feedback on the digital tests, the students' answers in the assignment tasks provided sufficient feedback about their learning strategies and misconceptions ( $Mdn = 4$ ,  $IQR = 3 - 5$ ). Teacher 11 wrote: “The digital test does not allow me to understand students' thinking process, while the assignment tasks do.” However, seven teachers pointed out that these tasks were highly time-consuming and difficult to schedule in the curriculum. Furthermore, five teachers specifically

pointed to the need for additional support in assessing assignment tasks. The need for more support by a scoring rubric was also indicated in the questionnaire ( $Mdn = 4$ ,  $IQR = 3.5 - 5$ ).

#### *Visualization of students' progress*

The teacher dashboard provided a clear view of the students' progress ( $Mdn \geq 4$ ,  $IQR = 4 - 5$ ), but the teachers indicated that the meaning of the domain progress bar in this dashboard was unclear. For example, a teacher maintained: *"It does not mean much to me yet. The student performed well in this domain and in this domain too. However, is it true that if the bar is completely full, then the student has mastered grade 3 and higher grades? That is not clear to me."* In the interviews, the teachers made a number of other suggestions for improvement, which are presented in Table 5.4.

Table 5.4 Suggestions for improving GM regarding "where students are in their learning"

Suggestions	Must have	Should have	Could have
<b>Assessment tasks</b>			
- Provide more in-depth feedback about students' learning strategies and misconceptions in the digital test	x		
- Provide a read-aloud function for dyslexic children in the digital test so as to reduce the required language skills		x	
- Provide a scoring rubric for assessing students' performance on the assignments		x	
<b>Visualization of students' progress</b>			
- Clarify the meaning of the learning domain progress chart		x	
- Specify the information about students' (non-)mastery in the dashboard, including the number of times the assessment was done, the number of errors recorded, space for observation notes, etc.			x
- Enable the possibility of printing students' progress for administration			x
- Show students' names in the plan screen of the teacher's dashboard			x
- Enable the possibility of adding one's own comments from student observations			x
- Enable the possibility of easily switching between student grades within a school			x

### *Overruling the assigned status*

The log file data show that 51.7% of the teachers did not overrule a mastery/non-mastery status and that 48.3% of them overruled a mastery/non-mastery status only a few times. In the questionnaire, the teachers strongly agreed that the possibility to overrule was very useful ( $Mdn = 5$ ,  $IQR = 4 - 5$ ), since a careless mistake by a student can easily occur. However, based on the interviews, 11 teachers would have liked to re-assess the student in order to allow the student the opportunity to experience success, to motivate the student, or because they were convinced that non-mastery was always indicative of inadequacy on the part of the student.

### *5.6.3 Question 3: To What Extent Are the Assessment Results Usable for Teachers in Establishing How Best to Get There?*

#### *Instruction*

If a learning objective has not been mastered, then the teacher needs to select appropriate instructional actions. Fifty percent of the teachers conveyed a preference for additional support from GM, of which 25% indicated that GM itself should offer instructional material, with another 25% indicating that a link to instruction material was sufficient. Only 25% of the teachers did not require further support in selecting instructional actions. The remaining 25% would have liked additional support but did not see it as a major issue.

#### *Planning*

If a learning objective has been mastered, teachers could plan new learning objectives. The log files show that 22 teachers planned separate learning objectives for the whole group and individual students. Four teachers planned only for the whole group because they wanted to try out the program in this way. Three teachers planned only for individual students because they were yet to discover the button for group planning. All the teachers agreed, however, that it was easy to plan learning objectives for individual students ( $Mdn = 4.5$ ,  $IQR = 4 - 5$ ) and for whole groups of students ( $Mdn = 5$ ,  $IQR = 4 - 5$ ). They made some suggestions for improvement (Table 5.5), for example: “*The user-friendliness can be improved by being more easily able to plan a learning objective for multiple students or re-planning a learning objective with greater ease*” (Teacher 1).

Table 5.5 Suggestions for improving GM regarding “how best to get there”

Suggestions	Must have	Should have	Could have
Instruction			
- Provide (a link) to instructional materials	x		
Planning			
- Enable the possibility to plan a learning objective for multiple students simultaneously			x
- Provide a button for cancelling a planned learning objective for the whole group			x
- Enable the possibility to plan multiple learning objectives for an individual student simultaneously			x

## 5.7 CONCLUSION AND DISCUSSION

In this study, validity evidence was collected to support assumptions regarding the intended use of formative assessment. The validity evidence focused on the question of whether assessment results are usable for teachers’ formative assessment practices. We formulated initial design principles from our conceptual framework and used the formative assessment system GM to critically evaluate whether these principles were met. The results show that GM was largely usable for establishing where the students were in their learning and where they needed to go and that it was somewhat usable in determining how best to get there. For example, the learning objectives were clearly formulated; the assessments could be used for establishing the students’ mastery of a learning objective; but the teachers had difficulty linking the results to instructional support.

There were also suggestions for the further development of GM, including a number of generic suggestions such as the provision of a read-aloud function for dyslexic children. Furthermore, there were a number of suggestions for improving the user friendliness of GM, such as displaying students’ names on the dashboard. Finally, some suggestions were deemed necessary for the successful implementation of formative assessment, which we formulated as substantive design principles for the development of formative assessment instruments.

First, if we want teachers to establish where students need to go in their learning, then a visualization of the learning trajectory is needed, which clearly shows the relationship between learning objectives. The current study showed that the visualization of the learning trajectory was unclear. Since other studies

have concluded that teachers have limited understanding of how knowledge and practice develop within a domain (e.g., Schneider & Andrade, 2013), the instrument needs to provide a clear visualization. Several existing studies have emphasized the importance of learning trajectories for teachers' formative assessment practices (Alonzo, 2011; Furtak & Heredia, 2014; Shepard, 2018). In this study, the teachers suggested the structuring of the learning objectives through sub-headings and overarching goals. This is in accordance with the design implications delineated by Heritage (2008), who suggested providing a big picture, multi-year progression that outlines overarching core ideas, from which more detailed descriptions are generated.

Second, if we want teachers to establish where students are in their learning, then in-depth feedback is needed in terms of students' learning strategies and their misconceptions regarding learning objectives. The need for fine-grained information has also been shown by Hopster-den Otter, Wools, Eggen, and Veldkamp (2017). Goertz, Olah, and Riggan (2009) suggested the use of multiple-choice item distractors to provide information about common errors in students' understanding. However, the current study showed that teachers have difficulty determining students' misconceptions and learning strategies on digital tests. Unless teachers collect additional information themselves, for example, in a follow-up discussion with the student, then the instrument needs to provide a fine-grained analysis of the misconceptions and learning strategies. This goes beyond showing a comparison between students' responses and the correct responses, such as in the current prototype of GM.

Third, if we want teachers to establish how best to enable students to get to their desired state of learning, then (a link to) instructional materials should be provided. The current study showed that GM provided insufficient guidance about selecting instructional activities and appropriate feedback. Although GM has the same learning objectives as most teaching methods, teachers have difficulty linking the assessment results to instruction. Several studies have showed that the translation from assessment results into actions seems to be challenging (Lai & Schildkamp, 2013; Marsh, 2012). It has been concluded that this transformation is a complex process, whereby teachers combine assessment results with other sources of information as well as their own expertise (e.g., Kippers, 2018). Given the complexity of this process, the question is whether a more direct link is sufficient to enable teachers to adapt their teaching repertoires to students' needs.



Therefore, it is important to note that the formulated design principles are focused on the development of formative assessment instruments. The process of using the results from instruments, however, is just as critical for the implementation of formative assessment. For this process, teachers need knowledge and skills regarding the use of assessment results, including subject-matter knowledge and pedagogical content knowledge (Falk, 2012; Gummer & Mandinach, 2015; Heritage et al., 2009; Sabel, Forbes, & Zangori, 2015). Several studies have mentioned the necessity of providing long-term professional development for teachers (e.g., Aschbacher & Alonzo, 2006; Heitink, Van der Kleij, Veldkamp, Schildkamp, & Kippers, 2016; Kippers, Poortman, Schildkamp, & Visscher, 2018; Popham, 2009; Timperley, 2009). For example, Heritage et al. (2009) argued that the visualization of learning trajectories in a formative assessment instrument is not sufficient. In order to make appropriate decisions that meet learners' needs, teachers need deep knowledge about how the concepts in a domain develop. To conclude, a thoughtful integration of both the instrument and the process are needed for a successful implementation of formative assessment.

### *5.7.1 Limitations of the Study*

This study evaluated GM in an authentic setting, which provided the opportunity to analyze the actual use and experiences of teachers. Although research in authentic settings has its benefits, it also raises a number of real-world challenges (McKenney, Nieveen, & van den Akker, 2006; Plomp, 2013). First, the teachers learned to use the program during the try-out. Their perspective on usability and how to use GM could thus have changed over time, thus requiring some caution in interpreting the results. For example, some teachers discovered the filtering options within the learning trajectory only after using the program for a month, making the use of the program significantly more intensive from that point. Second, the results showed that 84% of the teachers experienced digital problems during the try-out. These problems imply the need for caution in the interpretation of the log file results. Cognizant of possible threats to the study, we used different methods of data collection (log files, questionnaires, and interviews) as a main precaution to mitigate them.

### *5.7.2 Implications for Future Research*

The current study serves as an example of how to collect validity evidence regarding the intended use of embedded formative assessment. Since the actual use of formative assessment is essential to its effectiveness, a critical evaluation of its usability with intended users is needed. Moreover, the findings of the study enhance our understanding of the characteristics that make formative assessment instruments usable for these users.

The results and limitations can inform future research. First, the findings provided suggestions for improving the usability of the GM system itself. It seems worthwhile to examine whether the suggestions for improvement might lead to more support of the underlying assumptions. Second, the study evaluated usability from a teacher's perspective. Since students are also an essential part of formative assessment (Black & Wiliam, 2009), we suggest research on how students use and experience GM. For example, how do they understand the learning objectives? Are they able to analyze their mistakes, or do they also need more in-depth feedback? This suggestion is in accordance with the trend toward activating students as self-directed learners (Wiliam & Thompson, 2007). Third, further validation studies are needed to validate other assumptions underlying formative assessment as well as to combine them in a coherent validity argument. For example, the impact of formative assessment on students' learning processes and outcomes is still under debate (e.g., Bennett, 2011; Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012). Future research is needed to support the assumption that embedded formative assessment contributes to student learning.

### *5.7.3 Implications for Practice*

The results of this study show how we can support teachers in using evidence from formative assessment. This can be used in the design of GM and other formative assessment instruments. Practical implications include a clear visualization of the learning trajectory, in-depth feedback about students' learning strategies and misconceptions, and (a link to) instructional materials. This way, formative assessment instruments would support greater use of assessment results, thereby contributing to the potential of formative assessment as a way of supporting student learning.

## 5.8 REFERENCES

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Alonzo, A. C. (2011). Learning progressions that support formative assessment practices. *Measurement: Interdisciplinary Research & Perspective*, 9, 124–129. doi:10.1080/15366367.2011.599629
- Aschbacher, P., & Alonzo, A. (2006). Examining the utility of elementary science notebooks for formative assessment purposes. *Educational Assessment*, 11(3), 179–203. doi:10.1207/s15326977ea1103&4\_3
- Béguin, A. A. & Straat, J. H. (2019). On the number of items in learning goal mastery testing. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 121-134). Cham, Switzerland: Springer International Publishing.
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18, 5–25. doi:10.1080/0969594X.2010.513678
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31. doi:10.1007/s11092-008-9068-5
- Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and conclusions about efficacy of formative assessment. *Educational Measurement: Issues and Practice*, 31(4), 13–17. doi:10.1111/j.1745-3992.2012.00251.x
- Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.
- Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice*. (pp. 43–62). New York/Amsterdam: Teachers College Press.
- Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, NJ: Pearson Education.
- Brown, G. T. L., O’Leary, T. M., & Hattie, J. A. C. (2019). Effective reporting for formative assessment: The asTTle case example. In D. Zapata-Rivera (Ed.). *Score reporting research and applications* (pp. 107–125). New York and London: Routledge.
- Centraal Bureau voor de Statistiek (CBS), Dienst Uitvoering Onderwijs (DUO) en het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) (n.d.). *Onderwijs in cijfers* (Education

- in numbers). Retrieved from <http://www.onderwijsincijfers.nl>
- Cito. (n.d.). Groeimeter. Retrieved from <https://www.cito.nl/kennis-en-innovatie/citolab/citolab-projecten-po/citolab-po-groeimeter>
- Falk, A. (2012). Development in elementary science: Reciprocal relations between formative assessment and pedagogical content knowledge. *Science Education*, 96, 82–99. doi:10.1002/sce.20473
- Forbes, C. T., Sabel, J. L., & Biggers, M. (2015). Elementary teachers' use of formative assessment to support students' learning about interactions between the hydrosphere and geosphere. *Journal of Geoscience Education*, 63, 210–221. doi:10.5408/14-063.1
- Furtak, E. M. (2006). *Formative assessment in K-8 science education: A conceptual review*. Paper commissioned for the Committee on Science Learning, Kindergarten through Eighth Grade, National Research Council. Retrieved from [http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse\\_080104.pdf](http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_080104.pdf)
- Furtak, E. M., & Heredia, S. C. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching*, 51(8), 982–1020. doi:10.1002/tea.21156
- Goertz, M. E., Olah, L. N., & Riggan, M. (2009). Can interim assessments be used for instructional change? Policy brief. RB-51. *CPRE Policy Briefs*. Retrieved from [http://repository.upenn.edu/cpre\\_policybriefs/39](http://repository.upenn.edu/cpre_policybriefs/39)
- Gummer, E., & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record*, 117(4), 1–22.
- Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3), 365–379. doi:10.1080/0969594970040304
- Hattie, J. (2005, August). *What is the nature of evidence that makes a difference to learning?* Paper presented at the Australian Council for Educational Research Conference on Using Data to Support Learning, Melbourne, Australia.
- Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, 17, 50–62. <https://doi.org/10.1016/j.edurev.2015.12.002>
- Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Chief Council of State School Officers.
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. L. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31. doi:10.1111/j.1745-3992.2009.00151.x

- Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timss, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Technical Report 703). Los Angeles, CA: CRESST.
- 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. doi:10.1016/j.stueduc.2016.11.002
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. doi:10.1111/jedm.12000
- Kippers, W. B. (2018). *Formative data use in schools: Unraveling the process* (Doctoral dissertation). Enschede, the Netherlands, University of Twente. doi:10.3990/1.9789036546720
- Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention? *Studies in Educational Evaluation*, 56, 21–31. doi:10.1016/j.stueduc.2017.11.001
- Lai, M. K., & Schildkamp, K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 9–21). Dordrecht: Springer. doi:10.1007/978-94-007-4816-3\_1
- Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, 47(2), 71–85. doi:10.1080/00461520.2012.667064
- Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, 114(11), 1–47.
- McKenney, S., Nieveen, N., & van den Akker, J. (2006). Design research from a curriculum perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 62–90). London: Routledge.
- McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research*. Abingdon/New York: Routledge Taylor & Francis Group.
- National Research Council. (2001). *Classroom assessment and the national science education standards*. Washington, DC: National Academy Press.
- Nieveen, N., & Folmer, E. (2013). Formative evaluation in educational design research. In T. Plomp & N. Nieveen (Eds.), *Educational design research* (pp. 152–169). Enschede, the Netherlands: Netherlands Institute for Curriculum Development (SLO).
- Noteboom, A., Aartsen, A., & Lit, S. (2017). *Tussendoelen rekenen-wiskunde voor het primair onderwijs: Uitwerking van rekendoelen voor groep 2 tot en met 8 op weg naar streefniveau 1S* [Interim arithmetic-mathematics learning objectives for primary education: Description

- of arithmetic objectives for grade 0 through 6 on the way to target level 1S]. Enschede, the Netherlands: Netherlands Institute for Curriculum Development (SLO).
- Plomp, T. (2013). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational design research* (pp. 10–51). Enschede, the Netherlands: Netherlands Institute for Curriculum Development (SLO).
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, 48, 4–11. doi:10.1080/00405840802577536
- Sabel, J. L., Forbes, C. T., & Zangori, L. (2015). Promoting prospective elementary teachers' learning to use formative assessment for life science instruction. *Journal of Science Teacher Education*, 26(4), 419–445. doi:10.1007/s10972-015-9431-6
- Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action: Conclusions drawn from a special issue on formative assessment. *Applied Measurement in Education*, 26, 159–162. doi:10.1080/08957347.2013.793189
- Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design. *Kolner Zeitschrift für Soziologie und Sozialpsychologie*, 69, 107–131. doi:10.1007/s11577-017-0454-1
- Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, 31(2), 165–174. doi:10.1080/08957347.2017.1408628
- Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50(1), 99–104. doi:10.1111/jedm.12005
- Stiggins, R. J. (2001). *Student-involved classroom assessment* (3rd ed.). Columbus, OH: Merrill Prentice Hall.
- Timperley, H. (2009, August). *Using assessment data for improving teaching practice*. Paper presented the conference Assessment and Student Learning: Collecting, Interpreting and Using Data to Inform Teaching, Perth, Australia.
- Van Zanten, M., & Van den Heuvel-Panhuizen, M. (2018). Primary school mathematics in the Netherlands: The perspective of the curriculum documents. In D. R. Thompson, M. A. Huntley, & C. Suurtamm (Eds.), *International perspectives on mathematics curriculum* (pp. 9–39). Charlotte, NC: Information Age Publishing Inc.
- Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.
- Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, New Jersey: Lawrence Erlbaum Associates.



# CHAPTER 6





# THE VISUAL PRESENTATION OF A LEARNING TRAJECTORY<sup>1</sup>

While research shows the potential of learning trajectories for formative assessment, there has been a dearth of attention on their visualization. As teachers have shown difficulty understanding and using tabular learning trajectories, this chapter reports on two studies investigating the appropriateness of a graphical visualization. A prototype of a graphical visualization within an embedded formative assessment platform for arithmetic was used to collect data. Study A aimed to validate the graphical structure of the prototype by modelling student performance data using a Bayesian network analysis, and an assessment was administered to 787 students. The results showed multiple conditional dependencies in the data, confirming the graphical structure. Study B examined teachers' understanding and preferences regarding the graphical visualization. In total, 19 teachers used the prototype in a natural classroom setting for two months. The results from the interview data showed that the teachers had difficulty understanding the visualization in an appropriate manner. Several design principles for the graphical visualization were identified in relation to the content, structure, and usability of the learning trajectory.

Keywords: learning trajectory; formative assessment; visualization

<sup>1</sup>Dorien Hopster-den Otter, Remco Feskens & Saskia Wools

## 6.1 INTRODUCTION

There is widespread interest in developing and using learning trajectories or progressions (e.g., Confrey, Gianopulos, McGowan, Shah, & Belcher, 2017; Daro, Mosher, & Corcoran, 2011; Heritage, 2008). Learning trajectories (LTs) can be defined as empirically grounded frameworks about the likely progressions of students' reasoning regarding core constructs within a learning domain (Confrey, Maloney, & Corley, 2014; Corcoran, Mosher, & Rogat, 2009). This definition emphasizes that empirical research is used to validate the ways in which students' reasoning develops (Corcoran et al., 2009), that knowledge and skills in a learning domain are connected (Heritage, 2008), and that learning is conceived as a coherent process of increasing sophistication (Corcoran et al., 2009; Heritage, 2008). This differs from standards or curriculum frameworks that generally describe discrete objectives for the end of a grade level based on conventional wisdom and expert consensus (Bailey & Heritage, 2014; Kobrin, Larson, Cromwell, & Garza, 2015).

LTs can be developed for several purposes, including standards development, curriculum design, and formative assessment (Kobrin et al., 2015). Gotwals (2012) emphasized that an LT for one purpose may not be easily translated or used for another purpose. Different purposes result in different LT characteristics, for example, with regard to the time span and grainsize covered (Corcoran et al., 2009). These differences have implications for the LT's most appropriate use (Shepard, 2018). Therefore, there is a need to examine and evaluate the extent to which an LT is appropriate for a particular interpretation and use. This study explored the characteristics of LTs for formative assessment. Specifically, it investigated how to visualize an LT for this purpose.

### *6.1.1 LTs for formative assessment*

Formative assessment comprises three major questions (Broadfoot et al., 2002; Furtak, 2006; NRC, 2001; Wiliam & Thompson, 2007): (1) Where do the students need to go? (2) Where are they now? (3) How best to get there? The first question is about establishing the desired state of learning, such as the ultimate learning objective. The second question relates to identifying students' current state of learning, including determining their conceptions and misconceptions. The

third question uses the information gleaned from the first and second question to proceed students' learning from current to desired state.

To answer these questions effectively, teachers need a clear notion of how students' learning develops in a domain (Bennett, 2011; Daro et al., 2011; Heritage, 2008). This means that they need to understand the pathways along which students are expected to progress. However, many teachers are unclear about this continuum, making it difficult to locate students' current state and to then decide on instructional actions to move students' learning forward (Heritage, 2008; Schneider & Andrade, 2013).

Alonzo (2011) described the manner in which LTs might support teachers' formative assessment practices, arguing that LTs could serve as a road map during a journey. First, LTs present the final destination and several intermediate points toward establishing where the student is going. Second, LTs show which aspects of students' knowledge and skills might be particularly important. Formative assessments can be tied to these aspects, and the evidence elicited can help teachers in locating students' current learning status on the continuum of development. Third, LTs present suggestions for possible routes, which help students proceed to the final destination. Thus, LTs provide the big picture and can assist teachers in eliciting, interpreting, and responding to students' thinking.

### *6.1.2 Visualization of LTs*

There is a need for an understandable and useful visualization of LTs in order to support teachers' formative assessment practices (Daro et al., 2011; Gotwals, 2012; Ryan, 2006; Zapata-Rivera & Katz, 2014). This visualization is the bridge between the information captured by the LT and the decisions or actions of teachers. It should ensure that teachers are directly informed about their decisions.

Current LTs are typically depicted as a table, with core constructs in the columns and levels of achievement in the rows (Anderson, 2008). Each table cell is filled with the specific performances that belong to a particular core construct and achievement level. This way, students' learning process is shown as a relatively straightforward set of steps along a metaphorically linear path, starting at the so-called low anchor and identically developing their understanding to the upper anchor (Furtak, Thompson, Braaten, & Windschitl, 2012; Salinas, 2009).

Although tabular LTs initially seem quite simple and logical, teachers have difficulty using them for formative assessment (Hopster-den Otter, Wools, Eggen, & Veldkamp, 2019). This difficulty is illustrated by frequently asked questions from teachers, such as whether students jump ahead or whether they can work at more than one level at a time (Clements & Sarama, 2009). Moreover, the lack of visualizing connections between core constructs does little to support teachers in understanding how students' current performance fits within the larger LT and in deciding on the next instructional steps (Heritage, 2008; Hopster-den Otter et al., 2019).

These difficulties justify the need for a visualization that more explicitly presents how students' learning develops. Recent studies have suggested a graphical visualization as a promising way to communicate highly connected educational data (Kingston & Broaddus, 2017; Lobato & Walters, 2016; Salinas, 2009; Willcox & Huang, 2017). A graphical visualization allows for showing multiple possible paths between learning objectives. Such a visualization may reduce the need for teachers to make these connections on their own.

The current study explored the appropriateness of a graphical LT for the purpose of formative assessment. Two sub-studies were designed for investigation (Figure 6.1). Study A aimed to validate the graphical structure of the LT using student performance data. It investigated whether a graphic visualization was superior to a tabular visualization in better reflecting how student learning develops. Study B examined teachers' understanding and preferences of a graphical visualization for the purpose of formative assessment. In both studies, a prototype of an LT for arithmetic was used to collect data. This LT will be described in the next section.

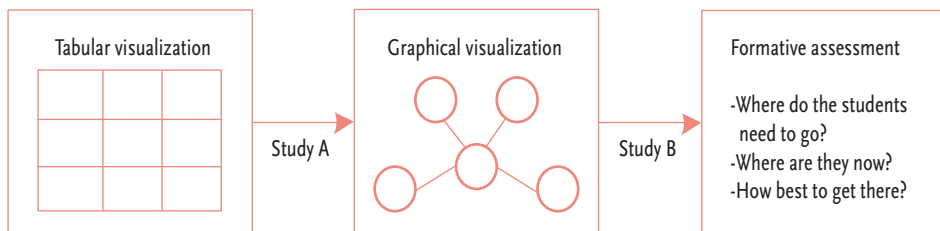


Figure 6.1 Conceptual Overview of Study A and Study B.

## 6.2 LT FOR ARITHMETIC IN PRIMARY EDUCATION

To address the study aim, we used a prototype of the arithmetic LT for primary education, as developed by the Cito Institute for Educational Measurement in the Netherlands. The arithmetic LT sketches learning pathways through which students (5–11-year olds) can achieve the standards determined for the end of primary school (Van Zanten & Van den Heuvel-Panhuizen, 2018).

The LT is developed by a bottom-up design approach, which means that the development starts from educational practice rather than from the scientific discipline (Heritage, 2008). Sources for the development include curricula, teaching methods, and literature about mathematics teaching (e.g., Verbeeck & Verschuren, 2010). Validation efforts were performed through focus groups comprising experts in mathematics (e.g., researchers, test developers, teaching method developers) and educational practice (e.g., teachers).

The LT consists of three hierarchical levels. The most fine-grained level entails the learning objectives in the arithmetic curriculum, as developed by the Netherlands Institute for Curriculum Development (Noteboom, Aarsten, & Lit, 2017). The specific formulation of the learning objectives enables teachers to link them to daily instruction. An example of a learning objective is “The student is able to split and group numbers up to 1,000 with hundreds, tens, and metric units.” The parent level of the learning objectives consists of core constructs. These are overarching learning units that cover several learning objectives. Examples of core constructs are quantities, number comprehension, addition and subtraction up to 100, and multiplication and division by 1,000. The core constructs are subdivided into a learning domain, for example, the core construct “quantities” belongs to the learning domain “numbers.” There are four learning domains: numbers, ratios, measuring and geometry, and data handling (Noteboom, 2017).

Within the level of the learning objectives and core constructs, prerequisite and co-requisite relationships are defined (Willcox & Huang, 2017). A prerequisite is another learning objective or core construct that must be achieved before starting the learning objective or core construct at hand. A co-requisite is a learning objective or core construct that can be met concurrently.

The LT is graphically visualized in a formative assessment platform called Groeimeter (GM), presented in Figure 6.2. For each of the four learning domains (1), the LT outlines the core constructs (2) and then drills down to the level of the learning objectives (3). The prerequisite relations between the core

constructs are visualized by arrows (4). The prerequisites between the learning objectives are visualized by an indented line (5). The co-requisite relations for both levels are shown by presenting the elements at the same vertical level (6).

The graphical visualization of the LT in GM is aimed at supporting teachers in their formative assessment practices. The learning domains and core constructs show the final destination and intermediate points that establish where the student is going. To determine a student's current level, a formative assessment is linked to each learning objective in GM. This assessment consists of a digital test or a hands-on assignment, depending on which of these is suitable for the learning objective to be measured. Each learning objective is clarified by means of an example of test item. The arrows and indented lines suggest possible pathways. For example, when the assessment information shows that a student has not mastered a learning objective, teachers would know which precursor learning objectives need to be developed to move that student forward. Similarly, teachers might focus their instruction on developing thinking aimed at higher learning objectives for students whose understanding outpaces that of their peers. Finally, GM allows teachers to vary from the suggested pathways by selecting or deselecting learning objectives from the LT. This way, teachers may determine their own sequence and adapt their instruction to students' needs.

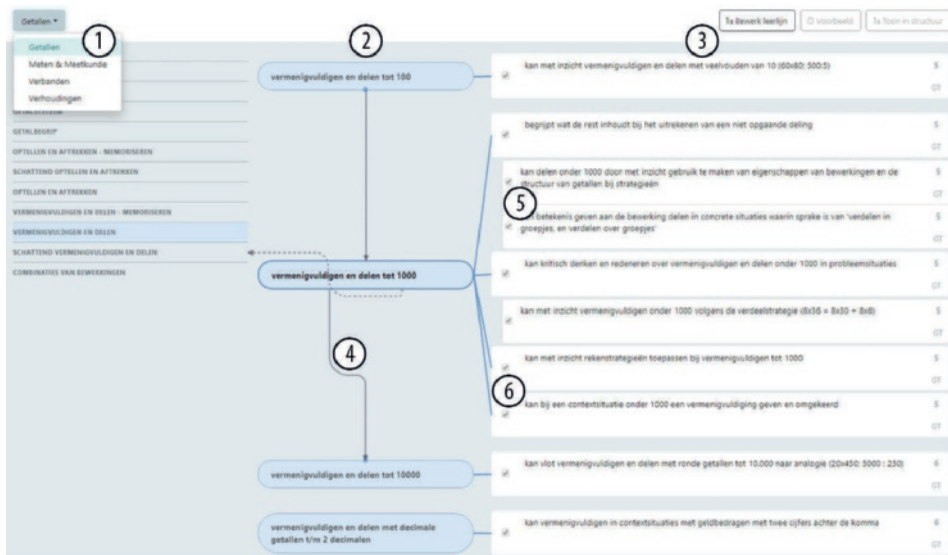


Figure 6.2 Graphical visualization of the LT in GM

## 6.3 METHOD STUDY A

### 6.3.1 Data collection

A digital test was developed to validate the graphical structure of the LT. The digital test covered 11 learning objectives, which were related to a core construct in the learning domain “Numbers” (Table 6.1). Students’ performance on this test could be used to analyze how the students were moving along the LT.

The digital test was a compilation of 11 tests from GM, each containing seven test items. This resulted in 11 tests times seven test items = 77 items in the digital test. Table 6.1 shows a description of the learning objectives covered by the 77 items. An example item for each learning objective is given in Appendix A.

Table 6.1 Description of the learning objectives

Code	Learning objective	Core construct
LO1	The student is able to count and count down from any number up to 1,000, even with jumps of 10 and 100.	Counting
LO2	The student is able to read, pronounce, and write numbers up to 1,000.	Concepts
LO3	The student is able to estimate, count, and show quantities up to 1,000, including by structuring.	Quantities
LO4	The student is able to compare and order (structured) quantities and numbers up to 1,000.	Quantities
LO5	The student is able to split and group quantities up to 1,000 with hundreds, tens, and metric units.	Number system
LO6	The student is able to indicate the position value of digits in numbers up to 1,000.	Number system
LO7	The student is able to split and group numbers up to 1,000 with hundreds, tens, and metric units.	Number system
LO8	The student is able to compare and order numbers up to 1,000 and place them on a number line.	Number system
LO9	The student is able to round off numbers to nearby hundreds.	Number system
LO10	The student is able to position numbers up to 1,000 with other numbers and compare the order of magnitude.	Number system
LO11	The student can indicate internal structures ( $4 \times 250 = 1000$ ) and external structures (998 is close to 1,000) from numbers up to 1,000.	Number system

The digital test was administered in two parts in order to maintain students' concentration. The first part consisted of 42 test items, measuring students' ability in terms of six learning objectives. The second part comprised 35 test items, covering five learning objectives. Furthermore, we developed eleven test versions to prevent an order effect. All test versions contained the same 77 items; however, each version started with test items from a different learning objective (see Appendix B).

The students completed the digital test individually at their school. They received a random assigned login number to make sure that their personal information remained anonymous. There was no time restriction for completion.

### *6.3.2 Data analysis*

The assessment results were scored automatically. A score of 0 indicated an incorrect answer, while a score of 1 indicated a correct answer. A score of 0 was also assigned to unanswered items. The eleven learning objectives were each measured with seven items. The cut score, demonstrating mastery of a learning objective, was set at an observed score of six for each of the learning objectives (Béguin & Straat, 2019). We started with descriptive analyses in order to provide insights into the nature of the items and learning objectives.

Moreover, we used an item response theory (IRT) model to test the appropriateness of a linear tabular structure. More specifically, we used the extended nominal response model (eNRM) implemented in the R package Dexter (Maris, Bechger, Koops, & Partchev, 2019). This model defaults to the well-known Rasch model in cases of dichotomous scored responses (Rasch, 1960). We first compared the model with Haberman's interaction model to assess how the Rasch model fit the data. We subsequently evaluated two important assumptions of the Rasch model: local independence and unidimensionality. Local independence means that the correlation between the item scores should disappear after controlling for the latent variable of interest. This assumption was evaluated by assessing the correlations between the item scores, conditional on the ability estimates. Unidimensionality means that only one latent factor should affect test performance. We evaluated this assumption by performing a factor analysis, which assesses how many factors are relevant in explaining the variation in the test scores. Furthermore, we estimated an IRT model assuming one and two latent traits and compared the model fit of both models.



Since the strict theoretical assumptions underlying the IRT model might not fully capture all the dimensionality in the data, we investigated the appropriateness of a graphical structure. We used a Bayesian network (BN) approach to model the data of student performance because these models explicitly incorporate dependence in the data. BNs are defined as directed acyclic graphs consisting of nodes and directed edges. The nodes represent discrete variables, such as learning objectives or core constructs, while the edges represent direct conditional dependencies between the variables. The BN analyses were conducted using the R package Bnlearn (Scutari, 2019). We started by estimating confirmative and explorative network models. The confirmative model describes the relationships between learning objectives, as defined in the arithmetical LT, while the relations in the explorative model are fully determined by the student performance data. Since these models showed different conditional dependencies, we developed them using input from two mathematical experts. We showed both models to the first expert and asked this expert to define a set of whitelist edges (representing real dependencies that should be present in the graph) and a set of blacklist edges (corresponding to impossible relations that should not be presented). We used the list edges to produce the first consensus network model. Model averaging techniques were used to improve the reliability of this model. This means that the model was constructed by producing 200 networks from the (bootstrapped) samples of the data and keeping the edges that appear at least 50% of the time (Efron & Tibshirani, 1993; Scutari, Auconi, Caldarelli, & Franchi, 2017). The first consensus network model was presented to the second mathematical expert, and the process was repeated. This resulted in a second consensus network model. We evaluated the predictive accuracy of this model using 10-fold cross validation. Cross-validation assesses the degree to which the model can be used to predict outcomes of new, independent observations. Finally, we calculated the probability of some variables, given the other variables ( $P(x|e)$ ) to draw inferences from the second consensus model.

### 6.3.3 Participants

A total of 787 students participated in the study, 24 of whom did not complete the test. Thus, 763 students were used in the analyses, all of whom were in grade 3 and came from 29 schools around the country. They learned from five different arithmetic teaching methods: 57.8% used the teaching method *Wereld in Getallen*;

19.3% used *Alles Telt*; 13.9% used *Pluspunt*; 5.5% used *Getal en Ruimte*; and 3.5% used *Wizwijs*.

In addition, two mathematical experts contributed to the development of the consensus network model. The first expert has arithmetic knowledge in the context of assessment, while the second expert has arithmetic knowledge in the context of curriculum development.

## 6.4 RESULTS STUDY A

### 6.4.1 Descriptive analysis

The 763 students each answered 77 test items. This resulted in 58,751 responses, from which 38,470 were scored as correct (score 1) and 18,779 responses were scored as incorrect (score 0). There were also 1,502 unanswered items, which were scored as incorrect. No significant differences with respect to the observed test scores were found between the 11 test versions,  $F(10) = 0.52$ ,  $p = 0.88$ . The different teaching methods did impact the observed test scores  $F(4) = 5.52$ ,  $p < .01$ . The 27 students following the learning method *Wizwijs* performed substantially better in learning objective LO11. It is important to note that these students were from a single school, so it is unclear whether these performance differences were caused by the school or by the teaching method. The performance on the other learning objectives was comparable across students following different teaching methods.

On average, the proportion of students who answered the items correctly was 0.65 ( $SD = 0.23$ ). Three items had a p-value below 0.2; thus, less than 20% of the students answered this item correctly. Since these items were too difficult for our target population, we removed them from further analysis.

Figure 6.3 shows the percentage of students who mastered and those who did not master each of the learning objectives. LO2 was mastered by 75% of the students, while LO11 was mastered by only 12% of them. Therefore, some of the learning objectives were more easily mastered than others.

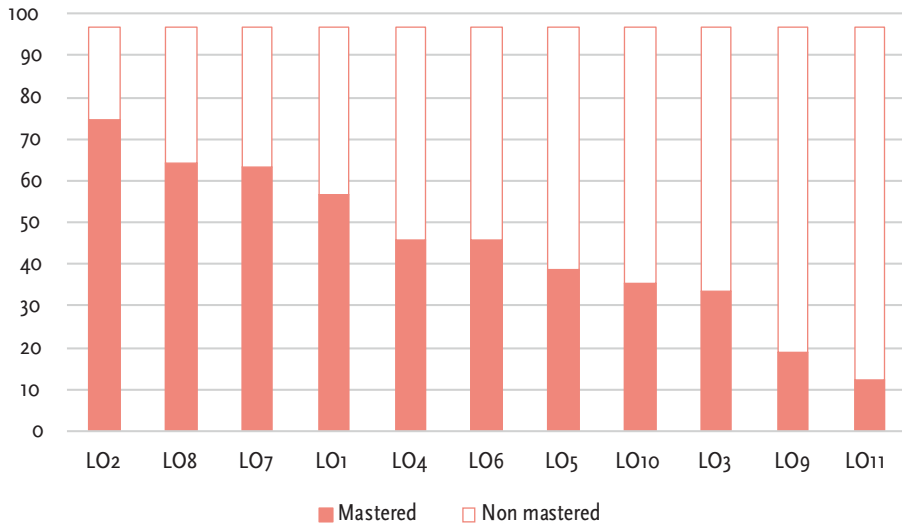


Figure 6.3 Percentage of the students who mastered and did not master a learning objective

#### 6.4.2 Tabular visualization

The evaluation of item fit was carried out by comparing the item regression curves of Haberman's interaction model and the Rasch model. The visual evaluation of both of these IRT models indicated that the Rasch model fit the data.

For the assumption of local independence, no more than five percent of the conditional correlations should be significant, and the distribution of p-values should be evenly distributed. However, the results showed that 32% of the item scores that correlated significantly and the p-values were not equally distributed. Therefore, the condition of local independence seems to have been violated.

For the assumption of unidimensionality, only one factor should be relevant in explaining the variation of test scores. However, the factor analysis showed that there were eight factors with an eigenvalue  $> 1$ . This means that multiple factors explained the variation of test scores. This evidence was further corroborated by estimating an IRT model that assumed one and two latent traits and comparing the model fit of both models. The results showed that the two-dimensional model (BIC = 51387.3) had a better model fit than the unidimensional model (BIC = 52627.5). This also confirmed that the unidimensionality assumption underlying the IRT was not met.

Since the strict theoretical assumptions underlying the IRT were not maintained, we concluded that a tabular visualization might not fully capture all the dimensionality in the data.

### 6.4.3 Graphical visualization

A BN approach was used to investigate the appropriateness of a graphical visualization. Student performance data were modelled using the prior knowledge of two experts as well as model averaging techniques, as described in sub-section 6.3.2. This resulted in the consensus network model shown in Figure 6.4.

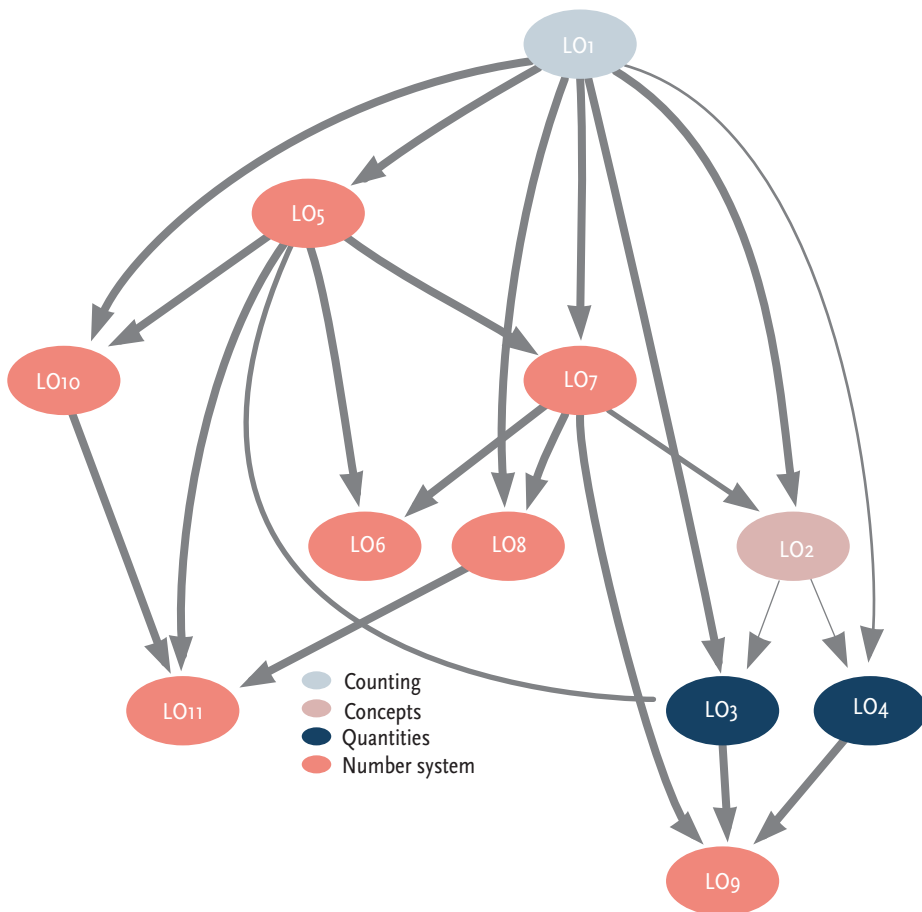


Figure 6.4 Consensus network model

In order to validate the consensus network model, the predictive accuracy was evaluated using 10-fold cross validation. The results showed that the prognosis was accurately predicted with a probability of 0.81 for LO9. All the directions of the edges seemed to be well established, which could probably be attributed to the use of the whitelist and blacklist, as they forced the directions of the nearby edges to cascade into place.

The consensus network model showed multiple dependencies between the learning objectives, confirming the appropriateness of a graphical structure. Moreover, the edges represented direct conditional dependencies between the learning objectives, which could be used by teachers to make inferences from the model. For example, Figure 6.4 shows that mastery of LO1 was a prerequisite for all other learning objectives. This means that the teachers needed to start by teaching number counting before teaching about the comparing, ordering, and splitting of numbers. Furthermore, it seems that LO2 did not strongly determine the mastery of any other learning objective. To illustrate, the chance of mastering LO4, given the mastery of LO1, was 0.57, while the chance of mastering LO4, given the mastery of LO1 and LO2, was 0.58. Such a small contribution of LO2 was also shown for the mastery of LO3 ( $P(\text{LO3}|\text{LO1}) = 0.47$ ,  $P(\text{LO3}|\text{LO1}, \text{LO2}) = 0.49$ ). This indicates that reading, pronouncing, and writing numbers did not seem to be a strong requirement for being able to calculate with them. LO9 and LO11 seemed to be the most difficult learning objectives. If a student did not master LO9, for example, it seemed most sensible to analyze the mastery of LO3, LO4, and LO7 ( $P(\text{LO9}|\text{LO3}, \text{LO4}, \text{LO7}) = 0.39$ ). Although this might still look like a low probability, overall, LO9 was difficult to master. Knowing which of the easier learning objectives are useful to master might be useful for formative assessment practice, as this knowledge would help increase the probability of mastering the more difficult learning objectives. Finally, the consensus model showed that the core constructs were not conditional.

## 6.5 METHOD STUDY B

### 6.5.1 *Data collection*

Although a graphic visualization better reflects student performance than a tabular visualization, we should also know whether teachers understand and prefer such a visualization. Therefore, Study B investigated teachers' understanding and preferences of the graphical visualization in GM by means of individual semi-structured interviews. The interview schedule consisted of two main parts (Appendix C). After an introduction in which the purpose of the interview was explained and background characteristics collected (e.g., grade level, teaching method), we continued by evaluating the teachers' understanding of the graphical LT. Their understanding was measured by three questions. The first question was about establishing where a student needs to go. We selected the core construct "number system up to 1,000" in the LT and asked the teachers which subsequent core construct they would choose if the selected core construct was mastered. The second question concerned interpreting students' current level. We selected a learning objective from the LT: "the student is able to split and group numbers up to 1,000 with hundreds, tens, and metric units." We then asked them which other learning objectives they expected the student to master if the current learning objective was mastered. The third question was about establishing the next learning objective towards mastering a core construct. We selected the same learning objective as in the second question and asked the teachers to decide on a subsequent learning objective. In the second part of the interview, we evaluated the teachers' preferences regarding the graphical LT. We asked them which aspects of the visualization provided understanding about the LT and which aspects could improve this understanding. The interviews were held by telephone, each lasting about 20 minutes.

### 6.5.2 *Data analysis*

All interviews were audio-recorded and transcribed verbatim. The transcriptions were then subjected to conventional content analysis (Hsieh & Shannon, 2005). This is an inductive way of category development, which means that categories emerged from the data during the analysis. For each transcript, we identified categories of understanding as well as categories of experience. The identification

was an iterative process whereby categories were assigned, combined, and split to ensure consistency and data coverage.

The final coding scheme (Appendix D) was used to double-code five transcripts, after which differences were discussed and ambiguities in the coding scheme clarified. Thereafter, two transcripts were double-coded. An inter-rater reliability analysis was performed on these transcripts to determine the consistency between the two raters. The agreement rate between the two researchers was 80.5% (Cohen's  $K = .777$ ), which is substantial (Landis & Koch, 1977).

### 6.5.3 Participants

In total, 19 teachers participated in the study, all of whom used the graphical LT visualization of GM in a natural classroom setting for a period of two months. Their experiences provided the basis to explore teachers' understanding and views regarding a graphical visualization. These 19 teachers came from 14 primary schools around the country, some of whom taught homogeneous classes ( $n = 17$ ) and others heterogeneous classes ( $n = 2$ ). Teaching methods for arithmetic were used by 17 teachers, such as the method *Wereld in Getallen* ( $n = 11$ ). Two teachers put together (digital) materials themselves.

## 6.6 RESULTS STUDY B

### 6.6.1 Teachers understanding

#### *Where do the students need to go?*

The teachers observed different parts of the LT in terms of their decision around the core construct of where a student needs to go. Four teachers correctly used the arrows for their decision. For example, teacher 10 reasoned: "*the program indicated 'number comprehension' by the arrow.*" However, seven teachers read from top to bottom and chose the core construct below the selected one. For example, teacher 11 said: "*Then I would take the core construct below, which is 'number system decimal numbers.'* That would be the logical sequence for me." Two teachers used both sets of reasoning in different parts of the interview. They noticed the arrows, but they also reasoned from top to bottom. For example, teacher 6 said: "*Within*

*the number system, I work from top to bottom during the years (...) but other learning objectives [from other core constructs] are learned simultaneously.*” Finally, six teachers deviated from the suggested routes in the LT and used their own knowledge. For example, teacher 15 said: *“I would choose to add and subtract up to 1,000. Then they also learn to work with number comprehension.”*

#### *Where are they now?*

The teachers also had different interpretations of pupils’ current level. Only two teachers correctly observed that the indented line represented conditional objectives, and one teacher correctly read the visualization as being a mind map. For example, teacher 6 said: *“...I see ‘number system up to 1,000’ as a kind of mind map with five learning objectives and other objectives underneath. So, then, you have five main objectives and some conditional objectives for a certain main objective.”* However, the majority of the teachers ( $n = 10$ ) read from top to bottom again. They indicated that all learning objectives above the selected learning objective had been mastered, while the learning objectives below still needed to be achieved. Two of these 10 teachers, as well as six other teachers, also used their own knowledge. For example, upon looking at the content of the learning objectives, teacher 7 reasoned: *“This learning objective is the same kind of exercise (...) I think that if you can split and compose numbers, you can also understand the position values of numbers.”*

#### *How best to get there?*

The selection of a subsequent learning objective toward the core construct was largely incorrectly understood. Twelve teachers chose the learning objective that was underneath the selected learning objective. They reasoned like teacher 2: *“If I look to the visualization (...) I would simply plan from top to bottom.”* Six teachers used their own knowledge, such as teacher 4: *“I think that I would choose addition and subtraction learning objectives (...) that is more from my own knowledge.”* Only teacher 1 correctly interpreted that there was no single correct sequence: *“there are all kinds of learning objectives that the students need to achieve and not necessarily the learning objective that is below the selected one (...) So actually, as I see it, there is the ‘number system up to 1,000’ with a lot of dashes to learning objectives, and the student needs to master them all before going to the next core construct.”*



### 6.6.2 Teachers' preferences

The teachers experienced the importance of LT understanding. Teacher 11 said: *"I think everything depends on LT knowledge, what you need to teach and in which grades."* They mentioned several aspects of the graphical visualization in GM that they liked or that needed to be improved, relating to the content, structure, and usability of the LT (Table 6.2).

#### Content

With regard to the content, ten teachers highlighted the importance of alignment between the LT and their curriculum. They expressed appreciation for the presentation of the grade level in GM, indicating which learning objective should be mastered in a certain period. However, teachers will be better equipped to adjust instruction if this is strongly linked to teaching materials. This can be done, for example, by indicating where the learning objectives are addressed in a teaching method. There was also the suggestion to use the same learning objective formulations in the LT and the teaching method.

Eight teachers liked the available tests, as they could help them determine a student's current level. For example, teacher 9 said: *"... the students could take the test, and I could determine whether they master the learning objectives, so that is what I liked."* Two teachers expressed a preference for additional exercises. For example, teacher 15 said: *"I want to have exercises for each learning objective so that students can learn something."*

#### Structure

Five aspects were related to the structure of the LT. First, seven teachers emphasized that the big picture should be made visible, which would outline the essential core constructs to be learned. For example, teachers 2 and 13 suggested a visualization in which all core constructs were connected in one screen.

Second, eight teachers liked the ability to drill down from the core constructs into more detailed descriptions. To illustrate, teacher 6 suggested: *"In this visualization, all learning objectives are shown behind the core constructs. Maybe, you can design a visualization that does not show all learning objectives directly, but you may have to click on the core construct before seeing them. Then you can show more information on a screen."*

Third, four teachers showed appreciation for the fine-grained information about the learning objectives. For example, teacher 12 said: *“If you clicked on the learning objective, you could also see an example. This lets you know exactly what kind of test items were asked and what the learning objective means.”* Teacher 17 noted the need to add information about students’ learning strategies because a teacher needs to know how to teach the learning objective.

Fourth, nine teachers mentioned that they did not understand the prerequisite and co-requisite relations of the learning objectives in the visualization. For example, as teacher 8 noted: *“I think this actually has to do with the primary learning objectives and prerequisite learning objectives, but that is how I fill that out myself. Anyway, it is not really clear to me.”* Teacher 16 suggested making this clearer by further drilling down, while teacher 18 suggested a distinction between the various levels through the use of arrows or different colors.

Fifth, fourteen teachers considered the importance of a clearly indicated route. They indicated that the arrows could appropriately be used to show where the student needs to go. However, students develop along numerous pathways simultaneously, such as a pathway about decimal numbers and one about monetary values. While these pathways are connected, this is not clear in the current LT visualization.

### *Usability*

Five aspects were related to the ease with which a teacher could use the LT for formative assessment and instruction. First, seven teachers liked the possibility of being able to select learning objectives. They could select learning objectives from the LT and add them to their own list in GM. This would allow them to skip or repeat learning objectives to match students’ needs.

Second, four teachers thought that the LT could be used even more intuitively. For example, teacher 13 said: *“then I think ‘how does it work’ while it should actually be very simple.”*

Third, two teachers expressed a preference for more visual organization because they found that the current LT contained too much text. Teacher 19 suggested making the layout cleaner by spreading the goals more widely.

Fourth, two teachers mentioned consistency as an important aspect. For example, teacher 14 said: *“The current LT is visualized as a tree diagram (...) so that is nice because it is a recognizable structure. You do not have to think about it because you recognize it directly from other experiences.”*

Fifth, eleven teachers mentioned the accessibility to a manual explaining the LT. They were convinced that the manual could help them understand and use the LT properly.

Table 6.2 Design principles for an LT for formative assessment

If we want teachers to use an LT for formative assessment and instruction purposes, then the LT should...
<p><b>Content</b></p> <ul style="list-style-type: none"> <li>...be aligned to the curriculum (grade division, teaching methods).</li> <li>...preferably contain tests and exercises.</li> </ul>
<p><b>Structure</b></p> <ul style="list-style-type: none"> <li>...present the big picture outlining the essential core constructs to be learned.</li> <li>...drill down from the core constructs to more detailed descriptions.</li> <li>...contain fine-grained information about the learning objectives (e.g., examples, learning strategies).</li> <li>...show the prerequisite and co-requisite relations of the learning objectives.</li> <li>...provide the (numerous) pathways that learners follow as their sophistication deepens.</li> </ul>
<p><b>Usability</b></p> <ul style="list-style-type: none"> <li>...make it possible to make a selection yourself.</li> <li>...be intuitive.</li> <li>...be visually organized and uncluttered.</li> <li>...be consistent with other (parts of the) visualization(s).</li> <li>...contain easily accessible resource materials required to understand the LT (e.g., manual).</li> </ul>

## 6.7 CONCLUSION AND DISCUSSION

This chapter reported on two studies investigating whether a graphical visualization of an LT would be appropriate for the purpose of formative assessment. A prototype of the arithmetic LT for primary education, visualized in the formative assessment platform GM, was used to collect data.

The general conclusion from both studies is that a graphical visualization of the arithmetic LT would be an appropriate representation. In Study A, multiple dependencies in the test data were found, which validated a graphical structure with connections between learning objectives. In Study B, several design principles were identified relating to the content, structure, and usability of the LT. These

design principles also support the suitability of a graphical visualization, for example, teachers prefer a clear visualization of the prerequisite and co-requisite relations between the learning objectives.

The study shows that the development of an LT requires a multidisciplinary approach involving subject experts, statisticians, and prospective users (e.g., Corcoran et al., 2009; Graham, Kennedy, & Benyon, 2000). In Study A, for example, differences were found between the structure of the explorative and confirmative network models. This can be caused by experts having incorrect assumptions or by existing student performance data that represent meaningless instructional sequences. Therefore, a consensus network model was estimated, in which the prior knowledge of experts and student performance data were combined. In Study B, the teachers experienced difficulty understanding the particular graphical visualization in GM. In particular, the connections between learning objectives and between core constructs were not clear. This shows that the validation of the graphical structure in Study A does not automatically validate its use. The intended audience should also be incorporated in the development of an LT to ensure an appropriate understanding of the LT (Lobato & Walters, 2016).

The study also demonstrated that the development of an LT is an iterative process in which a tentative LT is prepared, field-tested, revised, re-checked, and so on (e.g., Corcoran et al., 2009; Shepard, Penuel, & Pellegrino, 2018a). In Study A, several iterations were taken to estimate the consensus network model of Figure 6.4. The dependencies in this model were preliminary because we only used the input of two mathematical experts. Additional iterations are needed to further refine the arithmetic LT. In this regard, LTs and their visualization can be seen as hypotheses that are subject to substantial revision or replacement in the face of new evidence. Furthermore, the misunderstandings and design principles in Study B showed the need for additional iterations in the design of the LT visualization.

Significant differences were found between the observed test scores of students using different teaching methods. Although it was not clear whether these differences were caused by the school or the teaching method employed, it arguably matters that current instructional approaches influence the way in which students perform (Van Zanten & Van den Heuvel-Panhuizen, 2018; Wiliam, 2018). Moreover, the way in which the learning objectives were operationalized could have influenced students' performance on the learning objective, in turn influencing the network model. This limits the generalizability of the results

and calls for student and item samples that are independent of the curriculum and instructional practice to which they are exposed. Since this sampling is practically difficult or even impossible, at the very least, we need samples that are representative of the teaching methods used in order to formulate an LT that depicts the progression of the majority of students. Furthermore, we must examine whether the formulation of the learning objectives can be further fine-grained so that we can design test items for one skill.

Nevertheless, both studies illustrated how the appropriateness of an LT visualization could be investigated. Study A showed how a Bayesian network approach could be used to validate the structure of the LT. Since the approach allows for modelling multiple dependencies in the test data, it provides extensive information about the ways in which students' reasoning develops. Therefore, we could use this approach to investigate the relation between core constructs. Moreover, we could examine whether the graphical structure also applies to other subject areas, for example, science and language development. In this context, it is important to note that the items in the current study fit the target population quite well, indicating that IRT is still an appropriate way of assigning a total score to a student's ability.

In addition, Study B showed how an LT can be translated into usable tools for teachers. A prototype was used to let teachers experience a graphical visualization in a natural classroom setting. The design principles found can be used to improve the visualization in GM and create new visualizations. It would be necessary to evaluate whether these visualizations could lead to a better understanding of the LT and, ultimately, to improved instruction and student learning.

Finally, the design principles of this study are in line with those of Kobrin et al. (2015), except that their expert study only proposed a fine-grained structure and our study also identified the need for a big picture that shows how current teaching fits into the larger LT. This finding presents the possibility to investigate whether an LT visualization can be developed for multiple purposes simultaneously. This way, the LT may serve as the organizing framework in creating a coherent system of curriculum instruction and assessment (Gotwals, 2012; Shepard, Penuel, & Pellegrino, 2018b; Wilson, 2018).

## 6.8 REFERENCES

- Alonzo, A. C. (2011). Learning progressions that support formative assessment practices. *Measurement: Interdisciplinary Research & Perspective*, 9(2-3), 124–129. doi:10.1080/15366367.2011.599629
- Anderson, C. W. (2008). *Conceptual and empirical validation of learning progressions*. East Lansing, MI: Michigan State University.
- Bailey, A. L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly*, 48(3), 480–506. doi:10.1002/tesq.176
- Béguin, A. A., & Straat, J. H. (2019). On the number of items in learning goal mastery testing. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 121–134). Cham, Switzerland: Springer International Publishing.
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. doi:10.1080/0969594X.2010.513678
- Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.
- Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach*. New York: Routledge.
- Confrey, J., Gianopulos, G., McGowan, W., Shah, M., & Belcher, M. (2017). Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments designed around mathematics learning trajectories. *ZDM – Mathematics Education*, 49, 717–734. doi:10.1007/s11858-017-0869-1
- Confrey, J., Maloney, A. P., & Corley, A. K. (2014). Learning trajectories: A framework for connecting standards with curriculum. *ZDM – International Journal on Mathematics Education*, 46(5), 719–733. doi:10.1007/s11858-014-0598-7
- Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). Learning progressions in science: An evidence-based approach to reform. *CPRE Research Reports*. doi:10.1007/978-94-6091-824-7
- Daro, P., Mosher, F. A., & Corcoran, T. B. (2011). Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction. *CPRE Research Reports*. doi:10.12698/cpre.2011.r68
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Furtak, E. M. (2006). *Formative assessment in K-8 science education: A conceptual review*.

- Commissioned paper for the Committee on Science Learning, Kindergarten through Eighth Grade, National Research Council.
- Furtak, E. M., Thompson, J., Braaten, M., Windschitl, M. (2012). Learning progressions to support ambitious teaching practices. In: A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in Science* (pp. 405-433). Rotterdam, The Netherlands: SensePublishers.
- Gotwals, A. W. (2012). Learning progressions for multiple purposes. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions*. (pp. 461-472). Rotterdam, the Netherlands: Sense Publishers.
- Graham, M., Kennedy, J., & Benyon, D. (2000). Towards a methodology for developing visualizations. *International Journal of Human-Computer Studies*, 53, 789-807. doi:10.1006/ijhc.2000.0415
- Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington DC: Chief Council of State School Officers.
- Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). *Collecting validity evidence about the usability of an embedded formative assessment system*. Manuscript submitted for publication.
- Hsieh, H, & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277-1288. <http://dx.doi.org/10.1177/1049732305276687>
- Kingston, N. M., & Broadus, A. (2017). The use of learning map systems to support the formative assessment in mathematics. *Educational Sciences*, 7(1), 41, doi:10.3390/educsci7010041
- Kobrin, J. L., Larson, S., Cromwell, A., & Garza, P. (2015). A framework for evaluating learning progressions on features related to their intended uses. *Journal of Educational Research and Practice*, 5(1), 58-73. doi:10.5590/JERAP.2015.05.1.04
- Lobato, J. & Walters, C. D. (2017). A taxonomy of approaches to learning trajectories and progressions. In J. Cai (Ed.), *The compendium for research in mathematics education*, (pp. 74-101). Reston, VA: National Council of Teachers of Mathematics.
- Maris, G., Bechger, T., Koops, J., & Partchev, I. (2019). Dexter: Data Management and Analysis of Tests. R package version 0.8.5. Retrieved from <https://CRAN.R-project.org/package=dexter>
- National Research Council. (2001). *Classroom assessment and the national science education standards*. Washington, DC: National Academy Press.
- Noteboom, A. (2017). *Rekenen-wiskunde in het basisonderwijs: Domeinbeschrijving ten behoeve van peilingsonderzoek*. Enschede, the Netherlands: SLO.
- Noteboom, A., Aarsten, A., & Lit, S. (2017). *Tussendoelen rekenen-wiskunde voor het primair onderwijs*. Enschede, the Netherlands: SLO.

- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Mahwah, NJ: Lawrence Erlbaum.
- Salinas, I. (2009). *Learning progressions in science education: Two approaches for development*. Paper presented at the Learning Progressions in Science (LeaPS) conference. Iowa City, IA.
- Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action: Conclusions drawn from a special issue on formative assessment. *Educational Theory and Practice Faculty Scholarship*, 26(3), 159–162. doi:10.1080/08957347.2013.793189
- Scutari, M. (2019). Bnlearn: Bayesian network structure learning, parameter learning and inference. R package version 4.4.1. Retrieved from <https://CRAN.R-project.org/package=bnlearn>
- Scutari, M., Auconi, P., Caldarelli, G., & Franchi, L. (2017). Bayesian networks analysis of malocclusion data. *Scientific Reports*, 7, 1–11. doi:10.1038/s41598-017-15293-w
- Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, 31(2), 165–174. doi:10.1080/08957347.2017.1408628
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018a). Classroom assessment principles to support learning and avoid the harms of testing. *Educational Measurement: Issues and Practice*, 37(1), 52–57. doi:10.1111/emip.12195
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018b). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, 37(1), 21–34. doi:10.1111/EMIP.12189
- Van Zanten, M., & Van den Heuvel-Panhuizen, M. (2018). Primary school mathematics in the Netherlands: The perspective of the curriculum documents. In D. R. Thompson, M. A. Huntley, & C. Suurtamm (Eds.), *International perspectives on mathematics curriculum* (pp. 9–39). Charlotte, NC: Information Age Publishing Inc.
- Verbeeck, K., & Verschuren, M. (2010). *Het kwartje valt: Doelgericht rekenen in anders georganiseerd onderwijs*. 's-Hertogenbosch: KPC Groep.
- William, D. (2018). How can assessment support learning? A response to Wilson and Shepard, Penuel, and Pellegrino. *Educational Measurement: Issues and Practice*, 37(1), 42–44. doi:10.1111/emip.12192



- William, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Erlbaum.
- Willcox, K. E., & Huang, L. (2017). Network models for mapping educational data. *Design Science*, 3, e18. doi:10.1017/dsj.2017.18
- Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, 37(1), 5–20. doi:10.1111/emip.12188
- Zapata-Rivera, D., & Katz, I. R. (2014). Keeping your audience in mind applying audience analysis to the design of interactive score reports design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, 21(4), 442–463. doi:10.1080/0969594X.2014.936357

## APPENDIX A: ITEM EXAMPLES

Table A.1 Item examples for each learning objective

LO1: What is the next number?

Wat is het volgende getal?

920 - 910 - 900 -



LO2: How do you write this number?

negenhonderdachtenzestig



Hoe schrijf je dit getal in cijfers?

689 869 968 986

LO3: Click on 36 eggs in total

Klik in totaal 36 eieren aan.



LO4: Put the animals in order from light to heavy

Zet de dieren op volgorde van licht naar zwaar.




LO5: Nina has to pay 212 euros. Please select the correct amount

Nita moet 212 euro betalen.  
Waar ligt 212 euro?









LO6: A television costs 810 euros. What value does 8 have in this number?

Een televisie kost 810 euro.  
Hoeveel is de 8 waard in dit getal?

8

80

108

800

LO7: Hans needs 124 screws. How much does he have to buy?

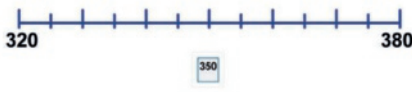


Hans heeft 124 schroeven nodig.  
Hoeveel moet hij er nog kopen?

2  
 20  
 200

LO8: Drag the card to the right place on the number line

Sleep het kaartje naar de juiste plaats op de getallenlijn.



LO9: Which child completes the numbers 339 and 453 correctly?

Welk kind rondt de getallen 339 en 453 op de juiste manier af?




LO10: Drag the balloons to the right place.


Sleep de ballonnen naar de juiste plek.



LO11: Nita buys these 300 pills. There are just as many in each jar. How many pills are in each jar?



Nita koopt deze 300 pillen. In elk potje zitten er evenveel.  
Hoeveel pillen zitten in elk potje?



## APPENDIX B: TEST VERSIONS

Table B.1 Order of the learning objectives in the eleven test versions

Test version	Order										
	Test part 1						Test part 2				
	1	2	3	4	5	6	7	8	9	10	11
1	LO1	LO3	LO4	LO5	LO2	LO6	LO7	LO11	LO8	LO10	LO9
2	LO3	LO4	LO5	LO2	LO6	LO7	LO11	LO8	LO10	LO9	LO1
3	LO4	LO5	LO2	LO6	LO7	LO11	LO8	LO10	LO9	LO1	LO3
4	LO5	LO2	LO6	LO7	LO11	LO8	LO10	LO9	LO1	LO3	LO4
5	LO2	LO6	LO7	LO11	LO8	LO10	LO9	LO1	LO3	LO4	LO5
6	LO6	LO7	LO11	LO8	LO10	LO9	LO1	LO3	LO4	LO5	LO2
7	LO7	LO11	LO8	LO10	LO9	LO1	LO3	LO4	LO5	LO2	LO6
8	LO11	LO8	LO10	LO9	LO1	LO3	LO4	LO5	LO2	LO6	LO7
9	LO8	LO10	LO9	LO1	LO3	LO4	LO5	LO2	LO6	LO7	LO11
10	LO10	LO9	LO1	LO3	LO4	LO5	LO2	LO6	LO7	LO11	LO8
11	LO9	LO1	LO3	LO4	LO5	LO2	LO6	LO7	LO11	LO8	LO10

## APPENDIX C: INTERVIEW

Table C.1 Interview schedule

---

Purpose and use
1. Which grade level do you teach?
2. Which teaching method do you use?
3. Have you seen the LT in GM? If yes, what did you use the LT for?

---

Understanding
Where do the students need to go?
4. Suppose a student has mastered all the learning objectives of the core construct “number system up to 1,000.” Which core construct would you plan next? How did you make this choice?
Where are they now?
5. Suppose a student is working on the core construct “number system up to 1000” and has mastered the learning objective “the student is able to split and group numbers up to 1,000 with hundreds, tens, and metric units.” Which learning objectives does the student also master? Why do you think that?
How best to get there?
6. Suppose a student has mastered the learning objective “the student is able to split and group numbers up to 1,000 with hundreds, tens, and metric units.” Which learning objective would you plan next? How did you make this choice?

---

Preferences
7. To what extent does the current visualization support your understanding of the LT?
8. Which aspects cause this understanding?
9. Which visualization aspects are unclear or uncomfortable for you to establish where the student is/where they need to go/how best to get there?
10. What else do you need in order to use the LT for establishing where the student is/where they need to go/how best to get there?

---

## APPENDIX D: CODING SCHEME

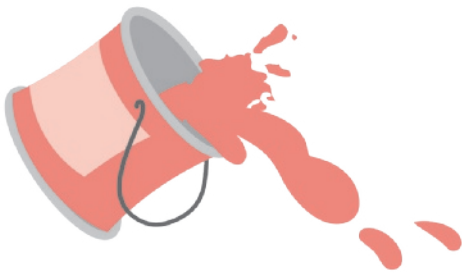
Table D.1 Coding scheme

Codes	Description	Example
Categories of understanding		
Where do the students need to go?		
Arrow	The teacher selected the core construct “number comprehension” due to the arrow	“The program indicated ‘number comprehension’ by the arrow”
Top to bottom	The teacher selected the core construct “number system decimal numbers,” which is the core construct below; reads from top to bottom	“Then I would take the core construct below (...). That would be the logical sequence for me.”
Own understanding	The teacher uses their own knowledge or teaching method instead of the visualization	“I know that grade 4 had to master calculations up to 10,000, and grade 5 had to master calculations up to a million. So that would be a logical next step for students to work on.”
Where are they now?		
Mind map	The teacher interprets the visualization as a mind map	“Of course, it is a kind of a mind map because the learning objectives below still has to do with it...”
Indented line	The teacher noticed the indented line as representing conditional objectives	“Take a look, I think this one because this learning objective is slightly more indented.”
Top to bottom	The teacher reads from top to bottom; mentions the learning objective above	“The student is also able to perform (...) because that is presented above this learning objective.”
Own understanding	The teacher uses their own knowledge or teaching method instead of the visualization	“I would at least evaluate whether a student mastered ‘up to 100.’”
How best to get there?		
Mind map	The teacher interprets the visualization as a mind map	“So actually, as I see it, there is a core construct ‘number system up to 1,000’ with a lot of arrows with sub-goals, and you have to master them all before you go to the next core construct...”
Top to bottom	The teacher reads from top to bottom; mentions the learning objective below	“Then I would take the learning objective (...). It is below, so that would be logical for me.”

Own understanding	The teacher uses their own knowledge or teaching method instead of the visualization	“The same learning objective, however, then extended to 10,000.”
Categories of preference		
Content		
Alignment to curriculum	Clear link between the LT and the grade division and teaching method. This also includes: formulation of learning objectives and inclusion of all learning objectives of the curriculum	“A lot of teaching methods used five learning domains, and this LT used four. (...) I can imagine that this can be very confusing.”
Tests and exercises	A test would help determine a student’s level. Exercise material would help students to work on it	“The big advantage of this system is that (...) it is seven test items and then you know whether a student has mastered the learning objective.”
Structure		
Big picture	A clear overview of all core constructs at a glance. This also includes: clear division into categories and different colors per core construct	“The advantage of that LT was that it was clear at a glance what the structure of the LT was. I could see that in one overview, in one screen.”
Drill down	The possibility of drilling down from the core construct to more detailed descriptions by clicking on it	“I liked that if you click on the core constructs on the left, you can open and close it. And then you see the corresponding learning objectives.”
Fine-grained	Providing fine-grained information, such as item examples and learning strategies	“We put in that LT the learning strategies, the strategies you need to perform well.”
Prerequisite and co-requisite relations	Clear visualization of relations between learning objectives	“I think this actually has to do with primary learning objectives and prerequisite learning objectives, but that is how I fill that out myself. Anyway, it is not really clear for me.”
Pathways	The sequence of learning objectives is clear, for example, by arrows. Even if there are multiple sequences or connections between core constructs	“But then I just wonder, you have this learning objective, but in addition, you have the objectives about addition and subtraction, and these are learned simultaneously. It is not that you first mastered this objective and then go to the next.”
Usability		
Selection	The possibility to make a selection of the learning objectives by yourself	“I just think it is very nice that you can select each learning objective that you would like to work on.”

Intuitively	The visualization works quickly, smoothly, and intuitively. Everything works well	"Then I think 'how does it work,' while it should actually be very simple."
Visually organized	Uncluttered and not too much text	"Yes, I thought it was very linguistic, (...) I think it is nice to see visual elements."
Consistent	The same symbols are used for the same elements in this visualization/other contexts	"The current LT is visualized as a tree diagram (...) so that is nice because it is a recognizable structure. You do not have to think about it because you recognize it directly from other experiences."
Resource materials	Accessibility of resource materials (e.g., manual).	"I think if there is more explanation about how to use it, it will be clearer."
Other		
Importance of LT understanding	Describing why it is important to have knowledge of the LT	"I think everything depends on LT knowledge, what you need to teach and in which grades."
Description LT	The teacher describes the LT on the screen	"Then I get a diagram with the core construct 'number system up to 1,000' in the middle, a lot of blue lines that point to several learning objectives, a dotted line that points to 'number understanding,' and the word 'number system' in blue."
Misconceptions	General misconceptions, such as the understanding of the arrows	"Yes, the arrow points to 'number understanding', which is the domain where this learning objective comes from."
General aspects of GM	The teacher mentions aspects, other than the LT, i.e., login codes, digital problems, the kind of test items	"I liked the system. We need a system that can follow students regardless of other tests."





# CHAPTER 7



## CONCLUSION AND DISCUSSION

This dissertation demonstrates the necessity and fruitfulness of designing and evaluating instruments with the intended audience to ensure that they understand and use the assessment results in an appropriate way. Moreover, it shows that the design and evaluation of formative assessment instruments is a significantly more complicated enterprise. It can be compared with a balancing act in which choices are constantly weighted.

The current chapter starts by describing the characteristics of a formative assessment instrument in supporting teachers' understanding and use, as investigated in chapters 2 to 6 (paragraph 7.1). The next section addresses several factors that should be balanced during the design and development of the formative assessment (paragraph 7.2). The chapter ends with some directions for future research (paragraph 7.3).

### 7.1 CHARACTERISTICS OF A FORMATIVE ASSESSMENT INSTRUMENT

Formative assessment instruments are intended to provide users with the information they need, in a way that they can understand, so that they can perform appropriate actions that support student learning. Actions by teachers include providing feedback, determining next steps in instruction, and selecting learning materials.

In this dissertation, characteristics regarding the content of formative assessment instruments were investigated. It was concluded that instruments should provide different kinds of information that enable teachers to perform appropriate instructional actions. First, the instrument needs to contain

fine-grained information about students' current performance, including their mastery of a learning objective and their learning strategies. Second, the instrument must contain information about where students need to go, including the learning objectives for each grade and the sequence of learning objectives in the learning trajectory. Third, the instrument should provide direction for follow-up actions, including a link to learning materials and suggestions for grouping students. To make a decision about instructional actions, teachers also need information about students' personal characteristics, such as their working attitude and motivation. This information shows the link between assessment results and other information about the student and combines different sources of information to make a decision.

In addition, characteristics regarding the visual presentation of the instrument were studied. Since measurement error is a difficult concept to grasp, different visual presentations of measurement error were compared. It was concluded that the visual presentation affects the teacher's decisions and preferences; however, several misconceptions were identified. Furthermore, the visualization of a learning trajectory was investigated, leading to the conclusion that a graphical visualization might be the most appropriate way to present possible learning pathways. Several design principles for visualization were identified in relation to the content, structure, and usability of the learning trajectory.

## 7.2 BALANCING ELEMENTS

### *7.2.1 Balancing Preferences and Comprehension*

When designing assessment instruments, it is important to consider both preference and comprehension, since the most preferred visualization is not necessarily the best understood (Zapata-Rivera, Kannan, & Zwick, 2019). Wainer, Hambleton, and Meara (1999) argued that comprehension outweighs preferences, as design choices are not a beauty contest. Consequently, this dissertation does not rely exclusively on user opinions; it also investigates users' actual understanding and use. For example, think-aloud protocols and log file

analysis provided interesting information about teachers' cognitive thinking processes and actual activities. Moreover, quantitative and qualitative research approaches were used to collect information from multiple points of view.

### *7.2.2 Balancing Intended Interpretation and Use*

Validation involves a critical evaluation of the intended interpretations and uses of assessment results. Kane and Wools (2019) considered two perspectives for validation: functional and measurement perspectives. The functional perspective essentially focuses on how well the assessment serves its intended purposes (i.e., its use and consequences), while the measurement perspective considers the technical quality of the assessment instrument (i.e., its content coverage and freedom from measurement error). Whichever one of these perspectives is emphasized depends on the specific goals and inferences. Arguably, however, both perspectives weigh equally in the validation of embedded formative assessment, as both are critical to the effectiveness of formative assessment.

### *7.2.3 Balancing Tradition and Innovation*

Formative assessment tasks must provide teachers with feedback about students' thinking, their learning strategies, and misconceptions. Traditionally, assessment tasks have offered feedback about students' correct answers on the basis of quantitative scores. However, these scores are of little informational value for teachers in determining the next instructional steps. Teachers typically use these scores to identify and reteach students who have obtained the greatest number of wrong items, but they are unable to adjust their instruction based on students' thinking. Increases in technological advancement have opened up possibilities for developing new forms of assessment that enable more qualitative insights (Shepard, Penuel, & Pellegrino, 2018b; Wilson, 2018). For example, simulations, multi-media-enhanced items, and hybrid tasks offer more authentic and elaborate opportunities to evaluate students' thinking. At the same time, difficulties may arise from the scoring and generalization of these tasks (Wools, Molenaar & Hopster, 2019). Consequently, we need tasks that combine the strengths of both innovative and traditional tasks.

#### *7.2.4 Balancing Instrument and Process*

Although there is a need for a well-designed instrument to support teachers' understanding and use, this dissertation shows that such an instrument is not sufficient. We also need skilled teachers who are able to accurately understand and use the instrument for student learning. For example, teachers need a certain level of assessment literacy (e.g., Mandinach & Gummer, 2016), as they have several misconceptions about the concept of measurement error. Furthermore, teachers' knowledge base is a crucial factor in transforming assessment information into instructional actions (Davenport & Prusak, 1998; Ebbeler, Poortman, Schildkamp, & Pieters, 2016; Marsh, 2012), since the instrument can never explicitly present teachers with precise instructional actions to perform. Teachers take into account students' individual circumstances and particular context (e.g., Harlen & James, 1997), and they combine the assessment information with disciplinary, curricular, and pedagogical content knowledge as well as an understanding of how children learn (Falk, 2012; Gummer & Mandinach, 2015; Heritage, Kim, Vendlinski, & Herman, 2009; Sabel, Forbes, & Zangori, 2015). This shows that a thoughtful integration of both the instrument and process is needed in order to successfully implement formative assessment.

#### *7.2.5 Balancing Formative and Summative Assessment*

Formative assessment exists within the larger educational context, which can hinder or support its effectiveness (Bennett, 2011). One influential component within this context is the use of summative assessment. Summative assessment can pose a serious threat to the learning objectives of formative assessment if it measures an impoverished representation of these learning objectives (Shepard, 2006). As a result, teachers and students may focus their attention and effort only on the graded portion of the curriculum, resulting in a reduction in the potential of formative assessment to engender deeper learning. Thus, a balanced assessment system is needed wherein formative and summative assessment are mutually supportive and coherently linked to the same learning trajectory (Marion, 2018).

### 7.3 FUTURE DIRECTIONS

The findings and balancing elements in this dissertation provide several directions for future research. First, the dissertation focused on the understanding and use of formative assessment results by teachers. Since students also constitute an essential part of formative assessment, we should also investigate what characteristics of formative assessment instruments support a student's understanding and use. The results of both strands of research should be combined into one instrument to achieve the maximum benefit of embedded formative assessment.

Second, the studies in this dissertation were performed in the context of mathematics. Formative assessment can be also applied to other domains of knowledge and skills. There has been increasing attention around 21st century skills, such as critical thinking, creativity, collaboration, and problem-solving. It would be interesting to investigate whether different characteristics of formative assessment instruments are required for these types of skills. For this investigation, methodological approaches and data collection materials similar to those undertaken in this dissertation could be used.

Third, while this dissertation highlights the validity of formative assessment results from a functional perspective, a critical evaluation from a measurement perspective is also needed to determine the validity of interpreting and using the assessment results in the proposed way. In this regard, I recommend further research into the development of items or tasks that provide more insight into students' learning process while preserving the power of the inferences drawn from traditionally used items. In addition, more research is needed to ascertain how psychometric criteria can be translated into the context of formative assessment.

Fourth, as mentioned earlier, formative assessment instruments are part of a larger educational context. To achieve success in implementing formative assessment, it is important to continue to invest in the professional development of teachers, as they are the main drivers in the teaching and learning process. Moreover, we need a coherent assessment system wherein formative and summative assessment are mutually supportive. The development and implementation of such a system represent an ambitious goal that needs further research, starting with the development and empirical validation of learning trajectories.

Finally, the current dissertation shows that test developers and researchers are both crucial in the implementation of formative assessment. They have a duty to provide teachers with understandable and useful instruments that will allow them to make valid inferences based on assessment results. By incorporating user characteristics and preferences into the instrument design, the implementation of formative assessment will be facilitated. This way, test developers and researchers will help teachers achieve their job description—i.e., supporting student learning by “easily, accurately, and appropriately identifying their learning needs and responding appropriately” (Hattie & Brown, 2008, pp. 199-200). It is by continuing to assume these responsibilities in future research that we can ensure balance in the design and implementation of formative assessment.

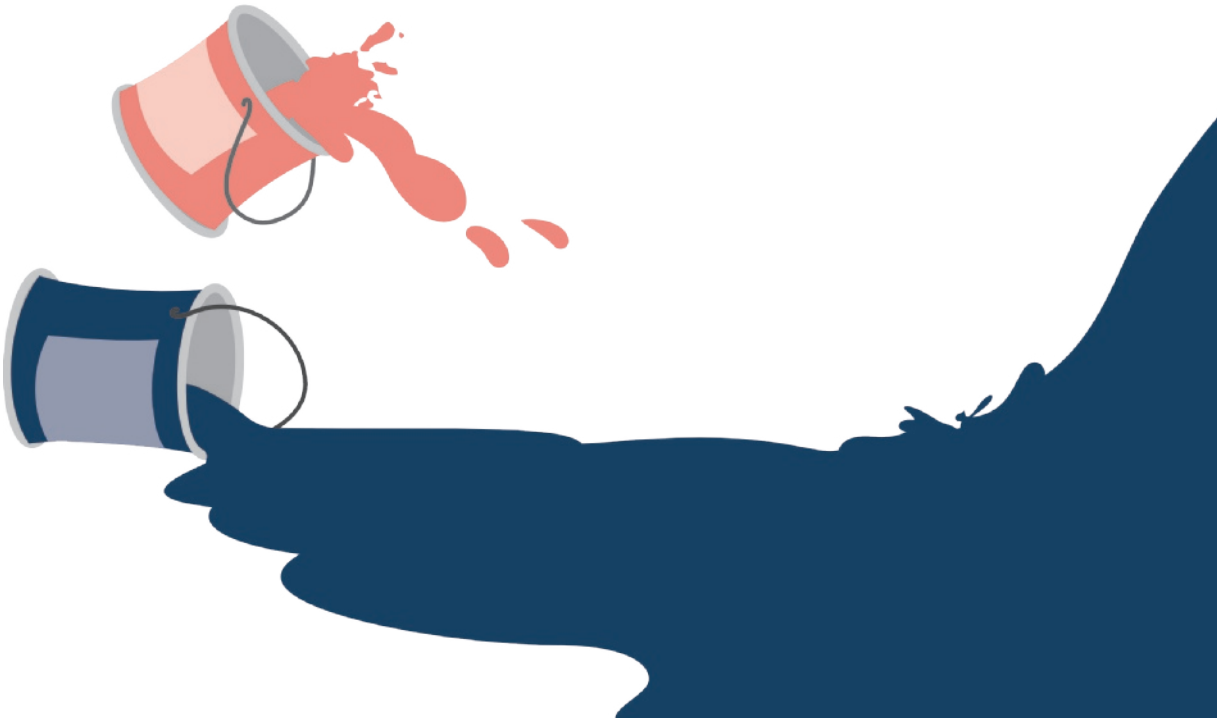
## 7.4 REFERENCES

- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. doi:10.1080/0969594X.2010.513678
- Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston: Harvard Business School Press.
- Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (2016). Effects of a data use intervention on educators’ use of knowledge and skills. *Studies in Educational Evaluation*, 48, 19–31. doi:10.1016/j.stueduc.2015.11.002
- Falk, A. (2012). Teachers learning from professional development in elementary science: Reciprocal relations between formative assessment and pedagogical content knowledge. *Science Education*, 96(2), 82–99. doi:10.1002/sce.20473
- Gummer, E., & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record*, 117(4), 1–22. Retrieved from <https://www.tcrecord.org>
- Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3), 365–379. doi:10.1080/0969594970040304
- Hattie, J. A. C., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, 36(2), 189–201. doi:10.2190/ET.36.2.g
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. L. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31. doi:10.1111/j.1745-3992.2009.00151.x



- Kane, M. T., & Woolls, S. (2019). Perspectives on the validity of classroom assessments. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom Assessment and Educational Measurement*, (pp. 11-26). New York, NY: Routledge
- 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. doi:10.1016/j.tate.2016.07.011
- Marion, S. F. (2018). The opportunities and challenges of a systems approach to assessment. *Educational Measurement: Issues and Practice*, 37(1), 45–48. doi:10.1111/emip.12193
- Marsh, J. A. (2012). *Interventions promoting educators' use of data: Research insights and gaps*, 114(11), 1–48. Retrieved from <http://www.rdc.udel.edu>
- Sabel, J. L., Forbes, C. T., & Zangori, L. (2015). Promoting prospective elementary teachers' learning to use formative assessment for life science instruction. *Journal of Science Teacher Education*, 26(4), 419–445. doi:10.1007/s10972-015-9431-6
- Shepard, L. A. (2006). Classroom Assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Washington DC: American Council on Education.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018b). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, 37(1), 21–34. doi:10.1111/emip.12189
- Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36, 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x
- Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, 37(1), 5–20. doi:10.1111/emip.12188
- Woolls, S., Molenaar, M., Hopster-den Otter, D. (2019). The validity of technology enhanced assessments - threats and opportunities. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Cham, Switzerland: Springer International Publishing.
- Zapata-Rivera, D., Kannan, P., & Zwick, R. (2019). Communicating measurement error information to teachers and parents. In D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. 63–73). New York and London: Routledge.







## SUMMARY

Formative assessment has been defined as both an instrument and a process that is intended to support student learning. The instrument provides information about a student's learning. During the process, the information is judged and used by teachers, students, or their peers for actions that support that learning. The existing research points to the potential of formative assessment, however, several studies have shown that teachers have difficulty understanding and using the evidence from assessment instruments. This dissertation investigates what content and visual presentation of a formative assessment instrument could help teachers. The central question is: What characteristics of a formative assessment instrument support teachers' understanding and use?

Chapter 2 introduced the concept of formative assessment and provided a general framework for the validation of formative assessment. Validity is one of the most important quality criteria for the evaluation of assessments and is often defined as the extent to which an assessment result is appropriate for its intended interpretation and use. The argument-based approach to validation is widely adopted for validating the appropriateness of assessment results. This approach consists of two stages. In the first stage, an interpretation and use argument (IUA) is developed by specifying the proposed interpretation and use of the assessment results. In the second stage, the IUA is evaluated, and a validity argument concludes whether it is valid to interpret and use the assessment results. The current study applied the argument-based approach to formative assessment. The focus is on embedded formative assessment, the most formal category, which consists of predefined tasks. This resulted in an IUA consisting of inferences regarding both a score interpretation and a score use.

With regard to score-interpretation inferences, we proposed a structure that is identical to the existing validation framework for summative assessment. This contained 1) a scoring inference, whereby students' performance is converted into interpretable information about their thinking; 2) a generalization inference, in which we draw upon the scoring of a limited sample of items to make inferences about the generalization of this score to all possible items in a so-called test domain; and 3) an extrapolation inference, in which the interpretation of all possible items is extrapolated to a more general claim about students' performance in a so-called practice domain. The practice domain is defined as the domain about which we would like to make a decision.

With regard to the score-use inferences, we proposed an extension to the validation framework for summative assessment. The existing 4) decision inference links students' performance regarding the construct in the practice domain to a decision about this performance. In addition, we proposed 5) a judgment inference because inaccurate understanding of the decision could lead to inappropriate action; 6) an action inference, since teachers and students are assumed to use the judgment for the selection of appropriate actions; and 7) a consequence inference because the implementation of these actions is expected to support student learning.

Furthermore, it was argued that the validity argument of the argument-based approach should focus on evaluating inferences regarding score interpretation as well as score use, since both are critical to the effectiveness of formative assessment. The proposed framework was illustrated by an operational example, including a presentation of sources of evidence that can be collected on the basis of the validation framework.

Chapter 3 focused on the characteristics regarding the content of formative assessment instruments. The study performed a needs assessment - an investigation of the type of instructional actions as well as the information needs for enabling these actions. This needs assessment provides the possibility to fit the content of formative assessment instruments to the actions and information needs of the intended users. The research questions were as follows:

1. Which types of actions would users choose as desired uses of test results, and how do these actions relate to actual uses?

2. What, if any, is the extent of the differences between teachers, internal coaches, principals, and parents with regard to desired and actual uses and corresponding actions?
3. What information from test results is needed to perform the desired actions?

The questionnaire data showed that teachers ( $n = 140$ ), internal coaches ( $n = 34$ ), school principals ( $n = 14$ ), and parents ( $n = 250$ ) want to use test results for actions that support learning, which amounts to a discrepancy relating to actual use. For example, they would like to create group action plans, individual action plans, or help students with homework.

Furthermore, the various users would perform actions on different levels and in different contexts. The teachers and parents reported that they would like to perform actions at the level of the individual student, with teachers acting in an educational setting and parents serving in a more informal capacity. The internal coaches and principals selected more actions relating to the school level. This result indicated the need for tailored reports that fit the information needs of individual users.

Based on these results, the decision was made to limit the third question to teachers. The results from the seven focus groups ( $n = 84$ ) showed the need for different kinds of information, for instance, relating to students' mastery of a learning objective, their strategy to solve an assignment, their working attitude, and motivation. This demonstrates that the teachers linked the test results to other information about the student and combined different sources of information to take actions.

Chapter 4 focused on the visual presentation of formative assessment instruments. Specifically, the study investigated the extent to which presentations of measurement error in score reports influenced teachers' judgments, since all assessment results are subject to a certain amount of measurement error. The influence of measurement error presentations was operationalized as teachers' need to gather additional information to enable decision-making regarding students. Two research questions were formulated:

1. To what extent do various measurement error presentation formats result in teachers' need for additional information compared to a presentation format that omits measurement error?
2. Which of the various presentation formats do teachers prefer?

Three presentation formats of measurement error (blur, color value, and error bar) were compared to a presentation format that omitted measurement error.

The results from a factorial survey analysis showed that the position of a score in relation to a cut-off score impacted most significantly on decisions. Moreover, the teachers ( $n = 337$ ) indicated the need for additional information significantly more often when the score reports included an error bar compared to when they omitted measurement error. The error bar was also the most preferred presentation format. The results were supported in think-aloud protocols and focus groups, although several interpretation problems and misconceptions of measurement error were identified.

Chapter 5 continued with an investigation of the characteristics of formative assessment instruments by collecting validity evidence to support assumptions regarding the intended use of formative assessment. The main question was: To what extent are assessment results usable for teachers' formative assessment practices? A prototype of an embedded formative assessment instrument, named Groeimeter (GM), was used to collect evidence. GM was used by 29 teachers in a natural classroom setting for three months, during which time, data were collected from log files, questionnaires, and interviews.

The results showed that the prototype was largely usable in terms of establishing where the students were in their learning and where they needed to go and that it was somewhat usable in relation to how best to get there. There were also suggestions for the further development of GM, including a number of generic suggestions and specific ones for improving the user-friendliness of GM. In addition, some suggestions were deemed necessary for the successful implementation of formative assessment, which were design principles formulated for the development of formative assessment instruments:

1. If we want teachers to establish where students need to go in their learning, then a visualization of the learning trajectory is needed, which clearly shows the relationship between learning objectives.
2. If we want teachers to establish where the students are in their learning, then in-depth feedback is needed in terms of students' learning strategies and their misconceptions regarding learning objectives.



3. If we want teachers to establish how best to enable students to get to their desired state of learning, then (a link to) instructional materials should be provided.

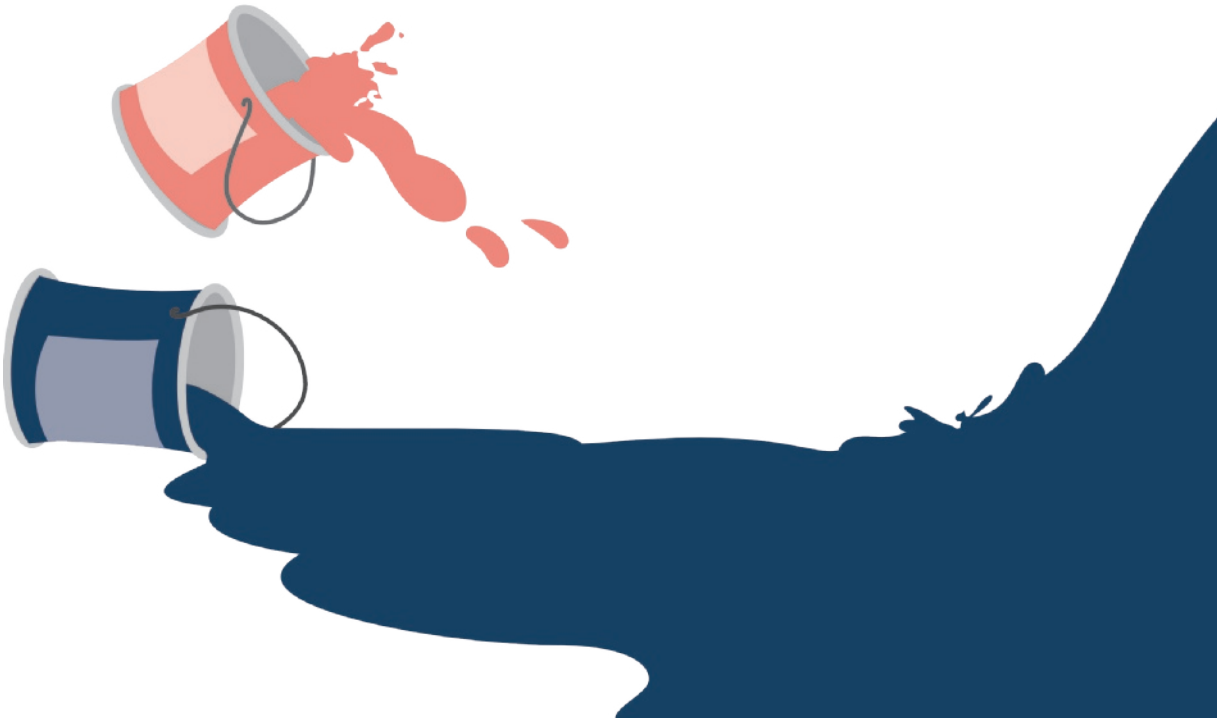
Chapter 6 focused on the visual presentation of a learning trajectory, which was one of the design principles in chapter 5. Specifically, the chapter reported on two studies investigating the appropriateness of a graphical visualization, since teachers showed difficulty understanding and using tabular learning trajectories. A prototype of a graphical visualization within GM was used to collect data.

The aim of Study A was to validate the graphical structure of the prototype by modelling student performance data using a Bayesian network analysis. An assessment was administered to 787 students, which measured student performance on 11 learning objectives. The results showed multiple conditional dependencies in the data, confirming a graphical structure.

Study B examined teachers' understanding and preferences regarding the graphical visualization. In total, 19 teachers used the prototype in a natural classroom setting for two months. The results from the interview data showed that the teachers had difficulty appropriately understanding the visualization. Several design principles were identified in relation to the content, structure, and usability of the learning trajectory. These design principles also showed that the teachers had a preference for a graphical visualization.

Overall, this dissertation shows that it is necessary and fruitful to design and evaluate assessment instruments with the intended audience in order to ensure that they will understand and use the assessment results in an appropriate way. With regard to the content of formative assessment instruments, we concluded that the instrument needs (1) fine-grained information about students' current performance (e.g., learning strategies, misconceptions); (2) information about students' desired performance (e.g., learning objectives, learning trajectory); and (3) directions to follow-up actions (e.g., learning materials, suggestions for grouping). With regard to the visual presentation of formative assessment instruments, we concluded that measurement error presentations affect teachers' decisions and that a graphical visualization is appropriate for presenting a learning trajectory.

Although the characteristics for formative assessment instruments make them appear easy to develop, their design and evaluation are highly complex. It can be compared with a balancing act in which choices are weighted constantly. For example, we should strike for a balance between teachers' preferences and comprehension. Furthermore, we must ensure an equal weight of the score-interpretation and score-use inferences during the validation of formative assessment. In addition, the need for fine-grained information shows that we need to develop test items that afford more qualitative insight while preserving the power of inferences drawn from traditionally used items. The interpretation problems and information needs also point to the need for a thoughtful integration of both the instrument and process of formative assessment. Finally, we must look for a balanced assessment system wherein formative and summative assessment are mutually supportive. The characteristics and balancing elements provide several directions for future research. Above all, these elements call on us to continue to take responsibility as test developers and researchers in developing understandable and useful formative assessment instruments so that the design and implementation of formative assessment can be balanced for all involved.





## SAMENVATTING

Formatief toetsen is zowel een instrument als een proces dat als doel heeft het leren te ondersteunen. Het instrument geeft informatie over het leren van leerlingen. Bij het proces wordt deze informatie beoordeeld en gebruikt door onder andere leerkrachten en leerlingen voor vervolgcities ter ondersteuning van dat leren. Formatief toetsen lijkt een veelbelovende ontwikkeling in het onderwijs, maar uit onderzoek blijkt dat leerkrachten moeite hebben met het begrijpen en gebruiken van toetsresultaten. Dit proefschrift gaat daarom na hoe inhoud en visualisatie van een formatief toetsinstrument leerkrachten kunnen helpen. De centrale onderzoeksvraag is: welke kenmerken van een formatief toetsinstrument ondersteunen het begrip van en het gebruik voor leerkrachten?

**Hoofdstuk 2** introduceert het concept van formatief toetsen en presenteert een algemeen kader voor de validering van formatief toetsen. Validiteit is één van de belangrijkste kwaliteitscriteria voor de evaluatie van toetsen en wordt vaak gedefinieerd als de mate waarin een toetsresultaat geschikt is voor de beoogde interpretatie en het beoogde gebruik. De argumentgerichte benadering van validiteit is een veelgebruikte manier om de geschiktheid van toetsresultaten te valideren. Deze benadering bestaat uit twee fasen. In de eerste fase wordt een interpretatie- en gebruiksargument (IUA) ontwikkeld, dat alle inferenties omtrent de voorgestelde interpretatie en het voorgestelde gebruik van toetsresultaten specificiert. In de tweede fase wordt de IUA geëvalueerd en geeft het validiteitsargument een conclusie over de validiteit van de toetsresultaten voor de beoogde interpretatie en het beoogde gebruik. De studie in hoofdstuk 2 past de argumentgerichte benadering toe op formatief toetsen. De focus ligt daarbij op het ingebedde formatieve toetsen, dat is de meest formele vorm en bestaat uit

vooraf ontwikkelde taken. Het resultaat is een interpretatie- en gebruiksargument (IUA) met score-interpretatie- en scoregebruikinferenties.

Wat betreft de score-interpretatie-inferenties wordt een structuur voorgesteld die gelijk is aan het bestaande valideringskader voor summatief toetsen. Dit bevat 1) een score-inferentie, waarbij de prestaties van leerlingen worden omgezet naar interpreteerbare informatie over hun denken; 2) een generalisatie-inferentie, waarbij de scores van een beperkte steekproef van items gegeneraliseerd worden naar een score op alle mogelijke items in een zogeheten toetsdomein en 3) een extrapolatie-inferentie, waarbij de interpretatie van alle mogelijke items wordt geëxtrapoleerd naar een meer algemene bewering over de prestaties van leerlingen in een zogeheten praktijkdomein. Het praktijkdomein is gedefinieerd als het domein waarover we graag een beslissing willen nemen.

Wat betreft de scoregebruikinferenties wordt een uitbreiding voorgesteld op het validatiekader voor summatief toetsen. De bestaande 4) beslissingsinferentie koppelt de prestaties van leerlingen met betrekking tot het construct in het praktijkdomein aan een beslissing over deze prestaties. Daarnaast worden de volgende inferentie voorgesteld: 5) een beoordelingsinferentie, omdat onjuist begrip van de beslissing tot verkeerde acties zou kunnen leiden; 6) een actie-inferentie, aangezien leraren en leerlingen verondersteld worden de beoordeling te gebruiken voor de selectie van geschikte acties en 7) een consequentie-inferentie, omdat verwacht wordt dat de implementatie van deze acties het leren van leerlingen zal ondersteunen.

Verder moet het validiteitsargument uit de argumentgerichte benadering zich concentreren op de inferenties van zowel de score-interpretatie als het scoregebruik. Beide zijn namelijk van cruciaal belang voor de effectiviteit van formatief toetsen. Het voorgestelde kader is geïllustreerd aan de hand van een concreet voorbeeld met ideeën voor bewijsmateriaal, dat op basis van het validiteitskader kan worden verzameld.

**Hoofdstuk 3** richt zich op de inhoudelijke kenmerken van formatieve toetsinstrumenten. Het onderzoek bestaat uit een behoefteanalyse waarin wordt nagegaan welke educatieve acties er zijn en welke informatie nodig is om deze acties uit te voeren. De behoefteanalyse geeft de mogelijkheid om de inhoud van formatieve toetsinstrumenten af te stemmen op deze acties en de informatiebehoeften van beoogde gebruikers. De onderzoeksvragen zijn als volgt geformuleerd:

1. Voor welke type acties willen leerkrachten, interne begeleiders, directeuren en ouders de toetsresultaten gebruiken en hoe verhouden deze acties zich tot het huidige gebruik?
2. Zijn er verschillen tussen de leerkrachten, interne begeleiders, directeuren en ouders met betrekking tot het gewenste en huidige gebruik en bijbehorende acties?
3. Welke informatie van toetsresultaten is nodig om de gewenste acties uit te voeren?

De resultaten afkomstig van de vragenlijsten laten zien dat leerkrachten ( $n = 14$ ), interne begeleiders ( $n = 34$ ), directeuren ( $n = 13$ ) en ouders ( $n = 250$ ) toetsresultaten willen gebruiken voor acties die het leren ondersteunen. Ze willen bijvoorbeeld graag handelingsplannen voor de groep of een individu maken, of ze willen leerlingen ondersteunen bij het huiswerk. Deze acties verschillen van het huidige gebruik, dat meer gekenmerkt wordt door acties die het leren evalueren.

Verder willen de verschillende gebruikers acties uitvoeren op verschillende niveaus en in verschillende contexten. Leerkrachten en ouders willen acties doen op het niveau van de individuele leerling, waarbij leerkrachten werken in een onderwijsomgeving en ouders handelen in een meer informele leeromgeving. Interne begeleiders en directeuren kiezen meer acties die aan het schoolniveau gerelateerd zijn. Dit resultaat laat zien dat er rapporten op maat moeten worden gemaakt die voldoen aan de informatiebehoeften van een individuele doelgroep.

Vanwege de verschillen tussen gebruikers is er besloten om de derde vraag te beperken tot de doelgroep van leerkrachten. Resultaten van zeven focusgroepen ( $n = 84$ ) tonen dat er behoefte is aan verschillende soorten informatie, bijvoorbeeld over het beheersen van een leerdoel, de strategieën om vragen op te lossen, de werkhouding en de motivatie van leerlingen. Leerkrachten combineren informatie uit toetsresultaten met andere informatie over leerlingen om tot een actie te komen.

**Hoofdstuk 4** richt zich op de visuele presentatie van formatieve toetsinstrumenten. Het onderzoek kijkt in het bijzonder naar de mate waarop meetfoutvisualisaties in scorerapportages invloed hebben op de beslissingen van leerkrachten, aangezien alle toetsresultaten onderhevig zijn aan meetfout. De invloed van meetfoutvisualisaties is gedefinieerd als de behoefte aanvullende informatie te verzamelen om beslissingen over leerlingen te nemen.

Twee onderzoeksvragen zijn geformuleerd:

1. In hoeverre hebben verschillende meetfoutvisualisaties invloed op de behoefte van leerkrachten om aanvullende informatie te verzamelen ten opzichte van een visualisatie zonder meetfout?
2. Aan welke visualisatie geven leerkrachten de voorkeur?

Drie meetfoutvisualisaties (vervaging, kleurintensiteit en foutenbalk) worden vergeleken met een visualisatie zonder meetfout.

De resultaten van een vignettenstudie laten zien dat de positie van een score in relatie tot de grensscore het meeste invloed heeft op de beslissingen van leerkrachten. Bovendien geven de leerkrachten ( $n = 337$ ) significant vaker aan dat ze meer informatie nodig hebben bij scorerapportages met een foutenbalkvisualisatie. Leerkrachten geven tevens de meeste voorkeur aan de foutenbalkvisualisatie. De resultaten worden bevestigd door hardopdenkprotocollen en focusgroepen. Hieruit blijkt echter ook dat leerkrachten verschillende interpretatieproblemen en misconcepties rondom meetfout hebben.

Hoofdstuk 5 gaat verder met het onderzoek naar de kenmerken van formatieve toetsinstrumenten door validiteitsbewijs te verzamelen over de inferenties met betrekking tot het beoogde gebruik van formatief toetsen. De hoofdvraag is: in welke mate zijn toetsresultaten voor leerkrachten bruikbaar ten behoeve van het formatief toetsen? Een prototype van een ingebed formatief toetsinstrument, genaamd Groeimeter, is gebruikt om bewijs te verzamelen. Groeimeter is door 29 leraren gedurende drie maanden in een natuurlijke klasomgeving gebruikt. In deze periode zijn er gegevens verzameld uit logbestanden, vragenlijsten en interviews.

De resultaten tonen aan dat Groeimeter vooral bruikbaar is om vast te stellen waar de leerlingen zijn in hun leren en waar ze naartoe moeten gaan. Daarnaast blijkt Groeimeter gedeeltelijk bruikbaar voor hoe leerlingen daar het best naartoe kunnen worden begeleid. Er zijn ook ideeën voor de verdere ontwikkeling van Groeimeter, waaronder een aantal algemene ideeën en ideeën om de gebruiksvriendelijkheid van Groeimeter te verbeteren. Daarnaast worden enkele ideeën noodzakelijk geacht voor de succesvolle implementatie van formatief toetsen. Deze suggesties zijn geformuleerd als ontwerpprincipes voor de ontwikkeling van formatieve toetsinstrumenten:



1. Als we leerkrachten willen laten vaststellen waar leerlingen naartoe moeten in hun leren, dan is er een visualisatie van een leerlijn nodig, waar de relaties tussen leerdoelen duidelijk wordt weergegeven.
2. Als we leerkrachten willen laten vaststellen waar leerlingen zijn in hun leren, dan is er diepgaande feedback nodig in termen van leerstrategieën en misconcepties van leerlingen.
3. Als we leerkrachten willen laten vaststellen hoe ze leerlingen het beste naar de gewenste doelen kunnen begeleiden, dan moet er (een link naar) instructiemateriaal worden gegeven.

Hoofdstuk 6 richt zich op de visualisatie van een leerlijn, die één van de ontwerpprincipes uit hoofdstuk 5 is. Het hoofdstuk beschrijft twee studies die de geschiktheid van een grafische visualisatie onderzoeken, aangezien leerkrachten moeite blijken te hebben met het begrijpen en gebruiken van een tabelweergave. Een prototype van een grafische visualisatie in Groeimeter is gebruikt om data te verzamelen.

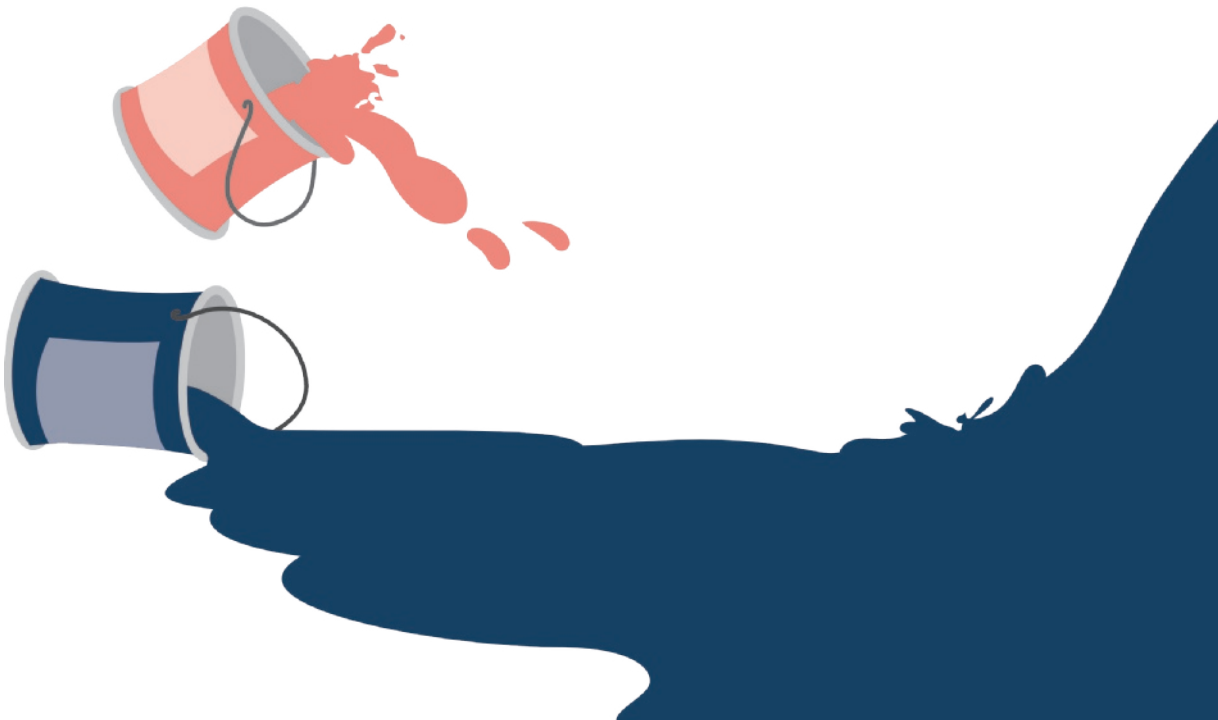
Het doel van studie A is de grafische structuur van het prototype te valideren door toetsdata van leerlingen te modelleren met Bayesiaanse Netwerk Analyse. Er is een toets afgenomen bij 787 leerlingen. Deze toets meet de prestaties van leerlingen op elf leerdoelen. De resultaten laten meerdere afhankelijkheden tussen de leerdoelen zien, wat de grafische structuur bevestigt.

Studie B analyseert het begrip en de voorkeuren van leerkrachten met betrekking tot de grafische visualisatie. Het prototype werd gedurende twee maanden door 19 leerkrachten in een natuurlijke klasomgeving gebruikt. Interviewresultaten laten zien dat leerkrachten moeite hebben om de visualisatie goed te begrijpen. Uit de geïdentificeerde ontwerpprincipes met betrekking tot de inhoud, structuur en bruikbaarheid van de leerlijn blijkt dat leerkrachten voorkeur geven aan een grafische visualisatie.

Samenvattend laat dit proefschrift zien dat het noodzakelijk en nuttig is om toetsinstrumenten te ontwerpen en te evalueren met de beoogde gebruikers, zodat zij de beoordelingsresultaten op een juiste manier begrijpen en inzetten. Er kan geconcludeerd worden dat formatieve toetsinstrumenten de volgende inhoudelijke kenmerken moeten bevatten: (1) fijnmazige informatie over de huidige prestaties van leerlingen (bijv. leerstrategieën, misvattingen), (2) informatie over de gewenste prestaties van leerlingen (bijv. leerdoel, leerlijn) en

(3) richtingen voor vervolgacties (bijv. leermaterialen, suggesties voor groeperen). Met betrekking tot de visuele presentatie van formatieve beoordelingsinstrumenten kan er geconcludeerd worden dat de meetfoutvisualisatie invloed heeft op de beslissingen van leerkrachten en dat een grafische visualisatie geschikt is voor het presenteren van een leerlijn.

Hoewel de kenmerken voor formatieve toetsinstrumenten het idee kunnen geven dat het eenvoudig is om deze instrumenten te ontwikkelen, laat dit proefschrift zien dat het ontwerp en de evaluatie veel complexer is. Het is te vergelijken met een evenwichtsoefening, waarbij keuzes voortdurend moeten worden afgewogen. We moeten allereerst zoeken naar een juiste balans tussen de voorkeuren en het begrip van leerkrachten. Daarnaast moeten we zorgen voor een gelijke weging van score-interpretatie- en scoregebruikinferenties in de validering van formatief toetsen. Verder laat de behoefte aan fijnmazige informatie zien dat we toetsvragen moeten gaan ontwikkelen die voldoende inzicht geven in de leerling, zonder de kracht van inferenties uit traditioneel gebruikte toetsvragen te verliezen. De interpretatieproblemen en informatiebehoefte geven bovendien aan dat er een goede integratie van het formatieve toetsinstrument en het formative toetsproces nodig is. Tot slot moeten we op zoek naar een juist evenwicht tussen formatief en summatief toetsen in het onderwijs. Dit alles biedt verschillende mogelijkheden voor toekomstig onderzoek. Bovenal roept het toetsontwikkelaars en onderzoekers op om verantwoordelijkheid te blijven nemen in de ontwikkeling van begrijpelijke en bruikbare formatieve toetsinstrumenten, zodat ook de verantwoordelijkheden bij het ontwerp en de implementatie van formatief toetsen voor alle betrokkenen in balans zullen zijn.

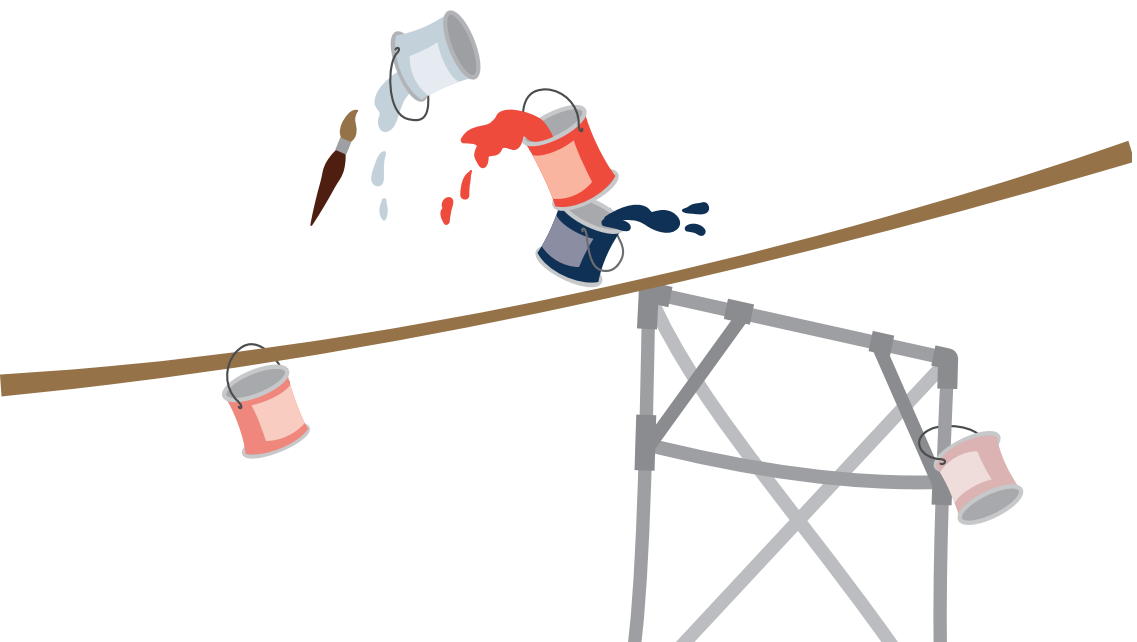




## DANKWOORD

In september 2015 ben ik gestart met het promotieonderzoek en dit proefschrift is het resultaat. De afgelopen jaren heb ik met veel plezier aan het onderzoek gewerkt. Graag wil ik Cito en Universiteit Twente bedanken voor het initiëren, faciliteren en financieren van mijn promotieonderzoek.

Het uitvoeren van een promotieonderzoek bleek ook een balancerende activiteit te zijn, waarbij het ging om een goed evenwicht tussen promotie- en nevenwerkzaamheden, aanwezig zijn op de Universiteit Twente en bij het Cito, en het lezen van wetenschappelijke artikelen (werk) en Jip & Janneke verhaaltjes (thuis). Daarom wil ik een aantal mensen bedanken die op verschillende wijze hebben bijgedragen aan het vinden en behouden van deze balans.



Theo,

In 2014 ontving ik een mailtje van jou met de vraag of ik een literatuuronderzoek naar toetskwaliteit wilde doen. Het zou een half jaar duren, niet wetende dat ik vijf jaar later nog steeds onder jouw begeleiding zou werken. Bedankt voor alle ontwikkelmogelijkheden die je me hebt gegeven. Je opbouwende feedback op globaal en vaak zeer gedetailleerd niveau en de ruimte voor persoonlijke gesprekken maakten je tot een superfijne promotor. Ik wens je een heerlijk en welverdiend pensioen!

Bernard,

Voordat ik met mijn promotieonderzoek begon, waren we op een dag samen op pad voor het RCEC. Toen we het hadden over een mogelijk promotietraject gaf jij aan dat je alle vertrouwen in mij had, maar dat ik zelf nog moest bedenken of ik het wel zou willen. Dit vertrouwen ben je ook altijd blijven geven in de jaren die volgden. Het binnenlopen voor een informeel praatje op de dagen dat ik op de UT werkte en je ontspannen levenshouding heb ik erg gewaardeerd.

Saskia,

Als dagelijks begeleider heb jij mij echt door mijn promotieonderzoek heen geholpen. Jij was mijn eerste aanspreekpunt en telkens weer beschikbaar om even mee te denken. Mijn eerste concepten gingen altijd naar jou, waardoor je enorm veel hebt moeten lezen. Na afloop van een voortgangsgesprek had ik steeds weer concrete ideeën om mee verder te gaan. Dank je wel voor je vele tijd, persoonlijke aandacht en gezelligheid!

UT-collega's,

Na een periode als student volgde een periode als medewerker van de universiteit. Het was erg leuk om deel uit te maken van de OMD-afdeling. Maaiké, we hebben een ontzettend fijne tijd als kamergenootjes gehad. Dank je wel voor alle gesprekken over ons promotie- en mamaleven. Jolien, wat leuk dat je mijn collega bent geworden. Ik vond het erg gezellig dat je regelmatig even binnen kwam lopen voor een kopje thee en dat we konden bijkletsen op weg naar huis. Ik vind het dan ook erg leuk dat jullie mijn beide paranimfen willen zijn. Martina en Nathalie, heel leuk dat ik opnieuw met jullie mocht samenwerken. Lorette en Birgit, dank jullie wel voor alle secretariële ondersteuning.

### Cito-collega's,

Dank jullie wel voor de leuke en leerzame tijd. Ik kijk met plezier terug op de gezellige afdelingsuitjes en goed gevulde snoeppot. Elise, Marije en Nikky, dank jullie wel voor de gezellige ViVjes en de hulp bij statistische analyses. Remco en Elske, het was leuk om samen met jullie aan een paper te werken. Marleen, Patricia en collega's van process support, bedankt voor jullie ondersteuning bij de werving van scholen. Servaas, bedankt dat je mijn Excel-expert wilde zijn. Floor, Judith en Anneke (SLO), bedankt voor jullie inhoudelijke bijdrage op het gebied van rekenen. Jos, fijn dat je altijd bereid was om mee te denken. Anke, bedankt voor het meelezen met één van mijn artikelen. Patrick, Marcel en Tjeerd Hans, dank jullie wel voor jullie technische hulp. Ivailo, bedankt voor je advies over de opmaak van het proefschrift. Rianne, bedankt voor je continue interesse in mijn algeheel welbevinden. Marica en Ilse, ik kijk terug op een leuke samenwerking tijdens het kennisdelingsproject. Ghita en Romy, wat leuk dat jullie het afgelopen jaar het Citolabteam nog leuker hebben gemaakt.

### Fontyscollega's,

Parallel aan mijn promotieonderzoek mocht ik helpen bij de ontwikkeling en uitvoering van de Master Toetsdeskundige. Dank jullie wel voor deze ervaring. Desirée, het was ontzettend fijn om met je samen te werken. De concrete taken waren een welkome afwisseling tijdens het onderzoek. Ook kijk ik terug op hele leuke congresweken in onder andere Finland en Canada.

### ICO-studenten,

Bedankt voor de gezelligheid tijdens de ICO cursussen en congressen. Leuk om jullie te leren kennen en onderzoek uit hele andere domeinen van het onderwijs te volgen.

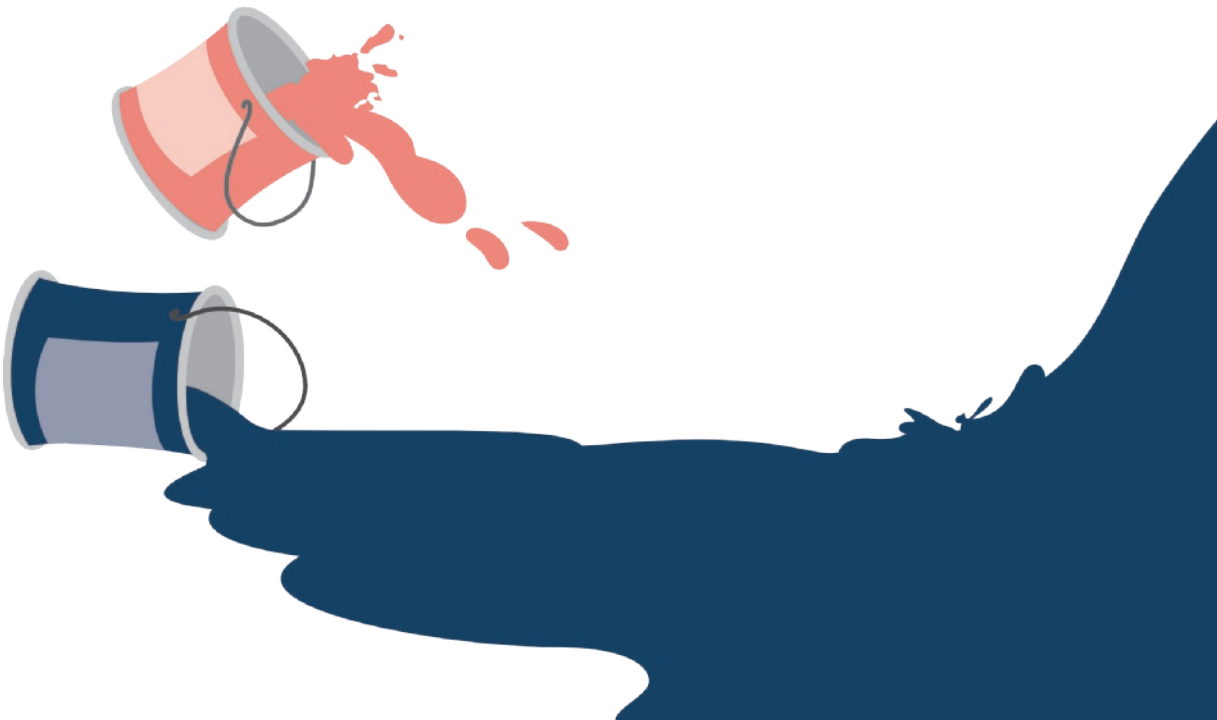
### Elke en Jorieke,

Samen hebben we onze interesse voor het onderwijs ontwikkeld tijdens de studie Onderwijskunde. Jullie waren ontzettend fijne studiegenootjes en zijn nog steeds hele lieve vriendinnen. Bedankt voor de gezellige tijd en jullie medewerking aan mijn onderzoek.

### Familie,

Bedankt voor jullie liefde en steun. Papa en mama, bedankt voor de wijze raad, huishoudelijke ondersteuning en het oppassen. Gert, bedankt dat je meerdere keren proefpersoon voor het onderzoek wilde zijn. Carlijn, dank je wel voor de gezellige en ontspannen zussendagen en spellingcontroles. Schoonouders, bedankt voor het verzorgen van de oppasdagen en heerlijke maaltijden. Tot slot, Mark, de start van mijn promotieonderzoek viel in dezelfde maand als ons huwelijk. Bedankt voor al je liefde, rust en relativerend vermogen. En wat een prachtig wonder dat we Renske hebben gekregen. Ik hoop dat we nog vele jaren samen mogen ontvangen.







## PUBLICATIONS AND PRESENTATIONS

### PUBLICATIONS IN SCIENTIFIC JOURNALS

Hopster-den Otter, D., Wools, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2019). A general framework for the validation of embedded formative assessments. *Journal of Educational Measurement*. doi:10.1111/jedm.12234

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

### BOOK CHAPTERS

Wools, S., Molenaar, M., Hopster-den Otter, D. (2019). The validity of technology enhanced assessments - threats and opportunities. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Cham, Switzerland: Springer International Publishing.

## CONFERENCE PRESENTATIONS

Hopster-den Otter, D. (2019). Meetfout in toetsrapportages. Maakt het verschil? Presentation at the Nederlands Platform voor Survey Onderzoek (NPSO), Tilburg, the Netherlands.

Hopster-den Otter, D. (2019). Ontwerpprincipes van formatief toetsen. Paper presentation at the Onderwijs Research Dagen (ORD), Heerlen, the Netherlands.

Hopster-den Otter, D. (2019, april). A general framework for the validation of formative assessments. Paper presentation at the National Council on Measurement in Education (NCME), Toronto, Canada.

Hopster-den Otter, D. (2019). Presenting measurement error in test score reports: Does it matter? Poster presentation at the National Council on Measurement in Education (NCME), Toronto, Canada.

Hopster-den Otter, D. (2018). A general framework for the validation of formative assessments. Paper presentation at the Association for Educational Assessment (AEA)–Europe, Lent, the Netherlands.

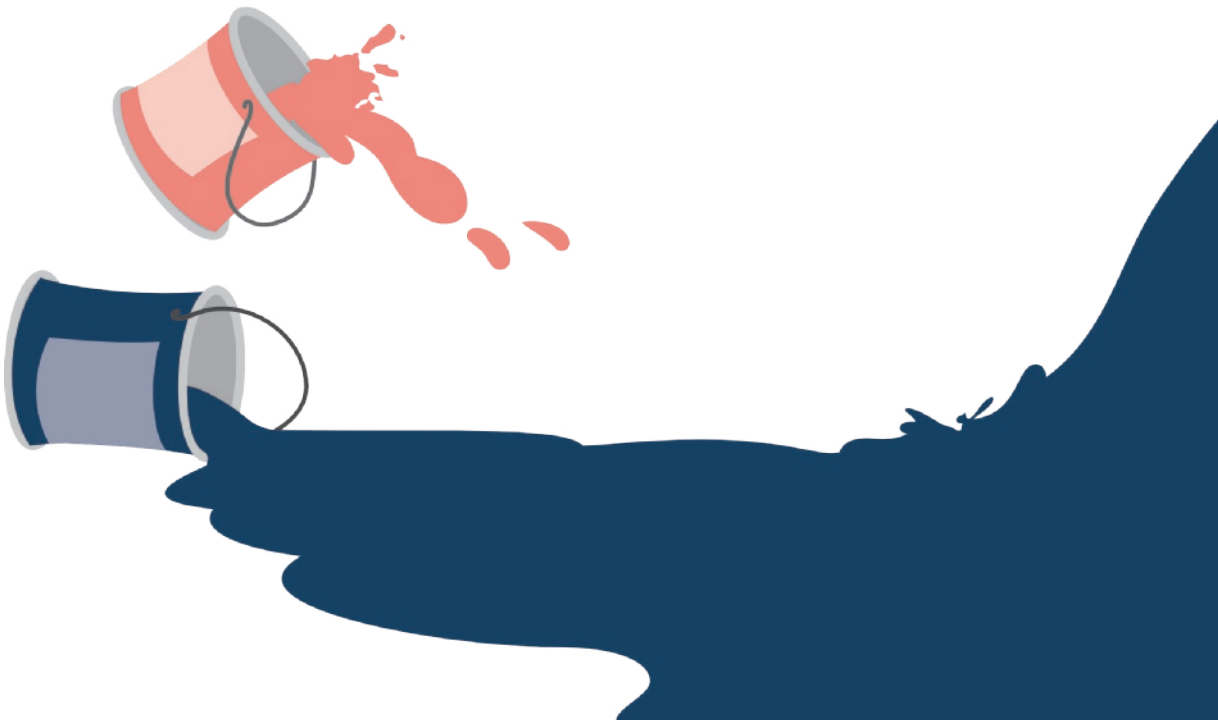
Hopster-den Otter, D. (2018). Een kader voor het valideren van formatief toetsen. Paper presentation at the Onderwijs Research Dagen (ORD), Nijmegen, the Netherlands.

Hopster-den Otter, D. (2017). Formative use of test results: A user's perspective. Paper presentation at the European Association for Research on Learning and Instruction (EARLI), Tampere, Finland.

Hopster-den Otter, D. (2017). Presenting measurement error in test score reports: Does it matter? Poster presentation at the European Association for Research on Learning and Instruction (EARLI), Tampere, Finland.

Hopster-den Otter, D. (2017). Feedback aan leerkrachten in toetsrapportages: De invloed van meetfout. Paper presentation at the Onderwijs Research Dagen (ORD), Antwerp, Belgium.

Hopster-den Otter, D. (2016). Effectieve rapportage van formatieve toetsen: Een behoefteanalyse. Paper presentation at the Onderwijs Research Dagen (ORD), Rotterdam, the Netherlands.





## ICO DISSERTATION SERIES

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University, and formerly University of Groningen (until 2006), Radboud University Nijmegen (until 2004), and Tilburg University (until 2002). The University of Groningen, University of Antwerp, University of Ghent, and the Erasmus University Rotterdam have been 'ICO 'Network partner' in 2010 and 2011. From 2012 onwards, these ICO Network partners are full ICO partners, and from that period their dissertations will be added to this dissertation series.

373. Bouwmans, M.H.C.F. (12-01-2018) The role of VET colleges in stimulating teachers' engagement in team learning. Wageningen: Wageningen University.

374. Jansma, D.J. (25-01-2018) This is wrong, right? The role of moral components in anti- and prosocial behaviour in primary education. Groningen: University of Groningen.

375. Okkinga, M. (02-02-2018) Teaching reading strategies in classrooms- does it work? Enschede: University of Twente.

376. Thomsen, M. (09-02-2018) Teachers Trust. Measurement, sources and consequences of teacher's interpersonal trust within schools for vocational education and training. Amsterdam: University of Amsterdam.

377. Van der Wurff, I.S.M. (09-02-2018) Fatty acids, Cognition, School Performance and Mental Well-Being in Children and Adolescents. Heerlen: Open University of the Netherlands.
378. Raaijmakers, S.F. (16-02-2018) Improving self-regulated learning: Effects of training and feedback on self-assessment and task-selection accuracy. Utrecht: Utrecht University.
379. Zhao, X. (07-03-2018) Classroom assessment in Chinese primary school mathematics education. Utrecht: Utrecht University.
380. Van Rooij, E.C.M. (15-03-2018) Secondary school students' university readiness and their transition to university. Groningen: University of Groningen.
381. Vanlommel, K. (26-03-2018) Opening the black box of teacher judgement: the interplay of rational and intuitive processes. Antwerp: University of Antwerp.
382. Boevé, A.J. (14-05-2018), Implementing Assessment Innovations in Higher Education. Groningen: University of Groningen.
383. Wijsman, L.A. (30-05-2018) Enhancing Performance and Motivation in Lower Secondary Education. Leiden: Leiden University.
384. Vereijken, M.W.C. (22-05-2018) Student engagement in research in medical education. Leiden: Leiden University.
385. Stollman, S.H.M. (23-05-2018) Differentiated instruction in practice: A teacher perspective. Leiden: Leiden University.
386. Faddar, J. ( 11-06-2018) School self-evaluation: self-perception or self-deception? Studies on the validity of school self-evaluation results. Antwerp: University of Antwerp.
387. Geeraerts, K. (25-06-2018) Dood hout of onaangeboorde expertise? Intergenerationale kennisstromen in schoolteams. Antwerp: University of Antwerp.
388. Day I.N.Z. (28-06-2018), Intermediate assessment in higher education Leiden: Leiden University
389. Huisman, B.A. (12-09-2018) Peer feedback on academic writing. Leiden: Leiden University.



390. Van Berg, M. (17-09-2018) Classroom Formative Assessment. A quest for a practice that enhances students' mathematics performance. Groningen: University of Groningen.
391. Tran, T.T.Q. (19-09-2018) Cultural differences in Vietnam : differences in work-related values between Western and Vietnamese culture and cultural awareness at higher education. Leiden: Leiden University
392. Boelens, R. (27-09-2018) Studying blended learning designs for hands-on adult learners. Ghent: Ghent University.
393. Van Laer, S. (4-10-2018) Supporting learners in control: investigating self-regulated learning in blended learning environments. Leuven: KU Leuven.
394. Van der Wilt, F.M. (08-10-18) Being rejected. Amsterdam: Vrije Universiteit Amsterdam.
395. Van Riesen, S.A.N. (26-10-2018) Inquiring the effect of the experiment design tool: whose boat does it float? Enschede: University of Twente.
396. Walhout, J.H. (26-10-2018) Learning to organize digital information Heerlen: Open University of the Netherlands.
397. Gresnigt, R. (08-11-2018) Integrated curricula: An approach to strengthen Science & Technology in primary education. Eindhoven: Eindhoven University of Technology.
398. De Vetten, A.J. (21-11-2018) From sample to population. Amsterdam: Vrije Universiteit Amsterdam.
399. Nederhand M.L. (22-11-2018) Improving Calibration Accuracy Through Performance Feedback. Rotterdam: Erasmus University Rotterdam.
400. Kippers, W.B. (28-11-2018) Formative data use in schools. Unraveling the process. Enschede: University of Twente.
401. Fix, G.M. (20-12-2018) The football stadium as classroom. Exploring a program for at-risk students in secondary vocational education. Enschede: University of Twente.
402. Gast, I. (13-12-2018) Team-Based Professional Development – Possibilities and challenges of collaborative curriculum design in higher education. Enschede: University of Twente.





