DIAGNOSING ENGINEERING SKILLS IN PRIMARY CLASSROOMS

Dannie Wammes

dissertation

ISBN: 978-94-6473-163-7
DOI: https://doi.org/10.33540/1857
NUR: 841
Cover design and layout: Goedinbeeld.nl
Printed by: Proefschrift.nl

This research was supported by the
Netherlands Organisation for Scientific Research
(NWO) under project number: 023.007.027 and
by the HAN University of Applied Sciences.

DIAGNOSING ENGINEERING SKILLS IN PRIMARY CLASSROOMS

Vaststellen van technische vaardigheden in basisschoolklassen

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht ‘
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 1 september 2023 des ochtends te 10.15 uur

door

Daniël Franciscus Wammes
geboren op 5 juni 1956
te Utrecht

Promotor:
Prof. dr. L. Kester

Copromotoren:
Dr. W. D. Schot
Dr. B. Slof

Beoordelingscommissie:
Prof. dr. T.A.J.M. van Gog
Prof. dr. F. Janssen
Prof. dr. M.G. Kleinhans
Prof. dr. J.W.F. van Tartwijk
Dr. C.M.M. Vloet

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO/023.007.027) en de Hogeschool van Arnhem en Nijmegen.

TABLE OF CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

Engineering is included in the curriculum of primary schools in many countries. This relates to the increasing role of technology in contemporary society (Lachapelle & Cunningham, 2014). From birth, people use technical devices and constructions to conserve and prepare food, learn and play, transport, use energy, shelter, work, communicate and amuse themselves. Therefore, people must learn how to handle, operate, assemble, and maintain these devices. Moreover, engineers will be necessary to design and develop future devices and constructions that will help meet the technical challenges in the future. So, primary schools should familiarise children with technical devices and raise their interest in engineering (Barlex et al., 2017).

Engineering aims to solve technical problems. Solving technical problems is an iterative process that involves determining the conditions for a proper design, constructing design models, developing the materials and techniques, testing and maintaining them. It is a systematic way of working focused on optimising materials and techniques (Carr et al., 2012). This implies that engineering includes various activities (see Cunningham & Kelly, 2017 for an overview). Problem-solving, considering problems in context, envisioning multiple solutions, designing, system-thinking, discussing implications of technologies and working in teams are examples of such activities that can be done even at the primary school level. The common ground for these activities is that they relate to systems that are used to satisfy human wants and needs (International Technology Education Association, 2007).

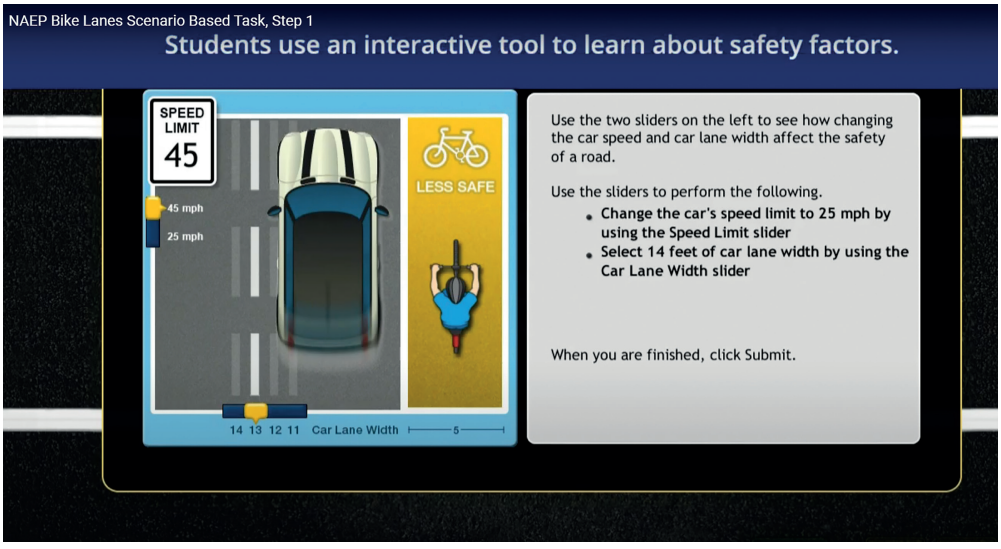
Although the relevance of engineering in primary education is widely acknowledged, its implementation in daily classroom practice is still limited. This relates to the focus on basic skills such as language and mathematics in primary education. When that focus dominates school policy, there is no incentive for teachers to develop the specific knowledge and skills required to teach engineering and other system-related subjects (Forbes et al.,, 2015; Guzey et al., 2014; Hartell et al., 2015; Hourigan et al., 2021; Hammack & Ivey, 2019). However, there are schools where principals and teachers consider engineering a valuable topic that offers many developmental opportunities. At these ‘engineering-minded’ schools, teachers try to optimise their lessons on the basis of the available didactical knowledge. As one of the obstacles in that process, teachers indicate that they find it difficult to establish their pupils’ technical skills, for which they depend on accidental observations within lessons. Insight into these skills and their development would offer them opportunities to adapt their teaching to differences in prior knowledge (e.g., skill level), evaluate the effectiveness of their lessons, inform pupils and parents about their skills and progress in this domain and communicate with colleagues about successful approaches.

Diagnosing skills relevant to engineering

The difficulties with identifying technical skills mentioned by the teachers of our 'engineering-minded' schools are also known from research (Culver, 2012; Moreland & Jones, 2000; Potgieter, 2012). These difficulties relate to the complex nature of skills that engineering requires: systems thinking and tacit knowledge (Niiranen, 2021). What should be learned about engineering has been analysed and described in the Standards for Technological Literacy (International Technology Education Association, 2007). Guidelines for assessment based on these standards (Garmine & Pearson, 2006) resulted in the Technology and Engineering Literacy (TEL) assessment, first administered in 2014 (National Assessment of Educational Progress, 2018). This TEL assessment is computer-based and includes interactive scenario-based tasks (see Figure 1 for an example).

Figure 1

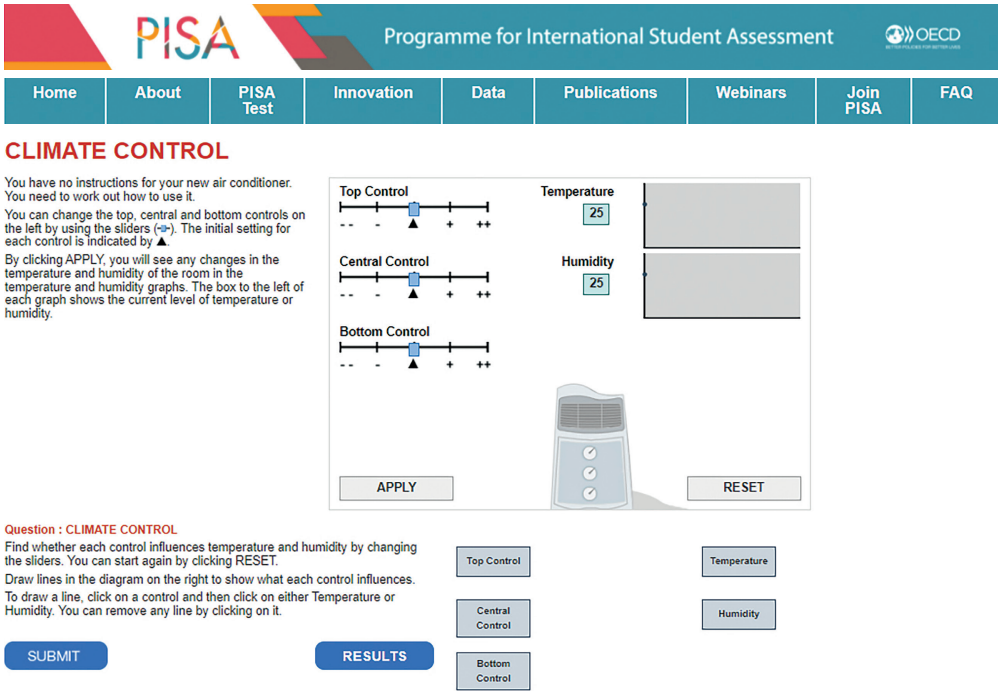
Screenshot of TEL Bike Lanes Scenario Based Task (National Assessment of Educational Progress, 2018)



Another assessment related to engineering and specifically directed to problem-solving is the PISA 2012 assessment on Creative Problem Solving (OECD, 2014; Csapó & Funke, 2017). This assessment included multiple interactive tasks (see Figure 2 for an example). Students could show their ability to explore, analyse and solve real-life technical problems with simulations of systems like the MP3 player, climate control and the ticket machine.

Figure 2

Screenshot of interactive PISA 2012 Creative Problem-Solving climate control task



Both TEL and PISA assessments were designed to identify a general technological literacy of 13-14-year-olds and a general problem-solving ability of 15-year-olds, respectively. These abilities were related to the background of the participants, their schools and their countries. Due to their design and purpose, these assessments are not suitable nor available for teachers, but the idea to create tasks that allow pupils to act on systems has been adopted in this dissertation.

This dissertation is not the first initiative to strengthen the diagnostic capacity of teachers. Such initiatives were undertaken at several places. For instance, in New Zealand, researchers used a training-on-the-job approach in which teachers learned to recognize pupils' technical knowledge and skills (Moreland & Jones, 2000). This enabled teachers to select tasks based on their value for the further development of knowledge and skills. Another approach was followed by van Niekerk et al. (2010). They supported teachers to develop rubrics related to several stages of the design activity that they engaged in with their students. The main advantage of these rubrics was that both teachers and learners could use them to set goals and evaluate progress. The disadvantage was that context-specific rubrics were needed for every design activity. Moreover, the teachers also needed considerable pedagogical content knowledge about engineering and design to make these rubrics. Unfortunately, most

teachers in primary schools lack such expertise (Compton & Harwood, 2005; Hartell et al., 2015; Nadelson et al., 2013; Rohaan et al., 2012; Svensson, 2018).

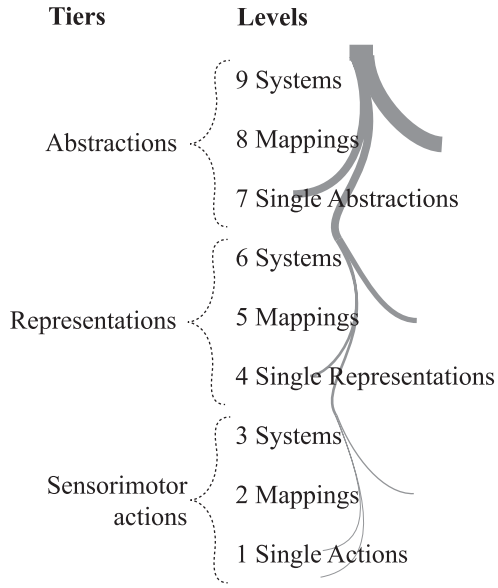
Although both initiatives were successful in improving the participants' diagnostic abilities, they lacked any follow-up. Costs and time investment of teachers are probably the main reasons (Altrichter, 2006) and should be taken into consideration in attempts to extend teachers' diagnostic abilities as these are barriers to application.

Skill development

Research related to the PISA 2012 assessment showed that generic problem-solving skills especially develop between the age of 12 and 15 (Molnár et al., 2013). This implies that in primary education, pupils' understanding of construction can differ considerably from their understanding of a pneumatic system. Skill development can only be compared within and between persons and specific contexts when using a common ruler. For skills related to systems, the Fischer scale (Fischer, 1980) has been advocated (Sweeney & Sterman, 2007). This scale has been used in research on social behaviour (Fischer & Bidell, 2007), language (Bassano & Van Geert, 2007) and science (Meindertsma et al., 2014). It also has been applied to identify skill development in the context of material-based systems (Parziale, 2002; Van der Steen, 2014). Therefore, this dissertation will use the Fischer scale as a common ruler to describe skill development.

Figure 3

Dynamic Skill Development, based on Fischer and Bidell (2007)



The Fischer scale (see Figure 3) starts with sensorimotor actions directly based on contextual information. For instance, a clamp makes it possible to connect; a bar makes it possible to stack. Representations are memories about events that are used to select actions. Representations are constructed from previously performed sets of sensorimotor actions. For instance, connecting a lamp with a battery in a circuit or ranking bars according to their length requires former experience. Such coordination of actions cannot be based on the properties of the materials only. Abstractions are based on a successive integration of representations, which leads to the use of rules that apply to a broad range of different objects and contexts. Using two objects to create an off/on switch in an electric circuit might indicate abstract thinking as it requires understanding the behaviour of electric current, which builds upon successive representations.

Fischer's model subdivides each of these main categories into three levels of increasing complexity: single use (of an action, representation or abstraction), mappings (a combination of two actions, representations or abstractions) and systems (combining all actions, representations or abstractions that are part of a system). From the studies based on Fischer's model, it is known that pupils seldom perform beyond the level of single abstractions before they enter secondary education.

Understanding and promoting skill development

Diagnostic data are important for teachers as it offers information about what pupils know and doesn't know (Oudman et al., 2018). Such information affects teachers' expectations about pupils' capabilities (Südkamp et al., 2012) which in turn may affect pupils' learning (Baudson et al., 2016; Lee et al., 2015). Improving teachers' diagnostic capability will reduce bias in those expectations and may improve pupils' learning, especially when teachers provide feedback that brings pupils' thinking forward (Behrmann & Souvignier, 2013; Helmke & Schrader, 1987). Training can improve teachers' diagnostic ability and their ability to provide adequate feedback (Thiede et al., 2018) and adaptive tasks (Ostermann et al., 2018). It has been demonstrated that offering adaptive tasks may result in better learning gains (Corbalan et al., 2008; Orvis et al., 2008).

Main research question and associated studies

For engineering, primary school teachers find it difficult to establish their pupils' technical skills. Insight into these skills and their development prior to their lessons would offer them opportunities to adapt their teaching to differences in prior knowledge (e.g., skill level), evaluate the effectiveness of their lessons and communicate about these skills and their development. We assume that a diagnostic tool feasible for classroom use might support teaching in this domain.

The main research question of this dissertation is, therefore: 'Which support can a diagnostic tool designed for classroom use offer teachers to infer and promote pupils' engineering skills?' This dissertation intends to answer this question for teachers of pupils aged 9 to 12 and for skills related to material-based technical systems.

The first step in answering our main research question was the design of a diagnostic tool. **Chapter 2** focuses on the development and validation of such a tool guided by the question: "How to assess primary education pupils' prior knowledge about technological systems in a valid manner?" We used the Evidence Centred Design method (Mislevy & Riconscente, 2005) to substantiate all design decisions for the assessment systematically. An important starting point for those decisions was that costs and time investment should be limited, as it has become clear that these are an obstacle to use. Due to that limitation, some of the design decisions had to be verified as they might affect the validity and reliability of the diagnostic tool.

Teachers find it difficult to establish pupils' engineering skills, which makes them insecure about their estimates. We assume that the diagnostic tool supports teachers in establishing engineering skills. We also know that a diagnostic tool will only be implemented when the value that teachers assign to the information provided by the tool outweighs the time and effort needed for test-taking (Chien et al., 2014). **Chapter 3** describes the results of a study carried out in eight classrooms to answer the question: "How do teachers value the application of the tool in their class?" Besides its practical relevance, this study is the first one to explore teacher judgement accuracy in a context where teachers had no knowledge of test results. Research on judgement accuracy has focussed on subjects like mathematics, language arts and science, subjects for which assessment results are known to the teacher who judges pupils' capabilities (Urhahne & Wijnia, 2021). In our study, teachers made and substantiated their estimates of their pupils' technical skills prior to test-taking. Afterwards, they reflected on the process of test-taking and what meaning they awarded to the results in view of their previous estimates.

Insight into engineering skills and their development is an important prerequisite for teaching the subject and communicating about these skills. With such insight, teachers are better able to provide tasks, instruction and feedback that offers pupils the opportunity to improve their skills (Van de Pol et al., 2010). One of the obstacles that teachers face when interpreting and reacting to pupils' actions may be their own knowledge. Providing adequate feedback, for instance, may be hindered by insufficient knowledge about the domain (Kramer et al., 2021) but also by teachers that assume that pupils think about systems as they do (Nickerson, 1999). That might result in providing information that is not

understood as it does not tap into pupils' prior knowledge. In order to provide effective instruction and feedback, a teacher should be able to take the pupils' points of view (Ostermann et al., 2018). **Chapter 4** describes a study designed to establish the effects of a short-term course, based on Nickersons' anchoring and adjustment model, on the diagnostic ability of pre-service teachers. In this course, the participants got an insight into their own understanding of material-based systems. Skill development was explained by the Fischer scale and associated theory (e.g. Edelman & Tononi, 2000; Thelen & Smith, 1994; Van der Steen, 2014). Results of the diagnostic tool were used to make the teachers aware of pupils' comprehension of material-based systems in various stages of skill development. The study used a pre-post test design to identify changes in the teachers' diagnostic ability and self-efficacy beliefs in relation to the course.

Insight into technical skills and their development prior to lessons creates the opportunity to adapt teaching to differences in prior knowledge (e.g., skill level). Reviews have shown the value of such an adaptive approach for pupils' learning in subjects like mathematics, reading comprehension and science (e.g., Deunk et al., 2018; Smale-Jacobs et al., 2019). However, no information about adaptive teaching is available for engineering in primary education. The diagnostic tool offers the opportunity to evaluate the effectiveness of adaptive measures in this context. A pre-post design was used to establish the effects of adaptive task selection on pupils' ability to reconstruct an electric circuit. **Chapter 5** describes the results of this study which was set up to answer the question: "What does adaptive task selection contribute to the learning outcomes of a lesson about electric circuits?".

To conclude, the initial question "how to identify pupils' skills in engineering" has been elaborated in four studies which all provided information to answer our main research question: 'Which support can a diagnostic tool, designed for classroom use, offer teachers to infer and promote pupils' engineering skills?' **Chapter 6** will present and discuss our final conclusions, their limitations and their practical and scientific value.

CHAPTER 2

PUPILS' PRIOR KNOWLEDGE ABOUT TECHNOLOGICAL SYSTEMS: DESIGN AND VALIDATION OF A DIAGNOSTIC TOOL FOR PRIMARY SCHOOL TEACHERS

This chapter is based on:
Wammes, D., Slof, B., Schot, W., & Kester, L. (2022a). Pupils' prior knowledge about technological systems: design and validation of a diagnostic tool for primary school teachers. International Journal of Technology and Design Education, 32(5), 2577-2609. https://doi.org/10.1007/s10798-021-09697-z

Acknowledgement of author contributions:
DW and BS designed the study; DW collected the data, planned the data analyses and analysed the data; DW and BS drafted the manuscript; DW, BS, WS and LK contributed to the critical revision of the manuscript and supervised the study.

Abstract

This study aimed to develop and validate, based on the Evidence Centered Design approach, a generic tool to diagnose primary education pupils' prior knowledge of technological systems in primary school classrooms. Two technological devices, namely the Buzz Wire device and the Stairs Marble Track, were selected to investigate whether theoretical underpinnings could be backed by empirical evidence. Study 1 indicated that the tool enabled pupils to demonstrate different aspects of their prior knowledge about a technological system by a wide variety of work products. Study 2 indicated that these work products could be reliably ranked from low to high functionality by technology education experts. Their rank order matched the Fischer-scale-based scoring rules, designed in cooperation with experts in skill development. The solution patterns fit the extended non-parametric Rasch model, confirming that the task can reveal differences in pupils' prior knowledge on a one-dimensional scale. Test–retest reliability was satisfactory. Study 3 indicated that the diagnostic tool was able to capture the range of prior knowledge levels that could be expected of 10 to 12 years old pupils. It also indicated that pupils' scores on standardised reading comprehension and mathematics tests had a low predictive value for the outcomes of the diagnostic tool. Overall, the findings substantiate the claim that pupils' prior knowledge of technological systems can be diagnosed properly with the developed tool, which may support teachers in decisions for their technology lessons about content, instruction and support.

Introduction

Technology affects many aspects of our social life, work and health care (Malik, 2014). Due to its importance, technology has been implemented in the curricula of primary schools in many countries (Compton & Harwood, 2005; Department for Education, 2013; Kelley, 2009; Rasinen et al., 2009; Seiter, 2009; Turja et al., 2009), initially as an independent subject, but recently as one of the cornerstones of an integrated STEM (science, technology, engineering, and mathematics) approach (Honey et al., 2014). The aim of technology education in primary schools is often twofold, namely a) evoking pupils' interest in technology (including its importance for society) and b) fostering pupils' understanding (concept and principles) of basic - e.g., electrical and mechanical - technological systems (De Grip & Willems, 2003; De Vries, 2005; Pearson & Young, 2002; Williams, 2013). Although the importance of technology education is acknowledged by teachers and school boards and, consequently, incorporated in many primary education curricula, a structural embedding in educational practices is often lacking (Chandler et al., 2011; Harlen, 2008; Hartell et al., 2015; Platform Bèta Techniek, 2013).

A possible explanation for this could be the limited pedagogical content knowledge and the low self-efficacy that many teachers experience when providing technology education (Hartell et al., 2015; Rohaan et al., 2012). Also, teachers who are confident in providing technology education often still experience difficulties when assessing (formative and summative) pupils' technology-related learning outcomes (Compton & Harwood, 2005; Garmine & Pearson, 2006; Moreland & Jones, 2000; Scharten & Kat-de Jong, 2012). A lack of knowledge about assessing and fostering pupils' understanding of technological systems properly, compromises the quality of technology education in primary schools (McFadden & Williams, 2020). It may also hinder a structural embedding of technology education curricula since policies on how to invest teaching-time are increasingly based on achieved learning outcomes in general (Slavin, 2002), for technology education as a specific subject (Garmine & Pearson, 2006) or within the context of STEM (Borrego & Henderson, 2014). Knowledge about learning outcomes does affect not only the composition of curricula at the national level (Harlen, 2012; Kimbell, 1997; Priestley & Philippou, 2018) but also the decisions taken at the school level (Arcia et al, 2011; Resh & Benavot, 2009) and the curricular practice at the classroom level, shaped by the day-to-day decisions on time-allocation taken by teachers (Siuty et al., 2018). This study tries to enhance the position and quality of primary technology education at the classroom level by supporting teachers in gaining more insight into their pupils' understanding of technological systems (Dochy et al., 1996). The study addresses this by developing and examining the validity of a diagnostic tool aimed at assessing pupils' prior knowledge of technological systems in primary schools. To this end, Mislevy's Evidence-Centered Design (ECD) approach (e.g., Mislevy et al., 2003; Oliveri et al., 2019) was utilised.

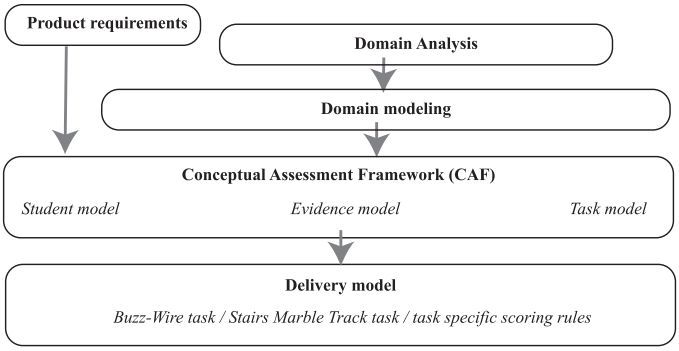
The Evidence-Centered Design approach

Evidence-Centred Design (ECD) was developed to facilitate a systematic design of large-scale assessments (Mislevy et al., 2003; Roelofs et al., 2021). However, its aim to substantiate validity by a systematic approach makes ECD valuable for the development of various kinds of assessments (see, for instance, Oliveri et al. 2019 and Clark-Midura et al., 2021).

The ECD approach is aimed at developing valid assessments (e.g., diagnostic tools) by utilising a stepwise four-layered design framework (see Figure 4). The design decisions made in preceding layers serve as input for the decisions made in subsequent layers.

Figure 4

Evidence-Centered Design (based on Riconscente, Mislevy, & Hamel, 2005)



The *domain analysis layer* focuses on describing the core characteristics of the (sub)domain for which the diagnostic assessment tool will be developed. This results in a general description of the type(s) of knowledge, skills and attributes (i.e., KSA's) that need to be assessed. The *domain modelling layer* addresses the operationalisation of the domain-related KSA's in terms of an interpretative validity argument, namely;

- What does the diagnostic tool specifically claim to assess?
- What are the underlying assumptions (i.e., warrants) for the claim?
- Which evidence (i.e., backings) can be provided to substantiate the assumptions?
- Which alternative explanation (i.e., rebuttals) might also be plausible?

Since the interpretative validity argument is the cornerstone for the decisions in the subsequent layers, it is vital that the decisions made in the domain modelling layer are properly substantiated by arguments (Kane, 2004; Kind, 2013; Zieky, 2014).

The *conceptual assessment framework* (CAF) layer focuses on operationalising the arguments into concrete design guidelines (i.e., an assessment blueprint). To this end, the student model (i.e., specifying the KSA's into observable performance behaviour), task model (i.e., selecting assessment task(s) that elicit the intended performance behaviour), and evidence model (i.e., formulating rules for scoring the performance behaviour) should be specified. As the tool is developed for classroom use, these models should also match the product requirements - addressing the contextual (e.g., classroom) opportunities and limitations.

The *assessment implementation layer* addresses the actual development and implementation of the diagnostic tool. For example, documents describing the intended performance behaviour, the assessment task(s), scoring rules, and instructions for applying these materials to educational practices will be made available for the assessments.

DIAGNOSING PRIMARY EDUCATION PUPILS' PRIOR KNOWLEDGE ABOUT MATERIAL-BASED SYSTEMS

Based on the ECD approach, a diagnostic tool aimed at gaining insight into pupils' prior knowledge of material-based systems will be developed for primary education teachers. This section first describes the product requirements and design decisions (i.e., including its theoretical substantiations) that were made in the domain analysis, domain modelling, and conceptual assessment layers. Thereafter, the tasks required for utilising the diagnostic tool (i.e., assessment implementation layer) will be described.

Product requirements

In the context of primary technology education (e.g., 25 pupils in a classroom), it is often not feasible for teachers to observe in real-time how all their pupils understand the technological system(s) at hand. Therefore, a diagnostic assessment tool designed for this context should preferably be based on outcome measures such as work products. This offers teachers the opportunity to diagnose and prepare appropriate remedial strategies, also after class hours (Van de Pol et al., 2014). Another requirement is that the tool can be used in a time-efficient manner. Since the time available for technology education is often limited, the time teachers require to carry out the diagnoses and prepare their lessons with adequate instruction and feedback should be carefully balanced.

Domain analysis

Technology is often characterised by the activities humans carry out to modify nature to meet their needs (Pearson & Young, 2002). Three frequently mentioned technology-related activities are crafting, troubleshooting, and designing (Jonassen, 2010). Crafting (e.g., bricklaying, cooking by a recipe, and mounting Ikea furniture) will be left outside the scope of this study since it is characterised by a clear, often stepwise pre-described process towards generating a well-defined work product. This structure makes it relatively easy to establish where pupils encounter difficulties and need support. Difficulties in diagnosing arise when it comes to troubleshooting and design activities since they require the use of knowledge in the context of dealing with material-based systems. Technological systems are defined as "a group of interacting, interrelated, or interdependent elements or parts that function together as a whole to accomplish a goal" (ITEEA, 2007). Understanding technological systems implies that pupils recognize the interrelationship between input, processes and output (De Vries, 2005) and are able to create (i.e., design) or restore (i.e., troubleshooting) these kinds of interrelationships. Gaining a proper insight into pupils' prior knowledge about technological systems is challenging since at least three aspects should be considered.

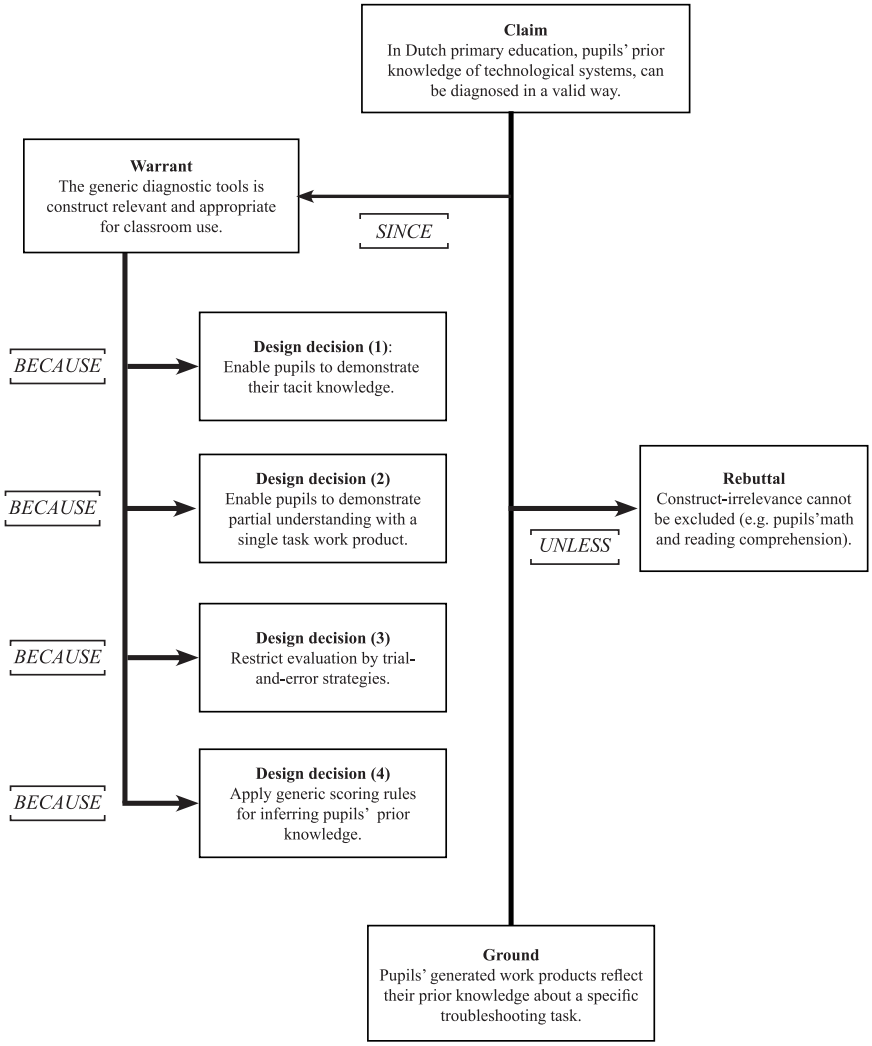
First, novice designers or trouble-shooters, like primary education pupils, often exhibit trial-and-error behaviour (Jonassen, 2010). This 'learning-by-doing' leads to 'knowing-that', which points to the often visual and procedural aspects of technological knowledge that cannot be learned by instruction or textbooks (De Vries, 2005). At the same time, trial-and-error behaviour complicates assessment: It is difficult to distinguish lucky guesses from prior knowledge-based actions without observation and questioning (Alfieri et al., 2011; Baumert et al., 1998). Secondly, understanding the interrelationship between input, processes, and output for a particular system does not automatically mean that pupils are also able to explain their knowledge adequately. Much technological knowledge is 'knowing-how' (De Vries, 2005). It includes procedural and visual knowledge, which is mostly tacit - and, therefore, difficult to verbalise (Hedlund et al., 2002; Mitcham, 1994). Thirdly, pupils' knowledge of technological systems is often limited to the ones they are already familiar with (Baumert et al., 1998; CITO, 2016; Jonassen & Hung, 2006). For most pupils, the development of a more general ability to understand the structure of technological systems by inductive reasoning takes place in the first years of secondary education (Molnár et al., 2013).

Domain modelling

The general characterisation of the technology domain has implications for formulating the interpretative validity argument (see Figure 5). The domain modelling layer focuses on explicating the design rationale behind the diagnostic tool in terms of the assessment argument; the more robust the underlying argument, the more valid the diagnostic tool's design. The interpretative validity argument starts with a *ground*, which usually is a score on a specific (performance) assessment. Based on the ground, a *claim* is made regarding the meaning and implications of the obtained score. In this study, the ground is a diagnostic score which represents a level of prior knowledge. To ensure that the diagnostic tool is valid, it is important to explicit the underlying *warrant(s)*. Here, the warrants address the question of why it is reasonable to assume that the diagnostic tool assesses construct-relevant (i.e., understanding of technological systems) pupil characteristics. The underlying assumptions should be explicated in the design decisions, which, in turn, should be substantiated by theoretical and, preferably also, empirical arguments. In the interpretative argument validity approach, this is coined as providing backings (i.e., evidence) for the warrants (i.e., design decisions). In the validation process, four design decisions were made. Below the underlying assumptions and associated theoretical arguments are provided. Based on this, the actual development and implementation of our assessment delivery model will be described.

Figure 5

Overview of interpretative validity argumentation (based on Oliveri et al., 2019)



Decision 1: Enable pupils to make use of their tacit knowledge. The domain analysis revealed that pupils' understanding of technological systems is often tacit. Since it is difficult for pupils to verbalise this kind of knowledge, they should be enabled to express their understanding in a manner that doesn't solely rely on verbalisation (Zuzovsky, 1999). To this end, it is essential that the diagnostic tool enables pupils to demonstrate their knowledge through their actions (Levy, 2012). By doing so, the design of the diagnostic tool aims to assess construct-relevant pupil characteristics.

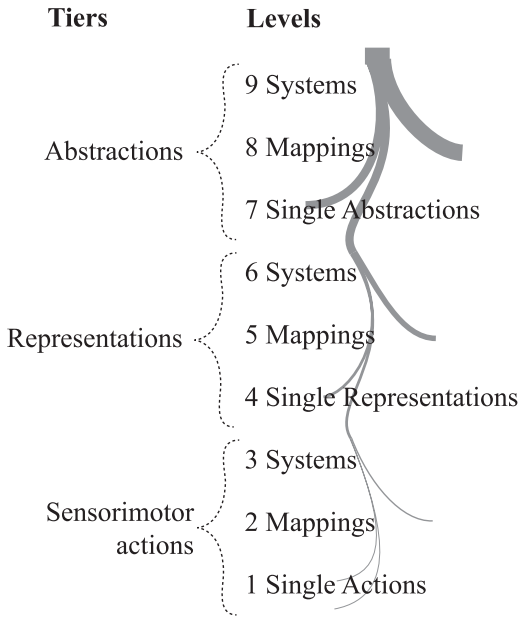
Decision 2: Enable pupils to demonstrate partial understanding with a single task work product. Administering performance-based diagnostic tools usually requires a considerable time investment (Davey et al., 2015). In technology education at primary schools, such time is limited, and therefore, a diagnosis should preferably be based on the work product of a single task. However, single tasks often limit demonstrating partial understanding after a mistake has been made (Greiff et al., 2015). The tasks' design should resolve such a limitation by allowing pupils to follow different pathways even after a mistake. That should result in a wide variety of work products, indicating differences in their prior knowledge. Since it is difficult to predict up-front whether pupils will generate such a wide variety of work products, empirical backings are needed to validate this design decision.

Decision 3: Restrict evaluation by trial and error strategies. Trial and error is for novices a dominant and valuable strategy to discover the behaviour and, through that, the structure of technical systems (Garmine & Pearson, 2006; Johnson, 1995; OECD, 2013). The domain analysis has indicated that it is hard to distinguish features of a work product generated by lucky guesses from aspects generated by prior knowledge. Because the diagnostic tool aims to assess prior knowledge, it should be plausible that a work product relates to knowledge gained from previous experiences and does not result from epistemic actions. Therefore the tool should limit 'learning from the task' by restricting the information about the systems' behaviour that trial-and-error might evoke (Klahr & Robinson, 1981; Philpot et al., 2017). The decision to restrict feedback could limit pupils' trial-and-error behaviour. That might affect the scope of options advocated by decision 2, that pupils consider while resolving a task. The empirical backing of decision 2 should indicate that this effect is limited.

Decision 4: Apply generic scoring-rules for inferring pupils' prior knowledge. In primary education, pupils' prior knowledge often differs per technological device (Baumert et al., 1998; CITO, 2015; Molnár et al., 2013). Comparing pupils' prior knowledge across different technological devices, thus, requires the utilisation of generic score-rules (Nitko, 1996). To this end, a framework for inferring pupils' understanding of different kinds of technological devices should be developed and substantiated with theoretical and empirical backing. From a theoretical viewpoint, Fisher's (1980) framework for describing the development of dynamic skills might offer relevant guidelines for developing generic scoring rules (see Figure 6).

Figure 6

Dynamic Skill Development, based on Fischer and Bidell (2007)



Dynamic skill development (e.g., developing an ability to understand, restore or create technological systems) evolves in three different phases (i.e., tiers). Each tier represents a specific kind of understanding which manifests itself in pupils' exhibited behaviour. In the first - sensorimotor - tier, pupils' behaviour (e.g., manipulations) is solely based on the sensorimotor information from the technological device. This implies that pupils do not have or use previous experiences to predict the consequences of their actions. In the second - representational - tier, pupils do apply knowledge, tacit or declarative, obtained from prior experiences to select their manipulations.

The main characteristic in which skills within the representational tier differ from those at the sensorimotor level is the need to apply knowledge about the behaviour of the system's components which cannot be observed on the spot. This difference is an important additional reason to restrict the systems' feedback on trial-and-error behaviour. Trials may occasionally evoke aspects of the system's behaviour that remain hidden for those who did not try a similar action. That would make it impossible to conclude, without continuous observation, that a pupil has used a representation or a visual clue.

In the third - abstraction - tier, pupils can apply general principles to guide their actions. Furthermore, Fischer's framework includes three sublevels for each tier to gain a more fine-grained insight into pupils' development. Each sublevel refers to

the extent to which pupils can interrelate the different device components properly. For instance, at the single-action level, pupils use the possibility to manipulate a device component without considering its interrelationship with the other components. This implies that these pupils have a less developed understanding compared to those who consider a component's relationship with another (i.e., mappings) or multiple other components (i.e., systems).

Although the Fischer scale might be a good model to describe the development of system-thinking skills (Sweeney & Sterman, 2007), and has been applied to infer levels of understanding in a non-verbal construction task (Parziale, 2002; Schwartz & Fischer, 2004), it has not yet been used to design a diagnostic tool for teachers. Such use is only valid when several conditions are met. First of all, the scale should be one-dimensional, requiring that work products can be reliably rank-ordered. Secondly, descriptions of the Fischer scale are highly abstract and difficult to interpret for teachers unfamiliar with Fischer's work. Therefore, the tool should have task-specific scoring rules corresponding with the original scale. Furthermore, scale validity should be demonstrated by comparing the levels generated by the task-specific scoring rules with an independent judgement about the quality of the work products. Finally, the work products resulting from a single task should be a reliable indicator of differences in prior knowledge (Novick, 1966).

Rebuttal: Considering construct irrelevant, alternative explanations. To ensure construct-relevance, it is also important to verify whether a similar diagnosis could be made by a teacher from sources of information that are already available, which would make the introduction of a new diagnostic tool unnecessary (i.e., *rebuttals*). For example, prior research revealed that primary education pupils' mathematics and reading ability scores are strong predictors of their academic achievement (Safadi & Yerushalmi, 2014; Wagensveld et al., 2014). Since understanding technological systems involves the application of scientific principles, it could be argued that pupils' math and reading abilities might predict the differences in pupils' levels of understanding technological systems. To ensure that the generic diagnostic tool has an added value for teachers, given pupils' math and reading ability scores, empirical backings are required (i.e., construct-relevance).

Conceptual assessment framework

Based on the validity argument in the domain model and the product requirements, concrete assessment design guidelines (i.e., an assessment blueprint) will be formulated in the CAF layer. This requires specifying the student, task, and evidence model.

Student model. The diagnostic assessment tool should be aimed at gaining insight into primary education pupils' general understanding of technological

systems. Pupils' levels of understanding manifest themselves by the behaviour that they exhibit at the sensorimotor, representational or abstract tier levels. Since pupils' levels of understanding differ substantially, it is important that the diagnostic tool's design includes a fine-grained scoring mechanism to capture this. For the intended target population - Dutch primary education - it is likely to assume that pupils (4 - 12 years old) are not yet able to reach the abstraction mapping level, which implies that they are generally not able to solve problems that require the combination of two different abstractions (Van der Steen, 2014). Consequently, a range of ability levels varying from the single sensorimotor actions level up to and including the single abstraction level should be adequate to diagnose pupils' prior knowledge.

Task model. The diagnostic assessment task should thus enable pupils to exhibit behaviour at the sensorimotor, representational, and single abstraction levels. This should result in different work products reflecting pupils' prior knowledge about the devices' technological system. Hence, in this context, pupils' understanding is inferred from the solution (i.e., the work product). Based on the design decisions in the domain model, this means that each task; (1) represents a technological system, (2) provides a rich diagnostic dataset (i.e., a variety of work products representing the different Fischer levels), (3) enables pupils to apply their tacit knowledge, and (4) restricts pupils experiencing the consequences of their trial-and-error behaviour (i.e., random manipulations). Preferably the diagnostic data can be gathered by administering a single diagnostic assessment task for a specific type of system, as this would limit teachers' time investment. Such a task should enable pupils to show their (partial) understanding of a single aspect or multiple functional aspects of the device without being able to reconstruct the whole system. To this end, the assessment task should be aimed at incorporating multiple device components (i.e., variables) which can be manipulated on their own and in combination. Only then does the generated work product manifests differences in pupils' understanding of its underlying technological principles.

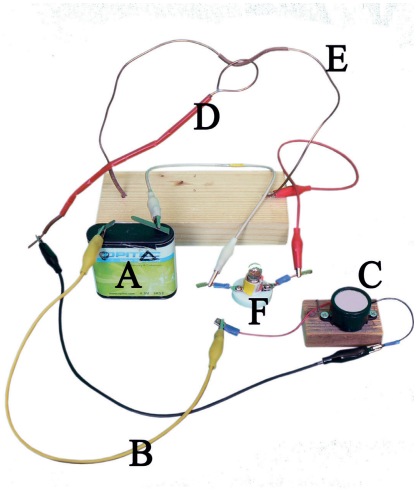
Evidence model. In addition to defining the different levels of understanding, rules for scoring them are required. The previously described Fischer levels are too abstract for directly inferring pupils' prior knowledge levels from the generated work products. To this end, specific scoring rules that match the generic levels should be utilised to determine which level best reflects the quality of the provided solution (i.e., the work product). Furthermore, the evidence resulting from the tasks and scoring rules should be considered from the psychometric viewpoint.

Assessment implementation.

This layer focuses on the actual development and implementation of the diagnostic assessment tool in educational practices. For the student-model, this means that

Figure 7

Buzz-Wire device



Buzz-Wire. A) Battery, B) Wire with crocodile clips, C) Buzzer, D) Loop, E) Copper spiral, F) Lamp with insulation on the tips of the connectors.

Figure 8

Stairs Marble Track device

Stairs Marble Track. A) Marble (not available in task), B) Camshaft with six eccentric wheels, C) Bar with slanted top, D) Handle (blocked).

all seven targeted levels of understanding should be described in a generic way. The task model states that the tasks should a) represent a technological system and b) enable pupils to manipulate (i.e., interrelate) its component in various ways. With some exceptions, like LEGO Mindstorms, most ICT-based devices used at primary schools are very restrictive in the possibilities to change the systems' properties and, therewith, do not fit the requirements of the task-model. Therefore two tasks were selected that are based on material-based systems that pupils can construct from scratch. The Buzz-Wire (BW) device (see Figure 7) is an electrical circuit that has been used in a Dutch national study on the quality of science and technology education in primary education (CITO, 2016). In the BW assessment task, pupils are asked to construct a Buzz-Wire circuit. Pupils can use different device components, such as a spiral with a fixed loop, an empty battery, a lamp, a buzzer and five wires with crocodile clamps. The Stairs Marble Track (SMT) is a mechanical device based on a camshaft (see Figure 8), which was used to examine pupils' scientific reasoning skills (Meindertsma et al., 2014). In the SMT diagnostic assessment task, pupils are asked to reconstruct the device by placing six bars in their correct position. Both tasks were slightly adapted so pupils would not experience the consequences of their manipulations. For the BW device, an empty battery was used, and for the SMT device, the handle was blocked, and the marbles

were left out. Pupils were informed about these restrictions, to avoid confusion when they did not notice an effect (does the device or a specific component operate properly?) of their actions. All device components were colour-coded in a way that resulted in unique combinations for different component states to allow for unambiguous coding of the pupils' work products. Based on the evidence model, a first draft of device-specific scoring rules was developed to infer pupils' general level of understanding from the work products (see supplementary material).

Study design and research questions

This study examines the validity of the generic diagnostic assessment tool's design by gaining more insight into the quality of the empirical arguments (i.e., backings). It aims to verify whether the theoretical arguments for the design decisions are backed by empirical evidence. That is, the tool's design should facilitate pupils to use their tacit knowledge (decision 1), enable them to demonstrate partial knowledge with a single task work product (decision 2), and restrict them from experiencing the consequences of trial-and-error behaviour (decision 3). It remains, however, to be seen to which extent pupils will use the possibilities that the assessment tasks (i.e., devices) provide them to generate a wide variety of work products, representing the differences in their prior knowledge. It could, for example, be that pupils of this age do not consider the various options due to the lack of opportunities to evaluate their actions (decision 3). They also might have similar notions about how to use some of the device's components (Defeyter & German, 2003; Matan & Carey, 2001). The first study addresses this by examining the variety of work products that pupils made when they were instructed to restore the SMT and BW device without being able to evaluate their actions.

The *second study* addresses the fourth design decision by examining the suitability of the scoring rules for inferring pupils' level of understanding of the material-based systems from their generated work products. Four requirements will be examined. a) Can work products be reliably arranged on a single dimension? b) Is it possible to construct task-specific scoring rules in compliance with the original Fischer scale? c) Do the levels generated by the scoring rules match an independent judgement about the quality of the work products? d) How reliable is it to use work products resulting from a single task as an indicator of differences in prior knowledge?

The *third study* explores whether the tool matches the student model, which states that pupils will demonstrate skill levels from a single sensorimotor action to a single abstraction. Moreover, this study aims to verify whether the formulated alternative explanations (i.e., rebuttal) can be rejected. Primary

education teachers' potential use of the developed diagnostic assessment tool also depends on their (perceived) added value of the tool. If a difference in pupils' understanding can be accounted for by other, already assessed constructs, it is probably not worth the effort to use the diagnostic assessment tool. As indicated in the domain model layer, pupils' math and reading ability scores might also predict their achievement in technology education. It is unclear yet how strong this effect is on pupils' understanding of electrical and mechanical systems. In case math and reading scores are strong predictors for pupils' understanding of these systems, teachers might not see the added value of using additional assessment instruments. The third study addresses this by examining the extent to which pupils' scores on standardised math and reading ability tests predict their diagnosed level of understanding of the material-based systems

In the following sections, the design and applied methodology for answering the research questions and the obtained findings will be described per study. This will be followed by an overarching discussion of the generic diagnostic assessment tools' validity, limitations, and implications for educational practices and research. Ethical approval for this study was provided by the faculty's ethical committee.

STUDY 1: VALIDATING THE VARIETY OF GENERATED WORK PRODUCTS

Participants and design

In total, 272 pupils (120 girls and 152 boys) from 17 different classrooms at seven Dutch primary education schools participated in this study. The pupils' average age was 11.0 years (*sd*=0.8, Min=8.9, Max= 13.6). Their schools were part of the first authors' professional network. The required parental consent was passive or active, depending on school regulations. When parents objected, which happened three times, no data for their child was collected. The assessment tasks were administered in an individual setting outside the regular classroom in the presence of the first author. Each pupil first watched a one-minute introductory video (made and provided by the first author) which briefly showed how the BW and SMT devices operated without revealing their configuration. Thereafter, the separate device components were shown, and each pupil was asked to re-configure the device so it would operate properly again. A maximum of five minutes was set to complete each assessment task. Pupils were informed that they would not be able to verify whether their actions were (in)correct due to the restriction of trial-and-error behaviour. The design was counter-balanced (half of the pupils started with the BW device and the other half with the SMT device) to minimise the risk of a sequencing effect (Davey et al., 2015).

Measurement and procedure

Registration of the work products. After pupils indicated that they had completed an assessment task, the configuration of the components in their work product was registered by the first author. For the BW task, the configuration of wires and components was drawn, with comments on whether a connection was on metal or on the isolating cable mantle that covers the copper wire. For the SMT task, each side of each bar had a colour code that remained unique even when the bars were in an upside-down position. The camshaft of the device had six cams, which were all wheels with an eccentric axis (see Figure 8). The colour code of the bar placed on each cam was registered. Configurations with bars that were not placed on a single cam were depicted. To allow verification afterwards, pupils' manipulations (i.e., hand movements) were videotaped.

Table 1

Overview of Buzz-Wire Variables

Indicators	Variables		
Components Battery (2 variables); lamp (3 variables); buzzer (2 variables); switch (loop and spiral, 3 variables)	*Connected to another component (battery, lamp and switch)* -1=not; 0: at one connection point 1: at both connection points	*Conduction (lamp, buzzer, switch)* -1: at least one non-conducting connection 0: not determinable 1: all connections by metal	*Circuit (battery, lamp, buzzer, switch)* -1: connected but not in a circuit 0: not connected 1: in a circuit
Wires Two variables with three values for the set of electric wires	*Connection* 1= at least two components connected with a wire 0= no components connected with a wire -1= no wire connected to any component		*Circuit* 1 = all wires are part of an electric circuit 0 = indefinite: no wires connected or almost all wires (>70%) are part of a circuit -1 = less than 70% of wires part of an electric circuit.
System Three variables, with five, six and three values	*Connections* Max number of interconnected components (values: 0 and 2 to 5)	*Circle* Number of components in a circle (values: 0 to 5)	*Circuit* -1 connection between battery poles without resistance (shortcut) 1 potential difference on all connected components 0 other situations

Table 2

Overview of Stairs Marble Track Variables

Indicators	Variables		
Bars on single cams Six cam positions, each with three variables and each variable with three possible values	*Bar length* 1= correct; 0= not present; -1= incorrect	*Slide rotation* 1= up and correct; -1= up 180° turned; 0= other positions or no bar	*Bar vertical position* 1= vertical slide up; -1= slide downward; 0= no bar
Bars not on single cams (combined)	*Bar position* 1= Vertical up- or downward; 0= not in the frame; -1= horizontally in the frame	*Horizontal solutions* Number of bars that are placed in the frame horizontally (0 to 6)	*Bar manipulations* 1= any combination of bars 0= no combination of bars -1= no change of bar position

Analysis of the work products. Each registration was converted into a record with numeric variable fields. Table 1 provides an overview of the variables of the Buzz-Wire work product. Not all the possible component-variable combinations were used. It was, for instance, not possible to connect clamps to the battery in any other way than on metal; therefore, the number of variables per component varies (as indicated in Table 1). The resulting BW record consisted of 15 variables. For the SMT work product, the bar on each of the six cams was described by three variables each, resulting in 18 variables. Three additional variables described work products of which the bars were not placed on a single cam (see Table 2. Together, the SMT record consisted of 21 variables. Ten BW and ten SMT work products were registered independently by the first author and an independent rater. For both assessment tasks, the interrater reliability (i.e., Cohen's Kappa) was computed. The Kappa scores were high (BW: $K = 0.988$, $p < 0.001$; SMT: $K = 1.000$, $p < 0.001$), indicating that the coding procedures were reliable (Landis & Koch, 1977).

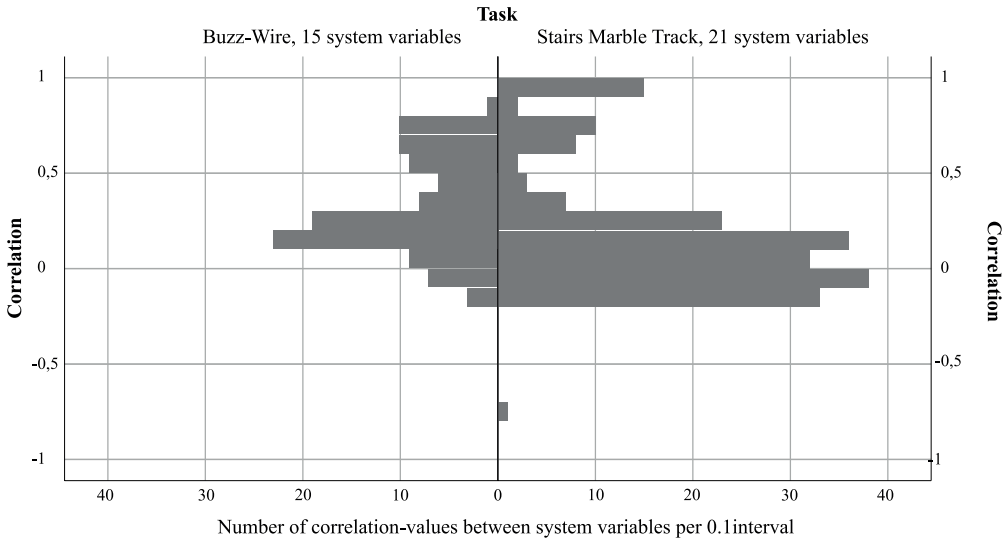
The SPSS aggregate function was used to compute the frequency of the different work products by using the variables of Table 1 (BW) and Table 2 (SMT) as break variables. A wide variety of work products would already indicate that pupils combine the device components in various ways. However, decision 2 implicates that such variation should reside in pupils' use of the opportunities that a task offers to combine its components in multiple ways. To back decision 2, the correlations between the BW (see Table 1) and the SMT (see Table 2) variables should be medium to low. A perfect correlation would indicate that only a single combination is considered. An SPSS bivariate correlation analysis was conducted on all BW and SMT variables. From the lower part of the correlation matrices (see supplementary material), the number of correlations per .1 interval was counted and displayed in a diagram to show.

Results

The 272 participants generated 145 different BW work products and 112 different SMT work products. For the BW device, there were seven pupils (2.60%) who did not make any connection between the components. For the SMT device, there was one pupil who did not combine any bar with another bar or the frame. Correct solutions were provided by 14 pupils (5.10%) on the BW device and 34 pupils (12.50%) on the SMT device. Figure 9 shows that 99% of the 86 BW variable correlations were below *r*=.8 and 65% below *r*=.5, indicating that pupils do combine the components in various ways. From the 209 correlations between the SMT variables, 92% was below *r*=.8 and 82% below *r*=.5. BW correlations above *r*=.50 were found between the circuit variables. This makes sense since the generation of circuits requires specific component combinations. However, even within the circuit variables, different combinations were made, as can be deducted from the fact that none of the BW correlation coefficients was higher than 0.80. Some SMT variables had correlation coefficients near one. These were the variables that indicated the vertical orientation of a bar for each cam. Although possible, not a single pupil put a bar upside down beside one right side up. This implies that for the SMT, pupils' choices on the six vertical-position variables are, in fact, represented by one variable accounting for the vertical position of all bars in the frame, which reduces the combinatory potential of the SMT to 16 variables. Except for this variable, pupils did combine all components of the SMT in several ways.

Figure 9

Left side: Distribution of correlations between the 15 BW variables described in Table 1. Right side: Distribution of correlations between the 21 SMT variables described in Table 2.



Conclusion

The results show that both assessment tasks (BW and SMT device) facilitated pupils to combine the device components in various ways and, consequently, generated a wide variety of work products. The variety for both assessment tasks represented different solutions, which differed from no change to the initial configuration (i.e., loose components) to the correct configuration. All in all, this offers an indication that tasks enable pupils to demonstrate their understanding of the tasks' system in various ways.

STUDY 2: VALIDATING THE SUITABILITY OF GENERIC SCORING RULES

Participants and design

The requirement that work products can be reliably ordered on a single dimension was explored by asking technology-education experts to compare work products on their perceived quality, i.e., which product displays the most aspects of a functioning device. The experts were invited by e-mail and phone by the first author. Nine out of fifteen were able to participate.

The second requirement for decision four was that task-specific scoring rules should comply with the Fischer scale. Researchers, known from their publications based on the Fischer scale, were invited by e-mail, ResearchGate and LinkedIn to react to the application of the Fischer scale in this study. Six researchers could participate during the timeframe of the data collection.

The third requirement was that the results from the scoring rules should match an independent judgement about the quality of the work products. This requirement was checked by comparing the levels resulting from the scoring rules with the ranking value of the same work products based on the independent judgements of the technology education experts.

The fourth requirement was that the levels generated by the scoring rules should reliably reflect pupils' level of prior knowledge. This condition was checked by examining the psychometric properties of the tool.

For this study, all technology education and Fischer experts were informed about a) the nature of the intervention, b) the data collection, data handling and data storage procedure, and c) the report in advance. All participating experts agreed by signing the informed consent form.

Measurement and procedure

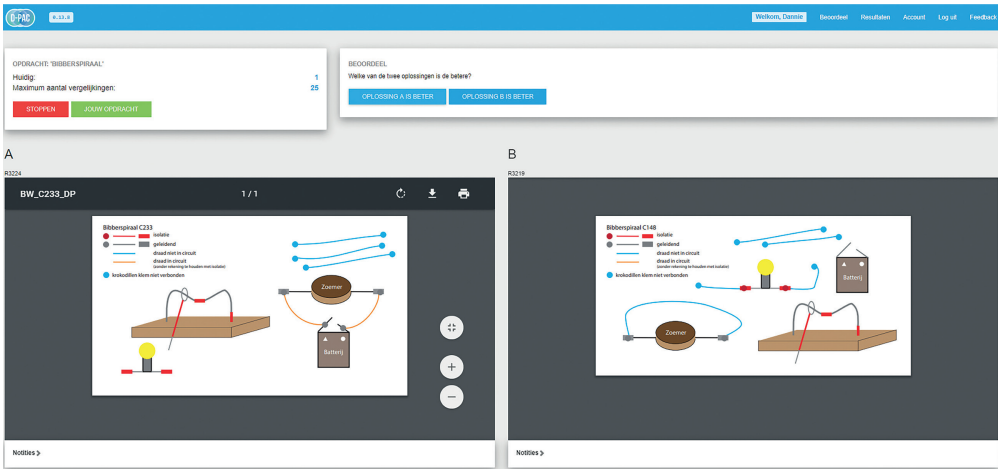
To explore the first requirement, the nine participating technology education experts compared 25 pairs of work products in terms of device functionality

utilising the Digital Platform for Assessment of Competences tool (D-PAC; Verhavert et al., 2019). The pairs were randomly chosen by the D-PAC tool from 19 BW and 17 SMT work products that were selected by the first author based on the criteria that they a) frequently occurred and b) ranged in terms of how many device components were (correctly) connected. The BW work products were represented by a schematic drawing, and the SMT work products with a photo. Per pair, the experts had to select the work product which, in their opinion, represented the best functionality of the device (see Figure 10).

The D-PAC tool uses the Bradley-Terry-Luce model to compute an overarching rank value and the 95% CI of its standard error per work product. This ranking value is used to establish a general rank order of the work products for both assessment tasks. D-PAC automatically computes the Scale Separation Reliability, which represents the interrater reliability between the experts (Verhavert, 2018). A high SSR-value indicates that the work products were rank-ordered in a reliable manner.

Figure 10

Screenshot D-PAC tool (Verhavert, 2018)



The second requirement that task-specific scoring rules should comply with the original Fischer scale was checked by consulting the researchers. First, they were asked to provide a written response to identify the similarities and the differences in their opinion about the levels of work products and, thereafter, a semi-structured interview (about 60 minutes) at their office to discuss the work products that were categorised differently, in order to sharpen the arguments for the final scoring rules. Due to availability during the time frame of this study, only four of the six experts could be interviewed. For the written response, a sample of work products (nine BW and 10 SMT tasks) was selected by the first author based

on the criteria that the work products a) were also used in the previous rank order study and b) represented all Fischer scale levels, as initially coded by the first author. The experts were asked to label the work products (BW and SMT) based on the Fischer level they believed best-represented pupils' level of understanding of the associated system. In addition, they were asked to substantiate their label by their knowledge of the Fischer scale. The Fischer experts received a brief instruction about the study design and a word file containing the 19 work products and textboxes to fill in the Fischer-level coding labels and their substantiations.

Unfortunately, only three Fisher experts provided a written response for the SMT device, and no written responses were received for the BW device. To (partly) remedy this, the semi-structured interviews started with the replication of the selected BW and SMT work products with the original materials. For each work product, the experts were asked to think aloud about the Fischer scale level label they believed was appropriate. During the interviews, arguments obtained from the written responses (i.e., SMT device) were put forward by the first author in case this provided another perspective on the matter. The think-aloud data was collected by audio-taping the semi-structured interviews. Finally, the arguments provided by the Fischer experts were used to revise the BW and SMT work-product scoring rules.

The third requirement was explored by correlating the work products' rating value, resulting from the ranking by the technology education experts, with their Fischer scale level, resulting from the scoring rules. The Intraclass Correlation Coefficient (ICC) was calculated to indicate the level of agreement between both approaches.

The psychometric properties of the tasks were explored in two ways. The test-retest reliability was explored by retesting eleven randomly selected pupils after six weeks. The ICC was calculated to quantify the relationship between the test and retest scores. The approach proposed by Hessen (2011) was used to establish the parameters of the extended Rasch model for the BW and SMT data and check the goodness of fit of this model using a Likelihood-Ratio (LR) test. For this, the scoring rules were considered as dichotomous items (e.g., was anything changed from the start, were all connections on metal). Whether these items were 'answered' correctly or not (i.e., whether a particular combination was present or not) was calculated from the BW or SMT variables (see Tables 1 and 2). For both sets of items, the SPSS aggregate function was used to create an extended Rasch model table of which the subsequent higher-order interactions were the coefficients of the covariates given by $yr=t(t-1)\ldots(t-r+1)/r!$, for $r=2\ldots a$, where a is the order of the highest constant interaction. Y2 being the first higher-order interaction, y3 the second etc. and t the sum score of each pattern of results. In SPSS GLM, the parameters of the extended non-parametric model

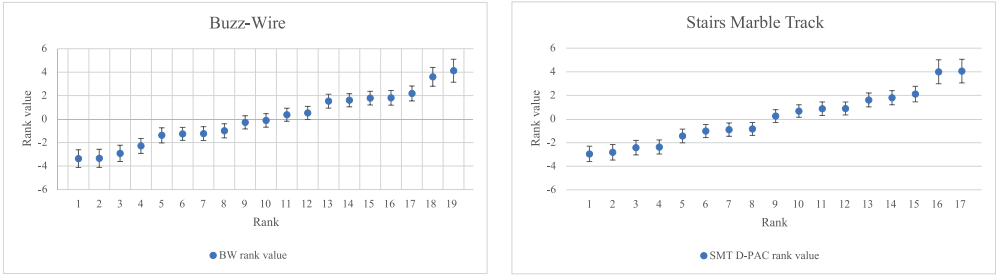
were analysed with a log-linear model and for an increasing number of constant higher-order parameters, of which the likelihood ratio was tested.

Results

Rank order of work products by technology education experts. The D-PAC tool provided an overview of the rank order per assessment task (see Figure 11). The order by which the work products are ranked on the X-axis is determined by their ranking value as depicted on the Y-axis. The whiskers show the 95% standard error of this ranking value. More than their rank, the ranking value of the work products and their 95% CI provides an indication of perceived difference. An overlap in the 95%CI implies that the experts do not consequently indicate one of these work products as the more functional one (e.g., the BW work products that are ranked at positions 5 to 8). No overlap between the 95% CI indicates that most or all experts consequently judge one work product as the better one (e.g., the SMT work products ranked at positions 8 and 9). This lack of overlap points to a clear difference in the perceived functionality of these work products. The SSR value was 0.91 for both tasks, indicating a high level of agreement between the experts (Verhavert et al., 2019).

Figure 11

D-PAC ranking of BW and SMT work products



Work-product levels by the Fischer experts. The comparison between the initial Fischer scale levels, as labelled by the first author using the initial scoring rules (see Supplementary materials, chapter 2, initial scoring rules), and the levels reported by the experts in their written responses (SMT device) is presented in Table 3. Although there are differences, this overview indicates agreement about which work products should be categorised at a higher level. Noteworthy here is that two of the experts used the highest level of understanding (i.e., Fischer level 7), while this was not included in the initial coding by the first author. So, there seems to be some disagreement about which level of understanding should be attributed to the highest quality work product (i.e., correct configuration).

Table 3

Overview of initial and expert scoring for SMT device-related work products

Case (SMT)	Initial scoring rules (supplementary material)	Expert 1	Expert 2	Expert 3
138	**1**	**1**	**1**	**1**
204	**2**	**2**	3	
184	**2**	**2**	3	**2**
147	**2**	**2**	**2**	
008	3	4	5	2
080	4	4	5	5
172	4	**5**	**5**	
100	**5**	**5**	**5**	4
127	**5**	**5**	**5**	
033	6	**7**	**7**	4

Interviews. The replication of the BW work products stimulated the Fischer experts to think and argue about the necessity of abstract reasoning for generating a fully functional device. For example, expert 5 stated: "Initially, I thought it (the correct SMT) should be level 7 because it requires the combined use of many representations, which is a complex skill; however, it does not require the application of a general physical law like for the correct BW solution." Other than for verbal accounts, in which level 3 can be distinguished from level 2 by the expression of a visible causal relationship, it was not possible to construct a comparable argument to distinguish level 2 and level 3 work products, as combining more than two components by their physical properties may be considered as a repetition of manipulations at level 2.

Based on the discussions with the Fischer experts and the suggestions they provided (e.g., expert 1: apply more formal scoring rules), the initial generic score rules (see supplementary material) were refined by the first author. The final heuristic (see Table 4) is based on downward reasoning, taking the correct solution as the starting point. If the work product does not meet those demands, the rules of the preceding, lower level should be considered. This approach has the advantage that the description can be limited to the essential difference with the preceding level and, consequently, the description of a level cannot be applied in isolation.

Table 4

Scoring rules for inferring pupils' understanding of two material-based systems. Text in grey represents the general rule (adapted from Van der Steen, 2014)

Level	Buzz-Wire	Stairs Marble Track
Single abstractions (Rp4/Sa1) An abstraction is used to realise the solution.		
7	Correct solution. Loop and spiral connected as a switch *and* correct connection of the battery *and* all connections on metal *and* all components will function when the circuit closes (no short circuit). If not: go to Rp3	Not used. The SMT task does not require abstract knowledge.
Representational system (Rp3) Relationships have been established between all components of the system.		
6	The work product demands the combination of multiple representations • An electric circuit in which both lamp and buzzer function by default *and* all connections on metal OR • Loop and spiral connected as a switch that sets a lamp or buzzer on/off, *disregarding* whether all connections are on metal If not: go to Rp2	The correct configuration. It demands the combination of all representations. • All bars are ordered according to their length. • The orientation of the slanted bar tops potentially allows a marble to roll onto the slanted top of an adjacent bar. • The slope of all slanted tops will cause the marble to roll down in the direction of the high roll-off point. • A correct estimate of the effect of a turning camshaft on the movement and height of adjacent bars.
Representational mappings (Rp2) Causal relationships with an intermediate step, linking single causal relations.		
5	The outcome contains a mapping: • A connected lamp or buzzer will function in an electric circuit, disregarding whether all connections are on metal. OR • All connections on metal, including both connection points of the lamp *and* the spiral and loop, are linked through a connection by one or more other components. If not: go to Rp1	The outcome contains a mapping: • All bars are in the correct order, *and* the direction of all slides at the bar top allows a marble to role upon the slide of an adjacent bar (any correct combination of correct or 180°-rotated slide positions) OR • The slope of at least five tops will cause the marble to roll down in the direction of the high roll-off point (but bars not in the correct order or one bar incorrect).

Single representations (Sm4/Rp1) A single representation (mental coordination of two or more sensory-motor systems) is part of the action. Single causal relationships.		
4	Use of a single representation. • There is a connection from one pole of the battery to the other pole by at least one other component (lamp, buzzer, loop, spiral). • Both poles of the lamp or buzzer connected to the battery. • All connections should be on metal, conducting electricity. Both poles of the lamp should be connected in this way • The ring and spiral are linked. Not directly but via at least one other component. If not: go to Sm3	Use of a single representation • All bars in the frame are ordered by their length at their correct position in the frame. OR • The direction of all slides in the frame potentially allows a marble to roll upon the slide of an adjacent bar. (any combination of correct or 180°-rotated slide positions)
Sensorimotor – system (Sm3) Observable causal relationships. A manipulation is linked to an observable consequence.		
3	All components are connected, treating the loop and spiral as a single component. If not: go to Sm2	Bars positioned to fill the gap in the frame between the roll-on and roll-off point, • There are bars vertically positioned in the frame with the slanted tops upward (but bars missing or at least one slanted top rotated by 90° or 270°). OR • There are at least five bars in the frame (filling up the space between the low roll-on and high roll-off point) but not with the slanted top upward (vertically top-down or horizontally)
Sensorimotor mapping (Sm2) Combining features of two objects.		
2	Any connection between two components with a wire. If not: go to Sm1	Combinations of single properties of bars and frame (single or repeated) • At least one combination [mapping] of a single property (e.g. length or top-shape) of two or more bars OR • At least one combination of a single property of a bar and a property of the frame (wheel-support, length of the gap between roll-on and roll-off point)
Single sensorimotor actions (Sm1) Use of a single feature of object or task. Observable.		
1	The work product has a connection but does not include a connection between two components by a wire. If not: no action = 0	Something has been changed, but the work product does not include a combination of bar or bar and frame features.

Alignment rank order technology education and Fischer level experts. With the aim to provide additional support for design decision 4, it was examined whether the task-specific rank values of the work products (i.e., device operationality) resulting from D-PAC aligned with the general levels of understanding as determined by the application of the refined scoring rules. There was a high and significant correlation between the Fischer scale level and the rating value of 19 BW work products (ICC=.875, p<.001, 95%CI[.704,.950]) and 17 SMT work products (ICC=.843, p<.001, 95%CI[.618,.940]).

Psychometric properties: test-retest reliability. Retesting a random sample of 11 pupils after six weeks showed a test-retest ICC (two-way mixed, absolute agreement) of .813, p=.002 for the BW task and .920, p<.001 for the SMT task.

Psychometric properties: One-dimensional Rasch model fit. Based on the scoring rules, 13 SMT and 12 BW items were constructed, each item indicating whether a particular kind of combination between the systems' components was present or absent in the work product. A good fit with the extended non-parametric Rasch model was found for the BW task when four constant higher-order interaction variables were added as covariates (LR-test: X^2=7.117, *df*=10, *p*=.71). For the SMT task, a good model fit was found when three constant higher-order interactions were added to the model (X^2=6.127, *df*=3, *p*=.11). (see supplementary material).

Conclusion

The results show that technology education experts were able to rank order the pupil-generated work products for the BW and SMT device in a quite similar and, thus, reliable manner. This provides an indication for the claim that the variety of work products can be rank-ordered in terms of the quality of their construction (i.e., device functionality).

The Fischer experts - after intensive labelling and discussions - provided concrete suggestions for refining the initially developed generic scoring rules. The levels resulting from these scoring rules showed a significant positive and high correlation with the rank orders provided by the technology-education experts. It, therefore, may be concluded that it is possible to indicate differences in pupils' ability to reconstruct a specific system with scoring rules that are based on a generic developmental model.

The test-retest reliability score suggests that the level resulting from the task relates to a person's level of prior knowledge. The goodness of fit of the extended Rasch model indicates that the items deduced from the work products and scoring rules relate to differences in a single latent variable, presumably pupils' prior knowledge about the tasks' system. Together, these results support decision four, using scoring rules based on a generic model to identify differences in pupils' prior knowledge about a specific system by a single-task work product.

STUDY 3: VALIDATING ABSENCE CONSTRUCT-IRRELEVANCE

Participants and design

The third step in the development and validation of the generic diagnostic assessment was to examine whether the levels inferred from pupils' work products matched the range expected from the student model. Furthermore, it was examined whether these results had an added value on top of already utilised tools. To this end, additional data (i.e., math and reading ability scores) were collected from the 272 pupils that generated the work products (see Study 1). Following privacy and data protection regulations, untraceable pupil identifiers were used at the school level to relate standardised math and reading ability test scores to the BW and SMT measures.

Measurement and procedure

Four different measures were used, namely the Fischer level scores for the BW and SMT device and the standardised test scores for reading and math ability. Pupils' *level of understanding* for both tasks was measured by scoring their work products based on the refined generic scoring rules (see Table 4). An automated SQL query was used for this. Pupils' *math and reading ability* are tested twice a year at their primary school. The scores are used to monitor a pupil's progress relative to previous test scores and relative to the average progression of other pupils (Feenstra et al., 2010; Janssen et al, 2010; Tomesen & Weekers, 2012). A yearly updated indication of the mean score is that each test is published on the test providers' website (CITO.nl). The math and reading comprehension tests were administered three months before or after the BW, and SMT tasks were administered. Due to absence during the test administration, scores for all four measures were available for 256 out of 272 pupils.

To examine the predictive value of mathematics and reading comprehension abilities on task performance and the predictive value of task performance on another system, two regression analyses were conducted with respectively the Fischer scale level score for the BW and SMT device as outcome variables. Predictors were pupils' reading comprehension and math ability scores, their SMT level for the BW outcome and their BW level for the SMT outcome. The regression analyses were conducted in a stepwise manner in order to establish the additive effect of each predictor. The assumption check (i.e., linearity, multivariate normality, multicollinearity, and homoscedasticity) showed a slightly skewed distribution of pupils' reading ability scores and their SMT Fischer level scores. To minimise this potential bias, the bootstrapping option with 1,000 iterations was used (Wu, 1986).

Results

Pupils' Fischer scale level scores for both devices are represented in Table 5 and show that pupils differed considerably in their understanding of the device's system. The cumulative percentage shows that the majority of the pupils did not reach Fischer level 5 (Representations, mapping) for both devices (SMT, 59.2; BW, 61.4). Pupils' average scores for math and reading ability are represented in Table 6 and show that their average scores are slightly higher than the national reference values.

Table 5

Distribution of Fischer-scale levels on the BW and SMT task (n=272)

Level	Buzz-wire			Stairs Marble Track		
	N	%	Cum %	N	%	Cum %
1 (Sm1)	28	10.3	9.9	20	7.4	7.4
2 (Sm2)	38	14.0	24.2	19	7.0	14.4
3 (Sm3)	35	12.9	37.1	48	17.6	32.0
4 (Rp1)	60	22.1	59.2	80	29.4	61.4
5 (Rp2)	65	23.9	83.1	71	26.1	87.5
6 (Rp3)	32	11.8	94.9	34	12.5	100.0
7 (Ab1)	14	5.1	100.0	-	-	-

Table 6

Reading comprehension and mathematics test scores

Assessment	Test grade	N	Reference[a]	Mean of participants		*sd*	Min	Max
Reading comprehension	4	23	32-33	35	[28,42]	18	7	98
	5	174	44-46	48	[46.51]	15	14	98
	6	67	55-61	60	[57,65]	17	29	119
Mathematics	4	23	86-87	92	[84,100]	20	55	164
	5	176	100-100	102	[100,104]	12	70	144
	6	62	112-112	116	[114,119]	11	86	154

Note. [a] Reference values for the mean of 2016 and 2017 (source: CITO.nl).
95%CaCI of participants' mean between brackets.

The regression analyses (see Table 7) for the BW device showed that pupils' Fischer level on the SMT task was the strongest predictor, accounting for 6.70% of the variance in BW Fischer level. Adding reading comprehension caused a significant but small (1.80%) improvement of the model. Adding math ability scores did not significantly improve the model. The regression analyses for the SMT device showed that pupils' Fischer level on the BW task was the strongest predictor, accounting for 6.70% of the variance in the SMT Fischer level. Adding mathematic ability scores caused a significant improvement of the model with 4.00%. Adding the reading ability score did not significantly improve the model.

Table 7

Multiple regression analysis, predictors of BW and SMT level

	Predictors	*b* [a]	BCa95%CI	*β* [a]	T [a]	Sig [a]	Fchange[b]	Sig. Fchange[2)]	Adj R^{2} [b)]
BW level	SMT level	.30	[.17, .43]	.25	4.03	<.001**	19.2	<.001**	.067
	Read comp	.020	[.01,.03]	.21	2.54	.012*	6.2	.012*	.085
	Math	-.01	[-.03, .01]	-.09	-1.04	.302	1.1	.302	.085
SMT level	BW level	.20	[.10,.30]	.24	4.03	<.001**	19.2	<.001**	.067
	Math	.021	[.01,.04]	.23	2.81	.005**	12.4	.001**	.107
	Read comp	-.00	[-.02,.01]	.-.03	-.380	.704	.14	.704	.104

Note. [a] Values of the final model with three predictors, with BCa95%CI.
[b] Value after predictor was added to the model
* significant correlation at the 0.05 level, **at the 0.01 level, ***at the 0.001 level (2-tailed)

Conclusion

The results show a low predictive value of both pupils' reading and mathematics ability scores on their obtained Fischer-level scores, meaning that math and reading ability tests should not be regarded as suitable alternatives for the generic diagnostic tool. Although pupils' Fischer level for the SMT device was predictive of their level on the BW device (and vice versa), it accounts for a very small part of the differences.

DISCUSSION

Findings

This study aimed at developing and validating a generic diagnostic tool for assessing primary education pupils' prior knowledge of technological systems. To this end, the Evidence Centered Design approach (Mislevy et al., 2003; Oliveri et al., 2019) was utilised. To properly validate the development of the diagnostic assessment tool, the design decisions (i.e., warrants) should be substantiated with theoretical as well as empirical evidence (i.e., backings).

Study 1 addressed the decisions related to the design of the assessment tasks based on an electrical (i.e., BW device) and a mechanical (i.e., SMT device) system. More specifically, it examined whether pupils could combine the system's components in various ways, allowing them to demonstrate partial knowledge and, by that generating a wide variety of work products. To this end, primary education pupils carried out both assessment tasks, which generated 272 individual work products per device. Results for both devices indicate that pupils interrelated the device's components in various ways, resulting in 145 different BW and 112 different SMT work products. This demonstrates that both tasks

allowed pupils to apply various aspects of knowledge about the interrelationship of the devices' components. All in all, these empirical findings corroborate the theoretical backings. More specifically, the assessment tasks enabled pupils to generate the necessary variety of work products (Davey et al., 2015) despite the restrictions in experiencing the consequences of trial-and-error behaviour (Klahr & Robinson, 1981; Philpot et al., 2017) and allowed them to make their tacit knowledge explicit (Levy, 2012; Zuzovsky, 1999).

Study 2 addressed the design decision that generic scoring rules can be utilised to infer pupils' prior knowledge about material-based systems from their generated work products. Since pupils' prior knowledge may differ considerably per device (Molnár et al., 2013; CITO, 2015), the theoretical backings favoured the development and utilisation of generic - device transcending - scoring rules (Nitko, 1996). Based on Fischer and Bidell's dynamic skill development framework (2007), seven generic levels were operationalised in level-specific scoring rules (see Supplementary materials, chapter 2, scoring rules). To examine the suitability of the generic scoring rules, two different types of expert groups were asked to qualify a representative sample of work products. Experts in the field of technology education (N = 9) rank-ordered, based on pair-wise comparisons, a representative selection of work products in terms of the quality of the construction (i.e., device functionality). Researchers in the field of dynamic skill development (N = 6) interpreted and substantiated the level of work products based on their experience with Fischer's framework on dynamic skill development. The semi-structured interviews yielded valuable insights and concrete suggestions, which were used to calibrate the task-specific scoring rules according to the principles of the generic scale for refining the generic scoring rules (see Table 4). After utilising the refined scoring rules, results for both devices show a significant positive and high correlation with the ranking value that resulted from the independent judgements of technology education experts. The correlation between test and retest scores was high. Pupils' results on items indicating specific combinations of components did fit an extended non-parametric Rasch model. All in all, these empirical findings align with the theoretical backings. Meaning that the diagnostic tool assesses construct relevant (i.e., prior knowledge about material-based systems) pupils' characteristics (Kane, 2004; Oliveri et al., 2019).

Study 3 addressed whether the tasks did indeed generate the differences in skill levels that were expected regarding the age of the pupils. For that, the generated work products from study 1 were scored according to the refined generic scoring rules. The outcomes were in accordance with the distribution of levels that was expected from previous studies with the Fischer-scale (Schwartz, 2009). The findings confirm those of studies indicating that pupils in primary

education find it difficult to understand technological systems (Assaraf & Orion, 2010; Ginns et al., 2005; Koski & de Vries, 2013; Svensson et al., 2012). A plausible explanation for this could be that pupils' ability to apply inductive reasoning strategies (i.e., Fischer level 7) is not sufficiently developed yet in primary education (Molnár et al., 2013).

By comparing pupils' levels on the tasks with their scores on reading comprehension and mathematics, it was also explored whether such scores might also be used as an indication of pupils' prior knowledge about the material-based systems. Prior research, for example, indicated that primary education pupils' math and reading ability scores are strong predictors of their academic achievement (Safadi & Yerushalmi, 2014; Wagensveld et al., 2014). To examine this, the levels of the work products from study 1 were related to pupils' math and reading ability scores obtained from National standardised tests. In contrast to the study of Safidi & Yerushalmi (2014), a neglectable effect of math and reading ability scores on task- achievement was obtained. A possible explanation might lie in the nature of the assessment task. Whereas Safi and Yerushalmi assessed pupils' understanding with multiple-choice questions, this study made use of performance assessments. By doing so, pupils were enabled to make use of their tacit knowledge (i.e., design decision 1), which differs from solely enabling pupils to use verbalisations (Cianciolo et al., 2006; Wagner & Sternberg, 1985). All in all, these empirical findings indicate that construct-irrelevance (i.e., assessing unintended/confounding pupil characteristics, see Kane, 2004; Roelofs, 2019) can be excluded.

The finding that the tasks reveal aspects of pupils' prior knowledge, which are not reflected by their scores on mathematics and reading comprehension, strengthens the importance of using such tasks in primary education. On the one hand, it can reveal that, preferably within integrated STEM, engineering activities are necessary to promote pupils' understanding of technological systems. On the other hand, it can also reveal the capacities of certain pupils that remain hidden by the current assessment practice.

Limitations

Although the obtained findings may sound promising, it is important to take the study's limitations into account when generalising their implications to other educational practices and research. A major limitation follows from the tools' purpose: enabling teachers to get information about their pupils' prior knowledge that can help them to prepare their lessons. The design decisions following that purpose limit the tools' application for formative use. By restricting the evaluation of trials, the tool does not enable pupils to show their problem-solving ability, i.e., the ability to infer a system's structure through interaction. See, for

instance, the Pisa 2012 assessment on creative problem-solving for such tasks (OECD, 2014). The use of a generic scale may suggest that the tool measures a generic ability. However, the generic scale only makes it possible to compare a pupil's prior knowledge of different systems. The level resulting from a work product only indicates prior knowledge about the system that the task represents.

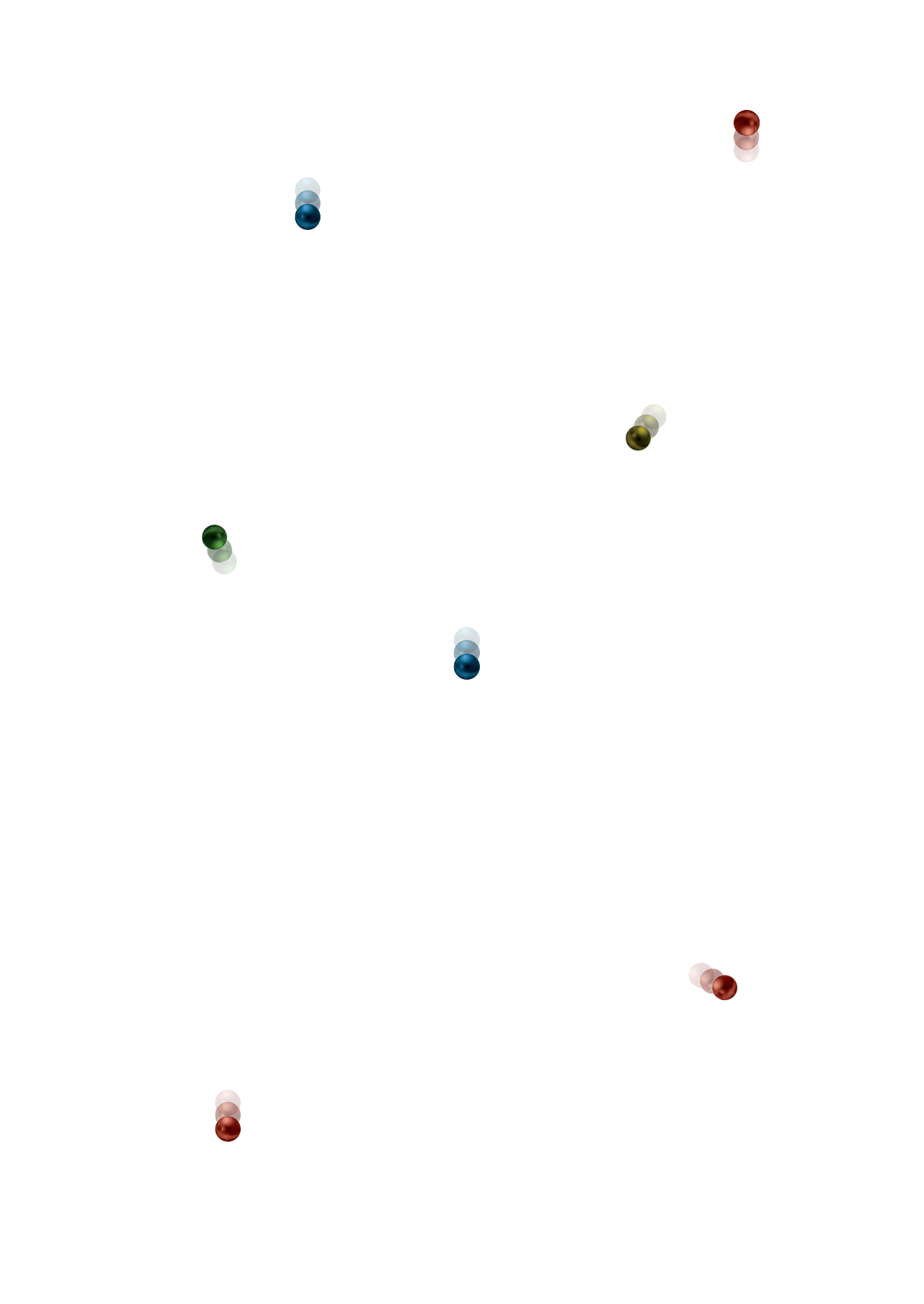
Other limitations reside in the methodology used in this study. First, whereas utilising the ECD validation approach has proven its value, this was - to the best of our knowledge - mainly the case for so-called high-stakes assessments such as standardised tests. Its utilisation for diagnostic assessment purposes is a yet unexplored area, and perhaps other validation approaches might be more suited for this end. To gain a broader perspective on the matter, the reader might, for example, also be interested in utilising design and validation approaches that have a stronger emphasis on formative educational practices (e.g., Black & William, 2018; Pellegrino et al., 2016). Second, as indicated by Study 1, it remains to be seen whether the current study was able to gain insight into the full range of work products pupils might generate. In case the range increases, this might have implications for the generic scoring rules. It, thus, remains to be seen if the current scoring rules are also suitable for a larger variation in generated work products. Third, as indicated by Study 2, the limited number of experts in the field of dynamic skills development indicated they found it difficult to utilise the scoring rules for the BW device. Although, after constructive discussions, an initial agreement about the generic score rules was obtained, further empirical backings (e.g., replication study with other technological devices) are required to substantiate this design decision further. Although the timeframe and availability of the experts did not allow it, it is also preferable to organise (multiple) calibration sessions in which the experts discuss the scoring rules with each other (O'Connell et al., 2016).

In addition, although work products are valuable assessment tasks, it can be questioned whether a full understanding of pupils' mental models (i.e., understanding of concepts and principles) can be inferred from them (Garmine & Pearson, 2006). As indicated by others, one should be aware that every assessment tool (e.g., purpose, task, scoring, outcomes) has its own merits and pitfalls and might want to consider the utilisation of a) multiple assessments with the same tool and b) different types of assessment tools (Gerritsen-van Leeuwenkamp et al., 2017; Van der Schaaf et al., 2019). Lastly, even though pupils from different schools and grade classes participated (Study 1 and Study 3), it remains to be seen if this specific sample properly reflects the entire population. This might have implications for the pupils' characteristics (i.e., math and reading ability) that were included in this study and their effect on pupils' understanding of material-based systems. It might, for example, also be

feasible to assume that a pupil's motivation affects his/her task engagement and academic achievement (Hornstra et al., 2020; Schunk et al., 2012).

Implications for educational practices and research

To conclude, this study's theoretical underpinning and its empirical findings support the validity of the generic diagnostic assessment tool. It is a first step in supporting teachers in assessing primary technology education-related learning outcomes (Garmine & Pearson, 2006) and - hopefully - warranting a more structural embedding of technology education in primary education curricula (Dochy et al., 1996; McFadden & Williams, 2020). As indicated by the study limitations, the diagnostic assessment tool requires more research to validate its utilisation. One potential direction for this could lie in replicating the current study with devices based on the current design decisions but which differ regarding the underlying physical principles. By doing so, future studies could examine whether the current design is robust enough to warrant its utilisation in other contexts. Another potential direction could be that triangulation techniques are utilised to examine whether tools aimed at assessing the same construct (i.e., understanding material-based systems) yield comparable results (Catrysse et al., 2016). More specifically, it would be valuable if pupils' verbalisation of their actions was measured a) during (i.e., think aloud) or after (i.e., stimulated recall) their task performance and related to the scoring of their generated work products. For educational practices, it is important to gain more insight into the tool's ecological validity (Kane, 2004). That is, can primary education teachers actually utilise the diagnostic tool to diagnose and enhance their pupils' prior knowledge about technological systems? Prior research indicates that teachers find it difficult to apply such formative teaching approaches (Heitink et al., 2016). Reasons for this could be that they often lack a) a clear understanding of these approaches (Robinson et al., 2014) and b) concrete - how to - examples indicating how such approaches can be utilised (Box et al., 2015). A potential first direction for addressing is to organise training (Forbes et al., 2015; Lynch et al., 2019) or calibration sessions (O'Connell et al., 2016; Verhavert et al., 2019) in which teachers learn how to utilise the diagnostic tool. When familiar with administering the diagnostic assessment tool and analysing the obtained results, (more) support could be provided regarding the adaptive enhancement of pupils' understanding of technological systems (Black & Wiliam, 2018; Van de Pol et al. 2010).



CHAPTER 3

TEACHER JUDGEMENT ACCURACY OF TECHNICAL ABILITIES IN PRIMARY EDUCATION

This chapter is based on:
Wammes, D., Slof, B., Schot, W., & Kester, L. (2023). Teacher judgement accuracy of technical abilities in primary education. International Journal of Technology and Design Education, 1-24. https://doi.org/10.1007/s10798-022-09734-5

Acknowledgement of author contributions:
DW designed the study, collected the data, planned the data analyses, analysed the data and drafted the manuscript; DW, BS WS and LK contributed to the critical revision of the manuscript; LK and BS supervised the study.

Abstract

Accurate teacher judgements can enhance pupils' learning about science and technology. This study explored primary school teachers' judgements about their pupils' ability to reconstruct an electrical and a mechanical system. The judgement accuracy of most teachers was poor, gender-biased, and underestimation was more common than overestimation. The teachers' gender or self-efficacy beliefs do not seem to affect their judgement accuracy, whereas greater technical knowledge and teaching experience might be beneficial. The teachers' judgements were primarily based on their estimation of pupils' cognitive abilities and learning behaviour, which both had less bearing on pupils' performance than the teachers had expected. Diagnostic tasks for technical abilities, like the ones used in this study, can be used by primary school teachers working with children aged nine and above to calibrate their judgement accuracy and adapt their teaching to their pupils' varying levels of prior knowledge. Pupils' performance on these non-verbal tasks can reveal unexpected abilities.

Introduction

The increasing importance of technology in society led several countries to introduce technology into primary education in the second half of the 20th century (Rasinen, 2003). The most important aim of technology education is to familiarise children with technology, as technology affects many aspects of everyday life (ITEEA, 2007). To this end, it is important to develop knowledge of, interest, and self-confidence in technology among young children. Although there are many opportunities to familiarise young children with technology and arouse their interest in it outside of education, the development of technical knowledge seems to depend strongly on the attention paid to it at school (Baumert et al., 1998; OECD, 2014a).

For the development of technical knowledge, it is important, as with other school subjects, that teachers link their instruction and feedback to their pupils' prior knowledge (Ausubel et al. 1968). This requires a correct assessment of this prior knowledge (Fahrman et al. 2020; Slangen et al., 2011; Van de Pol et al., 2010). In the case of subjects such as language and mathematics, teachers can calibrate their estimates based on test results, which leads to reasonable judgement accuracy (Südkamp et al., 2012). However, in primary education, technical abilities are not systematically tested (Hartell et al., 2015). As a result, primary school teachers, including those with a wealth of experience in teaching technology, are uncertain about the extent to which their instruction and feedback match the prior knowledge of their pupils (Scharten & Kat-de Jong, 2012).

This study explores the merits of using diagnostic tasks in relation to teachers' judgement accuracy regarding their pupils' technical abilities. First, the challenges

around teachers' judgement accuracy of their pupils' technical abilities will be discussed. Second, it is explained how teacher and pupil characteristics might influence teachers' judgements, pupils' performance and the associated judgement accuracy. Finally, teachers' views on the use of diagnostic tasks will be discussed.

Teacher judgement accuracy

The correctness of a teacher's judgement of pupils' abilities is usually defined as the extent to which this judgement correlates with pupils' test results (Südkamp et al., 2012). Therefore, teacher judgment accuracy has mainly been conducted for frequently tested subjects like language and mathematics. The deviations in teachers' judgements appear to result from random variation and systematic over- or underestimation (Timmermans et al., 2015).

In primary education, judgement accuracy for pupils' technical skills is difficult to establish due to a lack of testing. There are indications that teachers in primary education find it difficult to estimate pupils' technical abilities. Jones (1998) observed that in technology lessons, teachers reverted to feedback on areas they were more comfortable with, like cooperation. This might relate to the limited role of technology in the curriculum and the nature of technical knowledge. Many technical skills are largely based on tacit knowledge that is difficult to express in words (Mitcham, 1994). This implies that teachers cannot rely on their questioning to establish pupils' abilities. It also complicates the design of valid tests (CITO, 2016; National Assessment Governing Board, 2013; OECD, 2014b).

Scientific research has proposed tasks that enable pupils to apply their knowledge, including its tacit aspects, in the domain of science and technology to diagnose technical ability (Hast, 2020; Swaak & de Jong, 1996). This study used two tasks about technical systems that allow pupils to interact with the materials in various ways (see Chapter 2). Although these tasks can only capture a limited part of the technical skills spectrum, they provide objective data that gives us the first glimpse of teacher judgement accuracy in this domain. This is the first research question of the present study: How accurate are teacher judgements about pupils' prior knowledge of technical systems compared with the results of diagnostic tasks?

Teacher and learner characteristics

The availability of diagnostic tasks to obtain objective data about pupils' technical skills does not imply that teachers will accurately assess those skills. Even with objective data, teachers differ in their accuracy (Van den Bergh et al., 2010). It is still unclear what underlies these differences. Südkamp et al. (2012) point to factors such as teaching experience, years of exposure to the students rated, age and gender but note that they cannot determine their significance for teacher

judgement accuracy because, in the studies they examined, such data are reported incompletely or differently.

For engineering, it has been reported that teachers' ability to gain insight into their students' technical skills is related to their technical knowledge (Compton & Harwood, 2005; van Niekerk et al., 2010) and self-efficacy beliefs regarding teaching in this domain (Jones & Moreland, 2004). Teaching experience is likely of secondary importance (Nadelson et al., 2013).

Teachers also include student characteristics in their judgements. This can lead to bias if these characteristics exert a different effect on student performance than that expected by the teacher. Pupil characteristics that teachers include in their judgements include motivation (Kaiser et al., 2013), cognitive ability (Dompnier et al., 2006; Hoge & Coladarci, 1989) and the extent to which pupils are encouraged by their parents (De Boer et al., 2010). The ethnicity and socio-economic status of pupils may also play a role in teacher judgement accuracy (Timmermans et al., 2015). Moreover, it is known that primary school teachers systematically have higher expectations of girls than boys (De Boer et al., 2010). This is probably different for technology. In this domain, boys are often assessed more positively than girls (Plumm, 2008). This might be linked to teachers' knowledge of pupils' spatial insight, which they obtain from the mathematic tests. Boys generally show more spatial awareness than girls (Reilly et al., 2017; Wang 2017). Spatial awareness is important in construction and mechanics. It might be that teachers generalise this difference in performance to technology in general. Due to the lack of clarity regarding the significance of teacher and learner characteristics for teacher judgement accuracy, these are examined in this study using the question: How do teacher and learner characteristics relate to judgement accuracy of technical abilities?

Use of diagnostic tasks

The tasks used in this study were developed for formative use in primary schools (Wammes et al., 2022a). Whether primary school teachers will use these tasks depends partly on the effort and time required to use them in the classroom and partly on the value that teachers attribute to the data they produce (Kirton et al., 2007). That value will depend on the insight gained into pupils' technical skills, the possibilities a teacher sees for using that insight to adapt technology education, and the teacher's expectations about the impact of a more tailored approach on developing pupils' technical skills (Praetorius et al., 2017). The possibilities that teachers see for applying the acquired insight depend not only on personal didactic qualities but also on the freedom teachers have to design their lessons within the curriculum (Sach, 2015). The final research question of this study addresses the value of the diagnostic tasks from the teacher's

perspective: How do teachers value the diagnostic tasks for their judgement and teaching practice?

In summary, this study explores the value of diagnostic tasks for teachers' estimates of their pupils' prior knowledge of technical systems resulting in three research questions:

1. How accurate are teacher judgements about pupils' prior knowledge of technical systems compared with the results of diagnostic tasks?
2. How do teacher and learner characteristics relate to judgement accuracy of technical abilities?
3. How do teachers value the diagnostic tasks for their judgement and teaching practice?

METHOD

Participants

Two male and six female teachers and their classes at six primary schools in the Netherlands participated. Four teachers were in their first or second year of teaching, one had worked for five years in primary education, and three teachers had 20 or more years of teaching experience. Six teachers had a single, and two teachers had a mixed-age class. The participating teachers had no specific interest in teaching technology. They were asked to participate by students who did an internship at their school and were supervised by the first author. The eight classes had 87 male and 90 female pupils. Their age ranged from 7 years and six months to 13 years and four months, with an average of 10 and six months. Informed consent was obtained from all participating teachers and the parents of the pupils.

MEASUREMENTS

Pupil performance

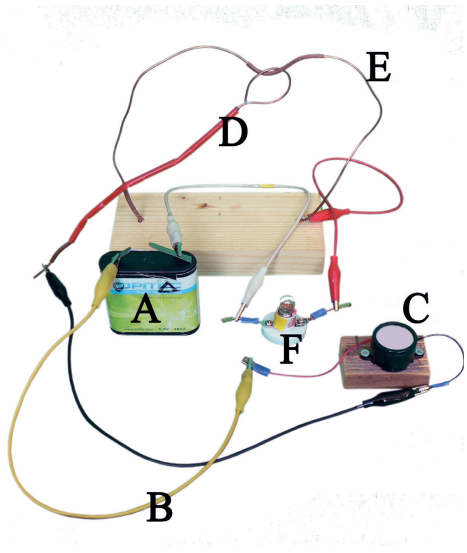
The question, 'How accurate are teachers' judgements about pupils' prior knowledge of technical systems as compared with the results of diagnostic tasks?' was explored by comparing teacher judgements with their pupils' performance on two diagnostic tasks: the Buzz-Wire and the Stairs Marble Track. These tasks are non-verbal, allow pupils to change system variables independently, take only a few minutes to accomplish and allow pupils' knowledge to be assessed on a generic scale (i.e., Fischer, 1980).

The Buzz-Wire (BW) is an electric circuit (see Figure 12). It has a switch made of a copper spiral and a ring with a handle. Touching the spiral with the ring closes the circuit and activates a lamp and buzzer. The Stairs Marble Track (SMT) is a mechanical device that transports marbles upwards to a descending track that brings the marbles back to the start (see Figure 13). The mechanism is a camshaft with six eccentric wheels and six bars of increasing length with a slanted top.

Both tasks are introduced with a one-minute video that shows how the devices should function without revealing their mechanism or construction. When pupils start the Buzz-Wire task, all parts are spread out on the table. The SMT task begins with the six bars scattered randomly on the table. In both cases, the pupils' task is to restore the devices.

Figure 12

Buzz-Wire device



Buzz-Wire. A) Battery, B) Wire with crocodile clips, C) Buzzer, D) Loop, E) Copper spiral, F) Lamp with insulation on the tips of the connectors.

Figure 13

Stairs Marble Track device

Stairs Marble Track. A) Marble (not available in task), B) Camshaft with six eccentric wheels, C) Bar with slanted top, D) Handle (blocked).

The scale used to evaluate the pupils' ability levels was developed by Fischer (1980). The task-specific elaborations (see Supplementary Materials, Chapter 2, Scoring rules) were validated by Wammes et al. (2022a). The Fischer scale describes how skills build up to more complex skills. A pupil's work product has

performance level 1 when it results from a single action based on sensorimotor information. A work product that results from a combination (mapping) of two sensorimotor-based actions indicates a level 2 skill. When the work product results from the combined use of multiple sensorimotor actions, it points to a level 3 (system) skill.

The repeated use of sensorimotor actions results in the ability to remember causal or other relationships between task components. In terms of Fischer: it results in a (causal) relationship *representation*. The tasks restrict the possibility of finding causal or other relationships between task components otherwise, for instance, by trial and error. When a work product reflects the use of a single representation, the pupil demonstrates a level 4 skill. Level 5 skills are evident in work products that require the combination of two representations. At level 6, pupils' work products show that their actions are based on the combined use of multiple representations. At the next level (7), the work products indicate that pupils can apply a generic feature of a phenomenon to solve an unfamiliar problem. The correct completion of the SMT task requires a level 6 skill, while the BW task allows pupils to demonstrate their understanding of electrical circuits up to level 7.

Teacher judgements

Two types of teacher judgement about their pupils' technical abilities were used: a relative and an absolute judgement. For the relative judgement, each teacher ranked the pupils by their presumed technical ability. This was done without further specification of technical ability, which makes this relative judgement, according to Südkamp et al. (2012), an uninformed type of judgement. This ranking procedure was also used to get information about the teachers' parameters for their uninformed judgements; they were all asked to think aloud, and their utterances were recorded while ranking (Loibl et al., 2020). The thinking-aloud recordings were obtained for six of the eight teachers. One teacher provided a general description of her considerations while ranking, and the recording of one teacher failed.

For the absolute judgements, the teachers predicted the level of their pupils' BW and SMT work products. These absolute judgements were an 'informed' judgment: the teachers made their judgments after seeing the tasks and being informed about the Fischer levels for classifying the pupils' work products (Südkamp et al., 2012).

Teacher and learner characteristics

Our second research question was: How do teacher and learner characteristics relate to judgement accuracy of technical abilities? First, the measurement and

scoring of the teacher characteristics and then the measurement of the pupil characteristics will be described.

For all teachers, their gender, teaching experience and the time they had worked with the pupils were recorded. The teachers' technical knowledge was assessed using a multiple-choice test. This Technical Knowledge Test (TKT) comprised 43 items about technology from the 2015, 2016 and 2017 editions of a Dutch national assessment on Science and Technology. This is an admission test for students who want to become primary school teachers. Cronbach's alpha of the TKT was .78.

Teachers' self-efficacy beliefs were administered using an adapted version of the Science Teaching Efficacy Belief Instrument (STEBI-b; Riggs & Enochs, 1990; Bleicher, 2004). The references to science in the statements were changed to references to technology. For example, "I will continually find better ways to teach science" was changed to "I will continually find better ways to teach technology". The instrument combines the Personal Teaching Efficacy scale (PTE, 13 items) and the Teaching Outcome Expectancy scale (TOE, ten items). The PTE is about personal efficacy beliefs, which relate to teacher judgment accuracy (Nadelson et al., 2013). The TOE refers to ideas about the effectiveness of science, technology, and education in general. The STEBI-b was fully administered in this study, but only the PTE score was used as that scale indicates the teachers' self-efficacy. Cronbach's alpha of the PTE was .72.

The learner characteristics measured were pupils' age, gender, and scores for reading comprehension and mathematics as an indication of their cognitive abilities. The participating schools administered the same reading comprehension and mathematics tests (CITO Assessment Institute, 2021).

Teachers' evaluation of the diagnostic tasks

To answer the third research question, 'How do teachers value the data from the diagnostic tasks for their judgment and teaching practice?', each teacher was interviewed about their experience with the diagnostic tasks. Results were reported beforehand. The interview consisted of three open and three multiple-answer questions. The open questions were: What did you notice when you compared the results of the children on the tests with the ranking you have made? Are there any other outcomes that stand out when you compare your predictions for the specific tasks with the students' results? How do the outcomes of the tasks contribute to the image you have of your students? The three multiple-answer questions were about the balance between the time and effort required to use the diagnostic tasks and their benefit, the intended use of such tasks when available, and how the teacher would use the information from such tasks.

Procedure

The teachers started with the TKT and the adapted STEBI-b questionnaire. Then, they ranked their pupils' technical abilities while their explanations were recorded. This resulted in a motivated, relative judgement for each pupil. After being informed about the diagnostic tasks and the associated Fischer levels, they predicted pupils' performance level for each task, which resulted in the teachers' absolute judgement for the Buzz-Wire task (judgementBW) and their absolute judgement for the Stairs Marble Track task (judgementSMT). Next, the teachers introduced the diagnostic tasks to their pupils in a classroom setting using PowerPoint. The pupils performed the diagnostic tasks individually without being disturbed. They were urged not to discuss the tasks with their classmates during the test-taking period. After finishing a task, a picture was made of the pupils' work product. In lower grades, this was done by a teaching assistant and in higher grades, by the pupils themselves. The teachers emailed these photos to the first author. Within a month, each teacher received a report about the performance of their pupils related to the relative and absolute judgements and pupils' reading comprehension and mathematical abilities. Finally, each teacher was interviewed, which took about 30 minutes.

SCORING AND ANALYSES

Teacher judgement accuracy

The performanceBW and performanceSMT of pupils' work products were determined using the scoring rules (see Supplementary Materials, Chapter 2, Scoring rules 1). A second rater independently assessed the work products of the pupils of one school. Inter-rater agreement was calculated in SPSS using the two-way mixed model of the Intraclass Correlation Coefficient (ICC) with the absolute agreement option. For 33 BW work products, the ICC was .85, and for 38 SMT work products, the ICC was .96, which is good.

SPSS two-way mixed ICC with the absolute agreement option was used to ascertain teacher accuracy, as this coefficient, unlike Pearson *r*, indicates differences in systematic deviations. For relative judgment accuracy, the ICC indicates the correlation between the pupils' rank resulting from their teachers' relative judgment and their rank resulting from their average performance level. All ranks are normalised to account for differences in class size. For the absolute judgement accuracies, the absolute judgementBW of the teacher was correlated with pupils' performanceBW, and the absolute judgementSMT was correlated with the performanceSMT.

The extent to which the teachers' judgment accuracy was biased by a systematic under or overestimation of their pupils' technical ability was explored by counting and tabularising all differences between the absolute judgements and pupils' performance on both tasks. The Paired-Samples T-test, with a Bonferroni correction accounting for two judgements for the same pupil, was used to identify significant systematic deviations between judgement and performance.

Teacher and learner characteristics

For the teacher characteristics, the gender, teaching experience, and the teachers' judgment accuracy were tabularized together with the percentage of correct answers on the TKT and their score on the PTE scale. The PTE score was calculated as in Bleicher (2004). Normalised TKT and PTE scores were used for the analyses. Teachers were coded as experienced when their teaching experience was more than three years, which is considered the time needed to develop basic teaching skills after graduation (van Eijk et al., 2015). The teachers' gender was not included in the analyses as there were only two male teachers who were also inexperienced.

For the learner characteristics, we coded age, gender and pupils' latent scores from the reading comprehension and mathematics tests. The latent scores indicate pupils' abilities regardless of age and associated test version. As the scale used for mathematics differed from the scale used for reading comprehension, normalised Z-scores were used.

There was a strong correlation between teachers' absolute judgements of their pupils' performance on the SMT and the BW (r=.811). Therefore, we decided to use the average teachers' judgements over the two tasks as the dependent variable for analysis in SPSS with a set of two-level hierarchical models that nested learners (level 1) within teachers (level 2). Model 0 was unconditional and used to establish the amount of variance in judgements between teachers and within the teachers' classes. Model 1 added pupils' performance on both tasks as predictors. Model 2 included the other learner characteristics, and Model 3 introduced the teacher characteristics. SPSS hierarchical regression analysis was used for the relative judgements.

To explore whether the various explanations given by the teachers for pupils' rank (relative judgement) were indicative of pupils' performance, a distinction was made between high-ranked, average-ranked, and low-ranked pupils for judgement and performance. High-ranked equals the top 25th percentile, and low-ranked below the 75th percentile range. The remaining pupils were coded as average-ranked. Subsequently, it was calculated which percentage of high-, average- or low-ranked pupils had a prediction of a certain category for judgement and performance.

Teachers' evaluation of the diagnostic tasks

How teachers valued the data from the diagnostic tasks for their judgements and teaching practice was established by summing up their answers to the multiple-answer questions. Answers to the open questions were categorised by open coding. One or two answers per category were selected as an illustrative example.

RESULTS

Teacher judgement accuracy

The ICC values in Table 8 show considerable differences in judgement accuracy. The relative judgements were generally the most accurate, followed by the absolute judgementsSMT. The absolute judgementsBW were the least accurate.

Table 8

Primary school teachers' judgement accuracy for technical ability

	Pupils			Relative judgement accuracy[a]		Absolute judgement accuracy			
Teacher	N (male)	Age				Buzz-Wire		Stairs Marble Track	
		$\bar{x}$	*sd*	ICC	*p*	ICC	*p*	ICC	*p*
1[b]	9 (4)	8.1	.4	.049	.447	.134	.353	-.274	.815
2	23 (13)	8.8	.4	.445*	.017	.426*	.013	.344	.054
3	29 (15)	9.9	.3	.263	.085	.251	.053	.357**	.001
4	25 (11)	10.7	1.0	.649**	<.001	.300	.074	.359*	.038
5	21 (13)	10.8	.6	.446*	.022	-.052	.676	.087	.229
6	17 (6)	11.5	.4	.460*	.032	.058	.413	.481*	.020
7	25 (11)	11.5,	.6	.231	.134	-.070	.646	.442*	.0.13
8	29 (14)	11.6	.4	.522*	.002	.126	.232	.305*	.022
All	178 (90)	10.5	1.2	.401**	<.001	.267**	<.001	.368**	<.001

Note. [a] accuracy= ICC two-way mixed, absolute agreement.
[b] teacher 1 had the youngest pupils and selected only those with high general performance.
*significant correlation at the 0.05 level, **at the 0.01 level, ***at the 0.001 level (2-tailed)

The frequencies of the deviations between judgement and performance in Table 9 reveal that underestimation of pupils' technical abilities occurred more frequently than overestimation. For three teachers, their underestimation was significant.

Table 9

Differences between absolute judgements and task performance

Teacher		Average Fischer scale levels								Deviation frequency					PST judgement - performance	
		abs-j$_{BW}$		perf$_{BW}$		abs-j$_{SMT}$		perf$_{SMT}$		over-estimate			under-estimate			
	N	$\bar{x}$	*sd*	$\bar{x}$	*sd*	$\bar{x}$	*sd*	$\bar{x}$	*sd*	<-1	-1	0	1	>1	*t*	*p*
1	9	4.8	*1.1*	4.2	*1.2*	3.1	*1.7*	4,1	*.6*	3	4	3	2	6	.489	.631
2	23	3.0	*1.7*	3.7	*1.6*	3.6	*1.7*	3.4	*1.4*	7	10	10	7	12	-1.067	.292
3	29	2.6	*1.0*	3.5	*1.4*	2.3	*1.0*	3.3	*1.0*	1	5	14	21	17	5.716**	<.001
4	25	4.6	*2.2*	4.7	*.9*	4.3	*1.9*	4.5	*1.0*	11	8	11	6	14	.462	.646
5	21	2.2	*.9*	4.0	*1.5*	2.5	*.9*	4.0	*1.1*		4	8	7	23	6.529**	<.001
6	17	4.8	*1.2*	4.5	*1.7*	4.2	*1.4*	3.9	*1.0*	9	8	7	5	5	-1.172	.249
7	25	4.5	*1.3*	3.8	*1.4*	4.2	*1.3*	4.1	*1.1*	12	13	13	5	7	-1.756	.085
8	29	4.4	*1.4*	5.1	*1.5*	3.8	*1.2*	4.7	*1.2*	4	8	16	11	19	3.622**	.001
All	178	3.8	*1.7*	4.2	*1.5*	3.5	*1.6*	4.0	*1.2*	47	60	82	64	103	4.800**	<.001

Note. PST = Paired Samples T-Test; Deviation frequency = absolute judgement minus performance level; abs-j.= absolute judgement; perf.= task-performance level
** significant at the 0.001 level (2-tailed).

Teacher and learner characteristics

The second research question explored which teacher and learner characteristics might relate to the differences in judgement accuracy. Table 10 provides an overview of the characteristics of the teachers and their judgement accuracy.

Table 10

Teacher characteristics ordered by relative judgement accuracy[a]

T	G	Exp.	Pm.	TKT	PTE	N	rel-j acc	abs-jBW acc	abs-jSMT acc
1	f	30	13	53%	3.2	9	.049	.134	-.274[b]
7	f	20	4	56%	4.1	25	.231	-.070[b]	.442*
3	m	2	6	74%	4.1	29	.263	.251	.357**
5	f	2	3	58%	2.8	21	.446*	-.052[b]	.087
2	f	1	3	72%	3.9	23	.445*	.426*	.344
6	f	5	11	72%	3.6	17	.460*	.058	.481*
8	m	1	7	93%	4.1	29	.522**	.126	.305*
4	f	30	6-26	63%	3.6	25	.649***	.300	.359*

Note. T = teacher, G = teacher's gender, N = number of pupils,
Exp.= Years of experience as a teacher, Pm.= months of exposure to pupils,
rel-j acc.= relative judgement accuracy, abs-j acc.= absolute judgement accuracy,
[a] ICC two-way mixed, absolute agreement,
[b] negative average covariance, violating reliability model assumptions,
* Correlation significant at the 0.05 level (2-tailed), ** Correlation significant at the 0.01 level (2-tailed), *** Correlation significant at the 0.001 level (2-tailed).

The unconditional model of the multilevel analyses in Table 11 shows, with an ICC of .25, that 25% of the variance in judgements can be attributed to differences between the teachers and 75% to within-class judgement variance. Pupils' performance introduced in model 1 explained 20.7% of the variance, but only the SMT task explained a significant proportion, implicating that the teachers' judgements are especially biased for the BW task. The introduction of the learner characteristics in model 2 reduced the unexplained between-learner variance, increasing the proportion of unexplained between-teacher variance to 34%. Being controlled for performance, model 2 reveals that pupils' gender biases the teachers' judgements. Teacher judgements for boys were higher than for girls. Boys did indeed outperform girls, but only on the SMT task where the mean difference in performance was less than expected (.61 level for performance, .84 for judgements).

Contrary to the teachers' expectations, there was no difference in performance on the BW task. Teacher judgements were also biased by pupils' test scores for reading comprehension and mathematics. Pupils' performance on the tasks was less related to these test scores, as expected by the teachers. The introduction of the teacher characteristics in model 3 reduced the unexplained between-teacher variance to 6% of the remaining 59.4% of unexplained variance. Teaching experience had a significant effect on teacher judgement. It thus might be that the less experienced teachers are, the more likely they are to underestimate the achievements of their students.

The hierarchical regression analyses on the relative judgements showed that introducing pupils' performance rank in model 1 explained 15.9% of the variance. The introduction of the learner characteristics explained a further 16.5%. The R-square change resulting from the introduction of teacher characteristics was 0.8% and non-significant. The final model showed that there was a similar gender bias (b=-.17, p<.001) and reading comprehension (b=.09, p=.002) as for the absolute judgements. In contrast with the absolute judgements, there was no significant bias for mathematics scores (b=.04, p=.194) but a significant bias for age (b=-, 06, p=.007), indicating that older pupils were less present in the higher ranks as judged by their teacher, whereas performance showed a small age-related increase.

Table 11

Multilevel regression models of teacher judgement

	Model 0		Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Student level variables								
Intercept	3.68***	0.29	1.44**	.45	1.83	1.26	3.55**	1.11
Performance BW			.08	.07	.10	.06	.09	.06
Performance SMT			.47***	.09	.25**	.08	.26**	.08
Age					.01	.12	-.07	.10
Gender					.79***	.17	.80***	.17
Reading comprehension ability					.34**	.12	.32**	.12
Mathematical ability					.48***	.12	.45***	.11
Teacher-level variables								
Technical knowledge							.31	.19
Self Efficacy beliefs							-.38	.18
Teaching experience							-1.62**	.34
Teacher-level intercept variance	.60	.34	.41	.24	.54	.32	.07	.07
Learner-level intercept variance	1.82***	.20	1.51***	.16	1.06***	.32	1.07***	.12
ICC	.25		.21		.34		.06	
df change			1		4		3	
χ^2 change			34.58***		65.06***		11.11*	
-2LL	628.684		594.101		529.043		517.938	
AICc	634.822		604.450		548.128		543.852	

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

For one class, the recording failed. Open coding of the thinking aloud recordings of the ranking of the remaining 145 pupils resulted in four categories: Learning behaviour (e.g., concentration, perseverance, posing questions), Cognitive ability (e.g., math scores; remarks like a 'clever' or 'average' pupil), Science and Technology (specific references to interest in science and technology), and Support at home (e.g., ' It's likely that there is no interest for technology at home'). Another feature of the explanations was their value; they were positive, e.g., 'a real go-getter', neutral, e.g., 'average cognitive ability' or negative, e.g., 'not interested in technology'. For 75% of the pupils, explanations were given that addressed more than one category, e.g., 'Not a smart boy per se, but he is really good at analysing

structures (categorised as Cognitive ability - neutral), but from him, I expect high performance in technical subjects' (categorised as Science and Technology - positive)'. Inter-rater agreement for the explanation categories was κ=.726, and for the value ratings (positive, neutral or negative), it was .857.

Table 12 shows that, in line with the bias found in reading comprehension and math scores, teachers tend to overestimate the importance of cognitive ability as an indicator of technical ability. In 60% of the explanations about high-ranked pupils by judgment, the teacher mentioned their high cognitive ability. Among those high-ranked by their performance, there were fewer (39%) pupils for whom a high cognitive ability had been mentioned. For the low-ranked judgements, the teachers mentioned a low cognitive ability for 39% and a high cognitive ability for 8% of the pupils. Contrary to these expectations, a low performance lacked any relationship to cognitive ability, as mentioned by the teachers (see Table 12: Low-rank; av-perf).

Teachers also frequently (37%) referred to positive learning behaviour when explaining high-rank judgements, while negative learning behaviour was seldom (6%) mentioned. However, based on performance, the number of high-ranked pupils with positive or negative learning behaviour explanations was comparable. Furthermore, positive explanations of interest in science and technology were less related to pupils' task performance than the teachers expected.

Table 12

Teachers' explanations by pupils' [predicted] and performance rank

Explanation Category	N [a]	Implication	High-rank [b] [rel-j]	av-perf	Average-rank [rel-j]	av-perf	Low-rank [c] [rel-j]	av-perf
Cognitive ability	61	High	[60%] [d]	39%	[26%]	23%	[8%]	26%
		Average	[3%]	8%	[24%]	21%	[11%]	21%
		Low	[0%]	5%	[3%]	10%	[39%]	21%
Learning behaviour	69	Positive	[37%]	24%	[18%]	21%	[0%]	6%
		Negative	[6%]	21%	[27%]	23%	[50%]	44%
Science and Technology	48	Positive	[37%]	26%	[20%]	22%	[8%]	15%
		Negative	[0%]	0	[15%]	18%	[14%]	9%
Support at home for technology	13	Positive	[14%]	5%	[4%]	7%	[3%]	6%
		Negative	[0%]	3%	[2%]	1%	[6%]	6%
Pupils [e]			[35]	34	[74]	73	[36]	38

Note. rel-j = rank by relative judgement, av-perf = rank by average task performance level
[a] Number of pupils with explanations of this category, [b] within the top 25th percentile range, [c] below the 75th percentile range.
[d] Percentage of pupils of this rank for whom their teachers expressed this type of explanation.
[e] Statements about 75% of the pupils included explanations from more than one category, causing the sum of column percentages to exceed 100%.

Teachers' evaluation of the diagnostic tasks

In the interviews, the teachers were especially surprised by pupils with non-expected high task performance. Illustrative is the reaction of teacher 5: "For K30, I thought it would be nothing. This was pure because of the image I have of her. She scored significantly higher than I expected". Another example comes from Teacher 7, "I immediately noticed K24. I had given her the lowest score for the BW task, and then she scored six". Teachers 3 and 8 found it striking that only a few students had lower scores than expected, which corresponds with their significant systematic underestimation shown in Table 9. Some teachers expressed that they were glad that the results confirmed their suspicions. "I knew that K36 was very interested in technology, but I found it difficult to estimate whether he could do it. From the results, I see that he is technically skilled." (teacher 7).
"I am delighted to receive confirmation that my view of the students is broadly correct." (teacher 4). Teachers also related pupils' task performance levels with their reading comprehension and mathematics abilities. Teacher 6 commented: "If you look at K17. Her test scores do not indicate that she is a brilliant student, but she is knowledgeable. She scored highly on both tasks. That certainly says something about her." Teacher 7 wondered why three pupils did the task so well while they had such low mathematics scores.

The teachers with pupils from nine years old were positive about the limited time needed for test-taking. Their pupils did the tests individually in a spare moment. They made the pictures with a tablet and got everything ready for the next one. Teachers with younger pupils were less positive about the effort needed. Their pupils needed the support of a teaching assistant to start the task and make the picture. All teachers reported that their pupils enjoyed test-taking. Seven teachers would use the results to differentiate tasks, instruction and feedback. One teacher found that she had insufficient knowledge and experience to use the results of the diagnostic tasks. Four teachers would communicate the results to pupils and their parents.

DISCUSSION

Teacher judgement accuracy

The participating teachers differed in their judgement accuracy. With one exception, these were below the mean of $r = .64$ reported by Südkamp et al. (2012) for language and mathematics. For three of the eight teachers, there was a significant systematic underestimation of the task performance level of their students.

The fact that teachers are less able to assess technical skills than language and mathematical skills may be related, on the one hand, to the nature of technical knowledge, which includes procedural and visual components that are difficult to identify, and, on the other hand, to the lack of objective data that teachers can use to calibrate their judgements.

The teachers were generally better in their relative than their absolute judgements. This is consistent with Schrader and Helmke's (2001) view that a ranking better reflects the individual teacher's perspective on student performance.

It is striking that pupils' task performance level was mainly underestimated. Studies on teacher judgement accuracy show that teachers usually overestimate the performance of their pupils (Loibl et al., 2020; Praetorius et al., 2013; Urhahne & Wijnia, 2021). Perhaps the limited attention paid to engineering in the curriculum leads to these low expectations.

Teacher and learner characteristics

Of the teacher characteristics, only the relationship between teaching experience and judgement accuracy proved significant. The average estimates of the experienced teachers correlated better with the average performance of their students than the estimates of their less-experienced colleagues. Such difference was not found in teacher judgment accuracy studies in mathematics (Stang, 2016) and foreign language teaching (Zhu & Urhahne, 2015). However, these are domains where teachers know their students' previous test results. The positive effect of teaching experience may be due to increased knowledge about the value of students' cues in relation to their performance in technical activities (Graney, 2008; Thiede et al., 2015). How long the teachers knew their students had no effect; however, this was at least three months for all teachers. The lack of a positive correlation between judgement accuracy and reported PTE scores (i.e., self-efficacy beliefs) seems to contradict the findings of Nadelson (2013) but was also reported by Klug (2016). The positive correlation of judgement accuracy with the TKT scores was not significant but in line with what is known about the importance of teachers' domain-specific knowledge (Compton & Harwood, 2005; Jones & Moreland, 2004; Nitko, 1996; Kramer et al., 2021).

The higher expectations of boys' performance compared to girls were partly confirmed by the better performance of boys on the SMT task. However, they did not correspond to the equal performance of boys and girls on the BW task. The higher estimation of boys' technical ability has been described before (Buccheri et al., 2011; Seiter, 2009). This might be domain-specific since primary school teachers usually estimate girls' skill levels to be higher (Timmermans et al., 2015). The better performance of boys on the SMT task may be related to the spatiotemporal skill that this task requires. Tasks that require Spatio-temporal

skills are known to be performed better by boys than by girls (Reilly et al., 2017, Wang, 2017).

The correlation of teacher judgements with pupils' proficiency scores in reading comprehension and mathematics, and the absence of such a correlation of these proficiency scores with pupils' task performance, resulted in a bias in judgement accuracy. The lack of such a correlation is in line with the estimates made by Wammes et al. (2022a), who reported that scores on reading comprehension and mathematics explained no more than 10% of the variance in pupils' scores on the BW and SMT task.

Remarkably, only reading comprehension explained a significant part of the variance in relative judgements and teachers' judgement accuracy. In contrast, mathematics proficiency scores were a better predictor of the variance of absolute judgements. That teacher judgements correlate with reading comprehension and mathematics skills is not surprising, as these are seen as good predictors of students' academic ability (Timmermans et al., 2015).

The reasons given by teachers for ranking pupils highly were mainly their strong cognitive performance or their positive learning behaviour, such as showing interest and perseverance. Low rankings were explained by weak cognitive performance and negative learning behaviours, such as poor concentration or giving up quickly. Only a quarter of the arguments mentioned pupils' relationship with technology, and 7% of the arguments related to the pupils' home situation.

The emphasis on cognitive ability and learning behaviour is not surprising since teachers rely heavily on these factors when assessing the academic abilities of their pupils. In the absence of specific knowledge about students' technical abilities, these considerations come to the fore for teachers (Dompnier et al., 2006).

High expectations of pupils were frequently supported by remarks about the positive learning behaviour of the pupils and rarely by remarks about negative learning behaviour. In contrast, strong performance did not show such a relationship. Further, pupils' cognitive ability seems to be less decisive for high performance than expected by teachers. Our results do not seem consistent with Bates and Nettelbeck (2001) and Feinberg and Shapiro (2003), who found that teachers' estimates of high-achieving students are the most accurate. This might relate to the important role of tacit knowledge in technical skills. This knowledge is not easily recognisable for teachers and does not play a role in their judgement, whereas it is reflected in pupils' performance on the tasks. The fact that the gender of the pupil was mentioned only once may indicate an explanation bias (Oort et al., 2009). Teachers may not be aware of the role that gender plays in their estimation of the technical ability of their pupils.

Teachers' evaluation of the diagnostic tasks

All teachers found the results of the diagnostic tasks to be an important addition to their perception of their pupils' technical abilities. In particular, it influenced their perception of pupils for whom they had low expectations but who performed well. In addition, some results confirmed their suspicions about pupils' abilities, but which were not reflected in the pupils' performance in other tests at school. The teachers with high judgement accuracy gained confidence in their judgements because they saw this confirmed by the results. This implies that diagnostic tasks like those used in this study can change teachers' expectations of their pupils. This might positively affect the self-esteem of these pupils and even their learning (Timmermans et al., 2018; Zhu et al., 2018).

The teachers of pupils from nine years were positive about the limited time and effort needed for test-taking. The teachers who had younger pupils were less positive, as supervision during test-taking was necessary. When asked about the value of their teaching, seven out of eight teachers mentioned the better possibilities for differentiation. The eighth teacher mentioned a lack of own knowledge and experience as a limitation for using the results.

Whether diagnostic tasks will be used will depend on the effort and time taken to complete them and how teachers value the results (Praetorius et al., 2017; Sach, 2015; Zeinz & Dresel, 2017). Especially in the upper grades, task collection seems easy to organise. From the factors that determined the attributed value, gaining a better understanding of students' technical ability stood out. The responses confirm that teachers calibrate their estimates based on the outcomes of the tasks. Most teachers saw opportunities to use the insight gained from the diagnostic tasks in their technology lessons. For those who did not, this was attributed to a lack of knowledge and experience, which confirms the findings of Jones and Compton (1998) and Nadelson et al. (2013). Which effects on pupils' learning may be expected from differentiation based on the results of the diagnostic tasks is still unknown.

Implications

This study is the first to confirm the suspicions that teachers are not accurate in their judgements about the technical proficiency of their pupils. The results also reveal the often-assumed gender bias. For technical education in primary schools, this implies that many teachers will be poorly tuned to different levels of prior knowledge when choosing assignments, instruction and feedback. The teachers who used the tasks in this study valued them as an opportunity to discover qualities in pupils that would otherwise not be evident. The tasks allow teachers to calibrate their judgements and differentiate their engineering lessons.

Limitations and future research

A major limitation of this study is the small number of teachers who participated. Therefore, the results can only be regarded as a first indication of how accurate teachers assess pupils' proficiency in this domain and the factors that may or may not influence said accuracy.

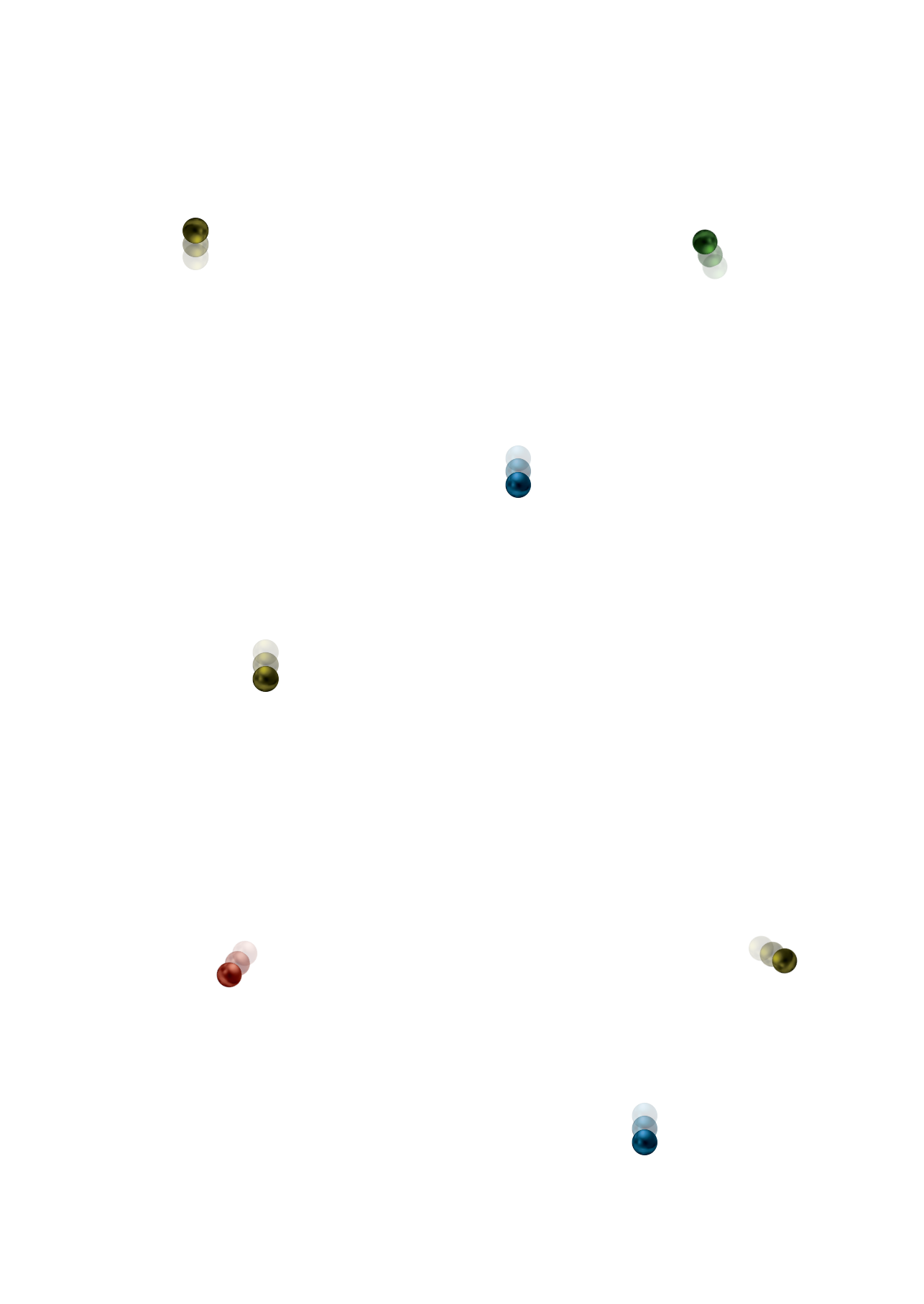
A second, equally important limitation is that only two tasks were used as references for pupils' technical abilities. Therefore, the results cannot be generalised to the full spectrum of technical skills. Studies with a broader range of diagnostic approaches might contribute to a more comprehensive understanding of the teachers' PCK and the amplifiers and filters that determine classroom practice (Doyle et al., 2019; Kärkkäinen & Vincent-Lancrin, 2013)

Another limitation is that the analysis of pupil results using the Fischer scale has yet to be carried out by teachers themselves. This makes it uncertain whether teachers will diagnose correctly with clear instruction or targeted training. A follow-up study will be needed to ascertain this. The significance of using such a diagnostic instrument for educational practice is also not yet clear. Therefore, it is recommended that research be carried out into the effect of differentiation in technology lessons based on the insight into pupils' prior knowledge that the new instrument offers.

Finally, it would be interesting to explore the value of tasks like those used in this study for a more unified assessment of student knowledge about technological systems (Hartell et al., 2015).

Conclusion

The current study confirms suspicions (Moreland & Jones, 2000; Scharten & Kat-de Jong, 2012; Südkamp et al., 2012) that teachers underestimate the technical skills of their pupils, particularly among girls. Teacher experience did relate to judgement accuracy, as did the teachers' technical knowledge, albeit to a lesser extent. The teachers' self-efficacy beliefs does not seem to affect judgement accuracy. The teachers were able to use the diagnostic tasks. They appreciated the results that emerged. The teachers were especially surprised by the unexpected high performances of some pupils. Contrary to the teachers' expectations, pupils' results on standardised reading comprehension and mathematics tests had a low predictive value for their performance on the Buzz-Wire and Stairs Marble Track task.



CHAPTER 4

FOSTERING PRE-SERVICE PRIMARY SCHOOL TEACHERS' ABILITY TO RECOGNISE DIFFERENCES IN PUPILS' UNDERSTANDING OF TECHNICAL SYSTEMS

This chapter is based on:
Wammes, D., Slof, B., Schot, W., & Kester, L. (2022b). Fostering pre-service primary school teachers' ability to recognize differences in pupils' understanding of technical systems. International Journal of Technology and Design Education, 1-20.
https://doi.org/10.1007/s10798-022-09774-x

Acknowledgement of author contributions:
DW designed the study, collected the data, planned the data analyses, analysed the data and drafted the manuscript; DW, BS WS and LK contributed to the critical revision of the manuscript; LK and WS supervised the study.

Summary

Pupils benefit from adaptive instruction and feedback from their teachers. A prerequisite for providing adaptive instruction is that teachers' diagnostic ability enables them to correctly perceive their pupils' skill level. A short course has been developed to improve primary school teachers' diagnostic ability for engineering. Based on Nickerson's anchoring and adjustment model, the participants became aware of the differences between their own and pupils' use of information when constructing technical systems. The Fischer scale was used as a model to understand and identify pupils' development in using such information. The participants were given examples of pupils' reconstructions of technical systems. They were asked to evaluate these work products in four ways: relative and absolute, combined with intuitive and explicit. The results reveal that relative and absolute diagnoses can differ considerably for the same teacher and between teachers, depending on whether they are implicit or explicit. Post-test results show that the course improved the ability to explain the differences between pupils' use of information to construct a technical system. The course also had a strong, significant, positive impact on teachers' self-efficacy beliefs about technology education.

Keywords: primary education; diagnostic ability, technical systems, training

Introduction

Technology education is part of the primary education curriculum in many countries, whether or not integrated into STEM. Learning outcomes in this domain can be improved when teachers align their instruction and feedback with pupils' prior knowledge (Behrmann & Souvignier, 2013; Hattie, 2013). Such alignment requires a correct diagnosis of pupils' ability level (Black & Wiliam, 1998; Shavelson, 1978; Van de Pol et al., 2010). Most primary school teachers, and even those with considerable experience in teaching technology, lack sufficient insight into the technical abilities of their pupils. As a result, they doubt the quality of their support for the optimal development of these abilities (Moreland & Jones, 2000; Scharten & Kat-de Jong, 2012). This study examines what a short course for prospective teachers can contribute to their diagnostic ability and whether this would impact their technology education self-efficacy. This is important since correct diagnoses of pupils' proficiency levels are the key to effectively adapting instruction, tasks and feedback to differences between pupils.

Diagnostic ability

In this study, we consider the diagnostic ability of teachers as a combination of their judgement accuracy and ability to explain and communicate their diagnoses. Teacher judgements can be relative or absolute (Südkamp et al., 2012). Ranking

pupils by their results is a relative type of judgement. The accuracy of a relative judgement is usually expressed as the correlation between the teachers' estimate and pupils' rank based on objective criteria. An absolute judgement is normative, for example, a teacher's estimate of the position of a pupil on a developmental scale or their test performance (Schrader & Helmke, 2001). The accuracy of absolute judgements can be expressed in different ways. Still, like relative judgement accuracy, it is usually expressed as a correlation here between the teachers' estimate and the pupils' results. Südkamp et al. concluded that teachers tend to be more accurate in their relative than in their absolute diagnoses. Only a weak relationship has been found between these types of judgement (Dunlosky & Thiede, 2013).

In addition to relative or absolute, a diagnosis can be implicit, based on intuition, or explicit, based on consciously and communicably weighing up the information (Wood, 2014). Wood argues that an implicit, intuitive judgement, which often arises from a first impression, could be based on the brain's fast, automatic System 1 processes (Kahneman, 2011), whereas an explicit evaluation will be primarily based on System 2 processes. Differences in accuracy could, according to Wood, relate to the interaction between the two processes. For example, there is evidence that a first impression (System 1) influences the choice of questions (System 2) asked during an assessment (Govaerts et al., 2013).

Pupils benefit most from teachers whose diagnoses are accurate and explicit (Edelenbos & Kubanek-German, 2004). Such diagnoses enable effective differentiated instruction and feedback (Van de Pol et al., 2010). The ability to diagnose correctly and explicitly requires knowledge about how a particular skill develops and how that can be recognised in pupils' activities (Gingerich, Kogan, Yeates, Govaerts, & Holmboe, 2014; Jones & Moreland, 2004).

Assessing pupils' technical skills is a complex endeavour for primary school teachers. Even within the subdomain of engineering, which is explored in this study, a wide variety of activities and associated skills exist (Pearson & Young, 2002). One of the overarching characteristics of engineering is that most of these activities relate to systems, e.g., constructions, pneumatic, mechanical and electrical systems, and ICT. In primary technology education, many activities are about such systems, ranging from building walls and roads to robotics with Lego Mindstorm (Brophy et al., 2008; Mullis & Martin, 2017; National Assessment Governing Board, 2013; Svensson et al., 2012).

Pupils' understanding of these systems develops hierarchically, from identifying the components of a system to the ability to imagine the systems' behaviour over time (Assaraf & Orion, 2010). This hierarchy is based on an increasing ability to combine knowledge about the system's components, interactions and functions. A similar hierarchy typifies Fisher's developmental model (1980). Therefore,

Sweeney and Sterman (2007) propose to use Fischers' model to interpret pupils' development in their understanding of technical systems.

Teacher characteristics and diagnostic ability

Teachers in primary education find it difficult to infer pupils' level of understanding of technical systems (Wammes et al., 2023). In their technology lessons, they tend to focus their feedback on other topics, like pupils' ability to cooperate or their mathematical skills (Moreland & Jones, 2000). Assumed causes include limited knowledge of technology (Jones & Moreland, 2004; Rohaan et al., 2012; Sanjosé & Otero, 2021) and insufficient knowledge about developing complex thinking skills (Retnawati et al., 2018), both of which are needed to understand technical systems. Other teacher characteristics related to diagnostic ability are self-efficacy beliefs (DePaulo et al., 1997), work experience (Ready & Wright, 2011; Wammes et al., 2023) and intelligence (Kaiser et al., 2012). These other characteristics only seem to explain teachers' diagnostic ability to a limited extent (see review Urhahne & Wijnia, 2020). It can be assumed that self-efficacy beliefs do not improve diagnostic ability; rather, they are affected by it.

Thus, opportunities to enhance teachers' diagnostic abilities in the context of technical systems lie primarily in broadening teachers' technical knowledge and their knowledge about how pupils develop their understanding of systems. Additionally, it can be expected that a course will be more effective when the content is linked to the participants' interests (Nauta et al., 2002). Generally, primary school teachers are particularly interested in developing their pupils' skills (Butler, 2012) and less in technical knowledge (Hsu et al., 2011; Knezek et al., 2011). Therefore, a course that aims to improve primary school teachers' diagnostic ability in the field of technology should make pupils' learning the focal point and introduce technical knowledge within that context.

Course design

Time for courses for teachers in primary education is scarce, but as Ostermann et al. (2018) have demonstrated, even short courses can positively affect diagnostic ability. Ostermann et al. designed their course about diagnosing pupils' ability to interpret graphs using Nickerson's (1999) anchoring and adjustment model. This model describes how we construct our ideas about what others know. It predicts that we tend to think that others think like us. Ostermann et al. used Nickerson's model to improve teachers' diagnostic abilities in three steps. First, they made the participating teachers aware of their strategies for interpreting graphs. Then, they showed them that their knowledge was incomparable to their pupils' knowledge. Finally, they created awareness of the task responses commonly shown by the pupils in their classes.

The course in this study followed the same steps as Ostermann et al. in three meetings of one and a half hours, including about 50 minutes for pre-test and post-test. First, the teachers were made aware of their understanding of how technical systems function by asking them to construct an electrical and a mechanical system while thinking aloud. How their thinking was incomparable to that of pupils in primary classrooms was demonstrated by showing them the results of their pupils on the same tasks. Finally, common responses from pupils in primary school were shown, categorised and explained by the different phases of Fischers' skill development scale (Fischer, 1980; Van der Steen, 2014; Wammes et al., 2023).

The Fischer scale consists of *three main phases*. The first phase (sensorimotor) is characterised by actions solely based on sensorimotor information. The second phase (representation) evolves out of repeated experiences and their neurological effects, which create the ability to remember what happened in previous situations and to include that knowledge in a choice of action (Edelman, 1992; Thelen & Smith, 1994). The third phase (abstraction) stems from the successive and repeated combination of multiple representations. In this phase, pupils are able, for a specific phenomenon, to identify its main characteristics and use these to choose an appropriate action in situations that have not been previously encountered. The first signs of actions or utterances based on reasoning at such an abstract level are usually seen between 12 and 14 (Fischer, 1980; Fischer & Bidell, 2006; Molnár et al., 2017).

Within each phase, Fischer distinguishes *three recursive levels*. The first level is a single piece of sensorimotor information, a single representation or a single abstraction that directs a pupil's action. The second level is known as 'mapping', which indicates the combined use of sensorimotor information in the first phase and the combined use of representations in the second phase. As the combined use of abstractions is very uncommon among pupils in primary education, the explanation of the Fischer scale was restricted to the level of a single abstraction. The third recursive level is called 'system' and indicates multiple combined sets of sensorimotor information or representations.

The Fischer scale was introduced in the course using examples of verbal utterances of learners described in several studies featuring the Fischer scale (Bassano & Van Geert, 2007; Meindertsma et al., 2014; Van der Steen, 2014) and then applied to pupils' attempts to restore an electrical and a mechanical system. The examples used provided an overview of the development of the use of information to reconstruct both systems.

Particular attention was paid to affordances (Gibson, 1977; Chemero, 2003). Affordances are found in the relationship between the properties of an object or situation and the possibility of an organism perceiving them. Affordances impose a strong, often unconscious, influence on a choice of action. An example of this

is the graphical objects on a computer screen. A button shape usually results in pressing, while a bar will elicit scrolling. Programming a button to react to scrolling would cause a lot of confusion.

Affordances play an important role in the way pupils interact with technical systems. They may result in effective actions but can trigger ineffective actions. Affordances play a major role in the first phase of the Fischer scale because, in this phase, actions are based on the available information. In the second phase, affordances are increasingly weighted by their possible function in the entire system (Svensson et al., 2012; Sweeney & Sterman, 2007). For instance, in constructing an electric circuit, most pupils tend to put clips on the outer points of connectors, even when these are insulated and even when these pupils correctly answer a multiple-choice question about the effect of conducting and insulating materials in an electric circuit. Through learning and experiences, pupils will gradually ignore their inclination to connect to outer points and only consider metal connections appropriate. During the course, participants learned to recognise the role of affordances within this developmental process.

When diagnosed on a certain level of the Fischer scale, what is needed to bring pupils' thinking forward was discussed in the course's third session. Making pupils aware of the unconscious role of affordances in decisions on actions is especially useful for pupils who are inclined to react without reflecting on the systems' function. Providing more experience and emphasising past experiences is especially needed for those with work products at the sensorimotor level. Encouragement and help to explain what guided the construction of work products are important for pupils who can construct work products at a high level. Such work products are often intuitively constructed, and discussing them helps pupils develop the language that can bring their understanding of technical systems forward.

Research question

This article presents the results of a small-scale study based on the question: What effect does a short-term course, based on Nickerson's anchoring and adjustment model and Fischer's model of dynamic skill development, have on the diagnostic ability of prospective teachers about pupils' comprehension of technical systems? The effects monitored were the teachers' relative-implicit (RI), absolute-implicit (AI), relative-explicit (RE) and absolute-explicit (AE) diagnoses and their self-efficacy beliefs about technology education at primary schools.

METHODOLOGY

Participants

The participants were 17 male and 34 female students of a university of Applied Sciences who followed a study to become a teacher in primary education. Their age ranged from 20 to 49, with a mean age of 25.4 years (*sd*=6.1). Eighteen participants followed the part-time program, which is meant for people who want to make a career switch to education. None of those participants had a technical background. All participants were in the final year of their study. In this phase of their study, they teach, under supervision, a few days a week in a primary school. For the participants, it was mandatory to follow several courses from a programme that included the current course. Therefore, it can be expected that the participants had some affinity with the subject. No post-test scores were available for two participants who did not finish the course.

Measurements, scoring and analyses

A pre-post design was used to evaluate the effects of the course on the diagnostic ability of the participants. The participants' technical knowledge was measured before the course as a covariate that might affect the participants' diagnostic ability and the course's impact on that ability (Pleasants & Olson, 2019; Sanjosé & Otero, 2021). An adapted version of the STEBI-b questionnaire was used to measure the self-efficacy beliefs of the participants before and after the course because we know that self-efficacy beliefs greatly influence teachers' teaching behaviour in general (Schipper et al., 2018; Hammack & Ivey, 2017). A diagnostic ability test was developed to determine the quality of the relative implicit (RI) and explicit (RE), and the absolute implicit (AI) and explicit (AE) diagnoses of the participants. The absolute-explicit diagnosis was not included in the pre-test as it required knowledge of the Fischer scale, which was first introduced in the course.

Technical knowledge

The test to determine teachers' technical knowledge was based on several editions (2015, 2016 and 2017) of a national admission test for students who want to become primary school teachers and lack sufficient qualifications (Admissiontests PABO, n.d). From these editions of the admission test, 35 items were selected with engineering-related content. All the items were multiple-choice questions with three or four answer options. For example, amperage is measured with an ammeter in a circuit with a resistor. Then another identical resistor is added to the circuit. What is the effect of the additional resistor on the current measured? The current (A) remains the same; (B) is doubled; (C) is halved; (D) is zero. Anderson's LR-test showed a good model fit (LR=23.697(31), *p*=.823) for the test. Latent scores were

calculated with eRM. The distribution of the knowledge test results deviated from a normal distribution. Therefore, the effect of technical knowledge on diagnostic skills was determined using Rfit (Kloke & McKean, 2012), a nonparametric, rank-based regression method.

Self-efficacy beliefs

The participants' self-efficacy beliefs were determined at pre-test and post-test using an adapted version of the Science Teaching Efficacy Belief Instrument - pre-service (STEBI-b: (Riggs & Enochs, 1990; Bleicher, 2004). The adaptation consisted of replacing the references to science with references to engineering. For example, "When a student does better than usual in science, it is often because the teacher exerted a little extra effort" was replaced with "When a student does better than usual in engineering, it is often because the teacher exerted a little extra effort." The modified version of the STEBI-b had a Cronbach's alpha of .85. The mean scale score was calculated to indicate the participants' self-efficacy beliefs, as in Bleicher (2004), using a 1-to-5-point value assigned to the Likert scale and a reversed value for negatively formulated items.

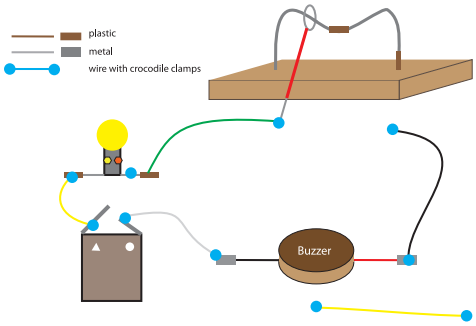
Whether the course had a significant influence on the self-efficacy beliefs of the participants was analysed by comparing the pre-test and post-test scores for the adapted STEBI-b with the Wilcoxon matched-pair signed-rank test. The pre-post correlation was considered in calculating effect size (Morris, 2008).

Diagnostic ability test

The diagnostic ability of the participants was tested by asking how they would interpret pupils' skills in constructing a technical system. Pupils' work products of two technical systems were used: the Buzz-Wire and the Stairs Marble Track. The Buzz-Wire (BW) is an electric circuit with a copper spiral and a ring with a handle. When the ring touches the spiral, it will activate a lamp and buzzer. The Stairs Marble Track (SMT) has a camshaft with eccentric wheels which support a set of bars of increasing height, each with a top that transports a marble in the direction of the roll-off point. Turning the camshaft allows the marble to roll onto the next bar and finally onto the slide, which brings the marble back to the roll-on point. Pupils from the upper primary classes were asked to construct these devices from their parts (see Chapter 2). Some of these work products, representing different developmental phases of the Fischer scale, were selected for the diagnostic-ability test. A schematic drawing represented BW work products (see Figure 14). SMT work products were represented by a photo with additional information on the bars' position (see Figure 15). The test consisted of three pairs of BW work products and three pairs of SMT work products. For each pair, 1) a relative-implicit, 2) an absolute-implicit, 3) a relative-explicit and 4) an absolute-explicit diagnosis was required.

Figure 14

Buzz-Wire device



Case 047 schematic drawing of a Buzz Wire work product. The pupil connects all parts. Clamps are only connected to the ends of the components' connectors (affordance), even though some of those ends are insulated (no application of knowledge about plastic being an isolator). A representation of an electric circuit is not used.

Figure 15

Stairs Marble Track device

Case 202 Stairs Marble Track work product, white numbers indicate correct bar positions, black numbers the bars as positioned by the pupil. The pupil applies the logic of the bar order in relation to the difference in height and the movement of the (blocked) camshaft. The slanted tops of bars 6 to 2 follow the virtual line (affordance) that connects the high roll-off point with the low roll-on point.

For the relative-implicit diagnosis (RI), the participants were asked which work product would reflect a higher level of understanding. The choice was correct when the work product represented the higher Fischer-scale level. The percentage of correct answers per participant was calculated for both pre and post-test. The significance of the difference was calculated in R with the related sample Wilcoxon Signed Rank test.

For the absolute-implicit diagnosis (AI), the participants were asked to indicate the difference between the levels of understanding reflected by the two work products of each pair on a seven-point Likert scale. Seven points was the maximum difference between no sign of any understanding represented by a lack of reconstruction and complete understanding represented by correct reconstruction. Per participant, the Intraclass Correlation Coëfficient (ICC; McGraw & Wong, 1996) was calculated (two-way random, single measure, consistency) for their estimates of the differences between the work products and the Fischer scale differences.

For the relative-explicit diagnosis (RE), the participants were asked to describe each pair of work products based on the knowledge they thought the pupils had applied. Two raters coded all the descriptions for three categories: system properties, technical terms and affordances. Per phrase, it was determined whether it referred to a system component or property. For example, the pupil does (not) know how to *connect* the *battery* too …; Both pupils know how to *connect* the *lamp*; The appropriate *order of the bars* is (not) recognised; Pupil A puts the *bars upside* down on the *eccentric wheels*, fitting the *slanted top* of the bar shape onto *the rim of the eccentric wheels*. The number of unique references to a particular system feature was tallied per comparison and participant. The agreement between the raters on which phrases referred to components or system properties was high (ICC3k=.88). The number of coded system properties (RE-sys) was used to indicate the participants' ability to explain their diagnosis using the observed differences.

The use of technical terms (RE-tech) was scored because domain-specific knowledge may play a role in teachers' diagnostic ability. First, the two raters independently determined which of the terms used could be deemed to be technical. The level of agreement was high (ICC3k=.88). After consultation, a list of 'technical terms' was drawn up, which included terms like electric circuit, poles (battery), metal, insulator, conductor, camshaft and axis. The number of different technical terms used per work product description was correlated with the results of the other diagnostic measurements.

References to affordances (RE-af) were scored because affordances play a major role in pupils' understanding of systems. The role of affordances in pupils' thinking and actions was introduced in the course. The level of agreement was .57. The same procedure as for the technical terms was used to create a coding list of phrases that were considered as referring to affordances, like 'Pupil A makes a slide', 'Pupil A fits parts like the pieces of a puzzle', 'Pupil B uses a virtual slope (from the highest to the lowest point) to determine bar positions- and 'Both pupils consider isolation on outer points (of lamp connectors) as suitable for attaching clips'.

The significance of the difference in the sums of the participant's RE-sys, RE-tech and RE-af references between pre-test and post-test was calculated in R with the paired t-test. Cohen's *d* was calculated as an indication of the effect size.

For the absolute-explicit judgement (AE), participants were asked to determine the level of skill development for each work product using the Fischer scale-based scoring rules (see Supplementary Materials, Chapter 4, Scoring rules Dutch). This absolute-explicit diagnosis was only solicited at the post-test because the participants did not know about the Fischer scale before entering the course. The Intraclass Correlation Coefficient (two-way mixed, absolute agreement) was used to indicate how the participants' rating matched the level calculated using an SQL version of the scoring rules described in chapter 2.

Procedure

Students took the knowledge test and the STEBI-b before the first meeting. The first meeting started with the first step of the course: becoming aware of one's perception of technical systems by constructing an electrical and a mechanical system. Then, they took the diagnostic ability pre-test that included the RI, AI and RE diagnoses. The diagnostic ability test was repeated at the end of the third meeting, with the addition of the AE diagnosis. Finally, the STEBI-b was filled-in again.

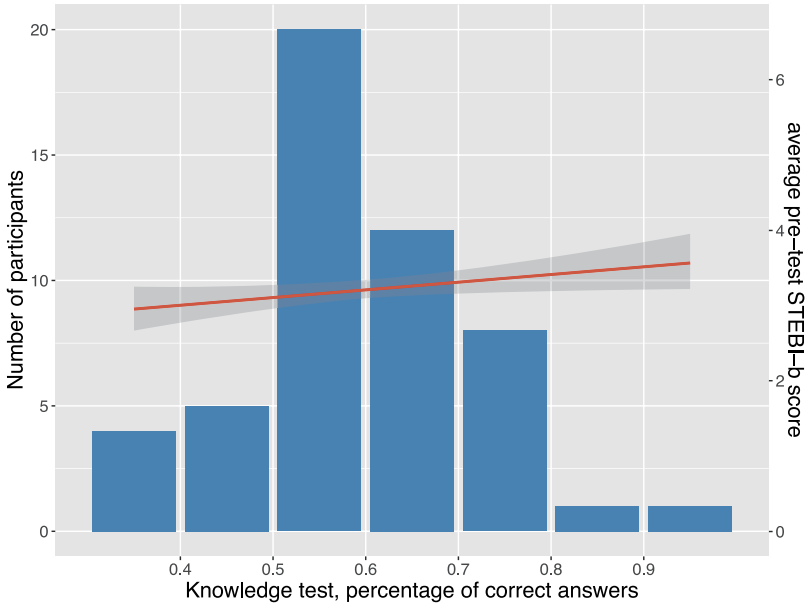
The course and the diagnostic test were piloted with six participants. The pilot resulted in some changes in the course and the diagnostic test. Some parts of the content of the course were removed, allowing more discussion about the remaining parts. Improvements were made in the wording of some of the questions and the pictures of the diagnostic test. Due to these changes, the data of these six participants were not included in the analyses. After these improvements, the course became part of the regular program. Data were collected in the first four sessions. Informed consent was requested from and given by all participants.

Results

The participants that opted for the course were interested in the domain. However, as Figure 16 shows, with 60% correct answers on the technical knowledge test, their subject-matter knowledge for the domain was limited, which is in line with other research (Culver, 2012; Ramaligela, 2021; Sanjosé & Otero, 2021). The pre-test scores on the adapted STEBI-b ranged from 2.26 to 4.48 with an average of 3.21 (*sd*=.43), which is below the average scores reported for the science domain, e.g., 3.62 (Riggs & Enochs, 1990), 3.58 (Bleicher, 2004). Figure 16 shows a weak, positive, non-significant correlation between the percentage of correct answers on the knowledge test and the participants' score on the adapted STEBI-b, r=.259, p= .072).

Figure 16

Average pre-test STEBI-b score related to the knowledge test score (n=34)



DIAGNOSTIC ABILITY

Relative-implicit diagnoses (RI)

For the relative-implicit diagnoses, a pair-wise judgement was used. The participants were asked to select the best from two work product pictures. No additional information was provided. In the pre-test, six participants were always correct in their choice. At the post-test, thirteen participants were always correct (see Table 13). The number of correct choices improved from 81% on the pre-test (76% BW, 86% SMT) to 90% on post-test (83% BW, 96% SMT). The related-samples Wilcoxon Signed Rank Test showed a significant difference (V=323, p=0,001) with a moderate effect size (r=.46). There was no significant improvement in correct choices for BW work products (V=187, p=.127), but a significant improvement for SMT work products (V=93, p=.010). It can be concluded that already at the pretest, for most comparisons, the participants could identify which work product did reflect a higher level of understanding. The course did improve this ability, but this improvement was only significant for the SMT work products.

Table 13

Participants' relative-implicit judgements about the best out of two work products [a]

Correct	Pre-test (n=49)	Post-test (n=47)
All six pair-wise judgements [b]	8	17
Five pairs	24	26
Four pairs or fewer	17	4

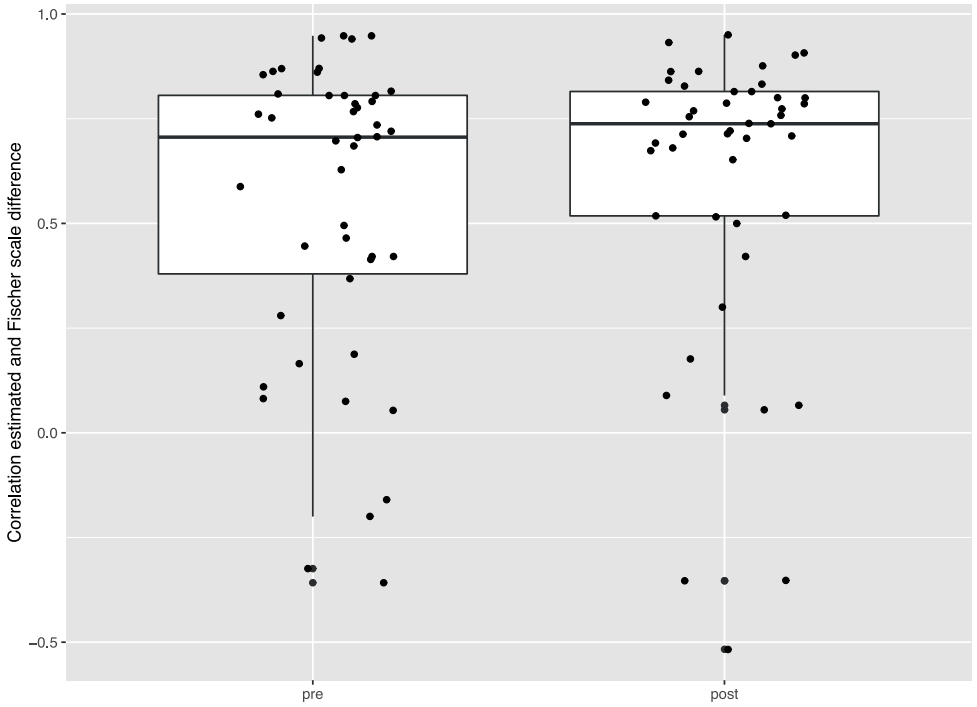
Note. [a] Three judgements on the best out of two BW work products and three judgements on the best out of two SMT work products.
[b] Correct when the participants' choice matches the difference as determined by the Fischer-scale-based scoring rules.

Absolute-implicit diagnoses

For the absolute-implicit diagnosis, the participants were asked to express the magnitude of the difference in understanding seen in the work products of each pair. At the pre-test, the correlation of their estimate with the Fischer-scale difference ranged from ICC(C,1) -.358 to .948 with a mean of .547. At the post-test, these correlations ranged from -.517 to .950, with a mean of .603. The boxplot of Figure 17 shows an overall improvement of the correlation between the estimates and the Fischer scale differences, but with considerable individual differences. At the post-test, there were 32 participants whose estimates were in line with the Fischer scale differences, of whom nine had low correlations at the pre-test. Six participants regressed to a substantially lower correlation at the post-test, and eight showed low correlations at both pre and post-test. There were missing values for three of the participants. The AI_{pre}-AI_{post} correlation was r_s=.480, p=.001.

Figure 17

Correlation between the participants' estimates of work product differences and the differences indicated by the Fischer scale (n=46)



The effect size of the course on the participants' ability to estimate differences between work products on a Likert scale could not be calculated due to the large standard errors of the correlations, which relate to the limited number of six comparisons per participant. From Figure 17, it can be concluded that the impact of the course on the absolute-implicit estimates was positive but limited.

Relative-explicit diagnosis

In their relative-explicit diagnoses, the participants used the differences between each pair of work products to infer and describe which knowledge had been used by both pupils. The system features, technical terms, and references to affordances were coded and counted for these descriptions. At post-test, there was a significant increase in the number of described system features, $t(45)=6.413$, $p<.001$, $d=1.04$, technical terms $t(45)=3.099$, $p=.003$, $d=.58$ and references to affordances, $t(45)=4.882$, $p<.001$, $d=0.67$. Table 14 summarises the differences between the pre-test and post-test for BW and SMT comparisons.

Table 14

References to system properties, affordances and technical terms in descriptions per device. Average per participant (n=48)

	System properties (RE-sys)				Affordances (RE-af)				Technical terms (RE-tech)			
	BW		SMT		BW		SMT		BW		SMT	
	Mean	*sd*	Mean	*sd*	Mean	*sd*	Mean	*sd*	Mean	*sd*	Mean	*sd*
Pre-test	7.5	2.4	6.2	2.4	1.2	1.3	2.4	2.2	3.8	2.9	0.4	.8
Post-test	9.0	2.6	9.2	2.4	1.9	1.8	3.6	2.0	5.1	3.3	1.4	1.7

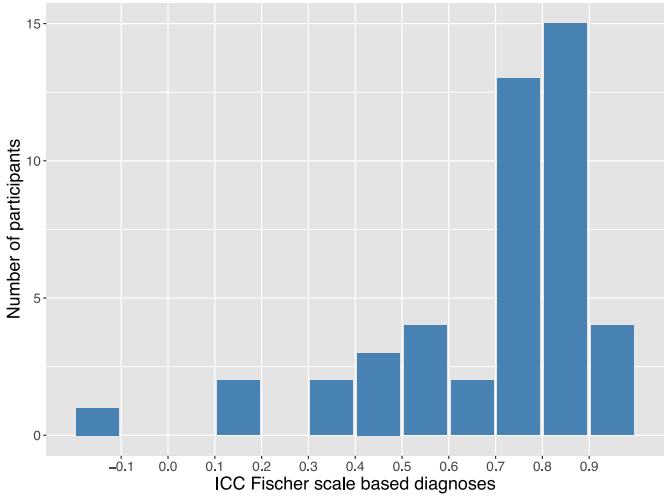
The results show that the participants recognise the role of affordances, especially in the SMT, which is obvious because the SMT system is based on differences in shape. The increase in affordances mentioned in the post-test can be attributed to the course, which made the participants aware of the role of affordances in pupils' actions. The more substantial increase in system properties for the SMT may relate to the fact that they were probably more experienced with simple electric circuits like the BW task than with a mechanical system as presented by the SMT task. The relatively large standard deviations indicate substantial differences between the participants, not only in their ability to recognise the differences but also in how they answered the question. Some participants answered in keywords and emphasised the most obvious differences, while others provided their answers in full sentences and mentioned both differences and similarities. The effect sizes indicate that the course improved the ability of the participants to notice and express the differences in terms of system features, technical terms, and references to affordances between the work products.

Absolute-explicit diagnosis

The absolute-explicit diagnosis required the application of the Fischer scale, which was introduced in the course. Therefore, it was only asked in the post-test, where the participants rated the six SMT and six BW work products using Fischer scale-based scoring rules. Figure 18 provides an overview of the distribution of the participants' ICC scores, which indicates the agreement rate between the participants' estimates of pupils' level on the Fischer scale and the scale level that the participants determined with the BW and SMT scoring rules (see Supplementary Materials, Chapter 4, Scoring rules). Out of 47 participants, 32 had an ICC above .7.

Figure 18

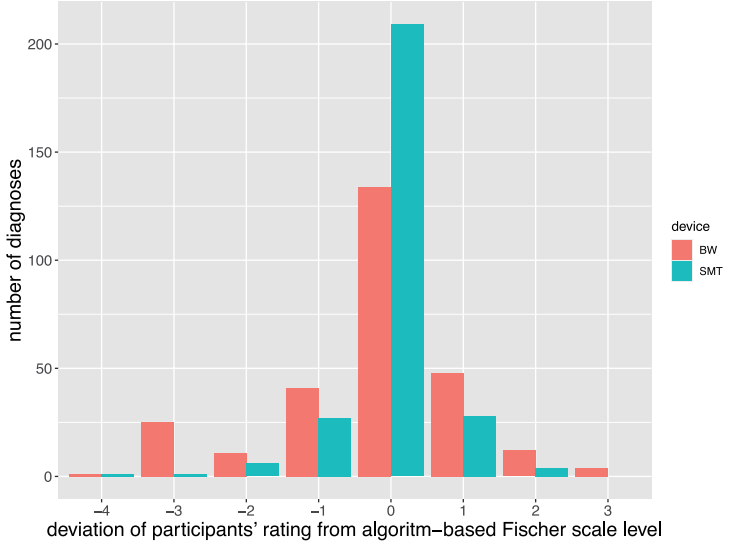
Distribution of ICC scores based on six BW and six SMT Fischer scale ratings (n=47)



The participants' ratings rarely deviated by more than one level from the algorithm-based reference (see Figure 19). It may be concluded that after the course, most participants were able to use the Fischer scale to interpret the developmental level reflected by pupils' work products.

Figure 19

Participants' Fischer scale ratings compared to an algorithm-based rating (n=47)



Participants' diagnostic ability

The results presented above indicate a positive effect of the course. However, not every participant benefited, and not all participants showed progress on all types of diagnoses. Table 15 shows the participants' post-test performances compared to the pre-test.

Table 15

Performance of participants by test compared to their pre-test result (n=47)

Type of diagnosis	Decline	Equivalent	Improvement
Relative-implicit accuracy (RI)	6	18	22
Absolute-implicit accuracy (AI)	6	23	15
References to system properties (RE_sys)	4	12	30
References to affordances (RE_af)	4	16	26
Use of technical vocabulary (RE_tech)	7	14	25

There were significant post-test correlations between the correct application of the scoring rules (AE) and the relative (RI) and absolute implicit diagnoses (AI) of the participants (see Table 16). Except for (RE_tech) and (AI), there was a lack of correlation between the diagnostic parameters (AE, RI and AI) and which differences in pupils' knowledge the participants identified.

Table 16

Spearman correlation post-test measurements

	RE_sys	RE_af	AE	AI	RI
RE_tech	.419**	-.171	.216	.313*	-.108
RE_sys		-.061	-.009	-.079	.033
RE_af			-.256	-.083	-.037
AE				.504**	.469**
AI					.211

*significant correlation at the 0.05 level, **at the 0.01 level, ***at the 0.001 level (2-tailed).

Rfit showed that the results of the knowledge test had a positive but non-significant effect on RE-sys$_{post}$ (b=7.78, t=1.20, p=.24), RE-tech$_{post}$ (b=9.09, t=1.57, p=.12), RE-af (b=7.50, t=1.64, p=.11) and AE, the correct application of scoring rules (b=0.13, t=0.72, p=.47). No effect was found on AI$_{post}$ (b=-.12, t=-.476, p=.64). This implies that the ability to identify differences has only a weak relationship with technical knowledge.

Self-efficacy beliefs

At post-test, participants' self-efficacy beliefs significantly increased from 3.24 to 3.60 on the five-point scale, $t(42)$=8.02, p<.001. The effect size d was 1.22, 95%CI [.82;1.61]. The STEBI-b showed a weak positive non-significant correlation with the knowledge test results on pre-test (r=.259, p=.072) and on post-test (r=.256, p=.086).

Discussion

Diagnostic skills enable teachers to adjust their instruction, tasks and feedback to their pupils' prior knowledge. It has been demonstrated that most teachers in primary education have limited diagnostic skills for technology education. To improve these skills, we developed a course for prospective teachers that consisted of three 90 minutes sessions, including about 50 minutes of testing. Like Ostermann et al., we used Nickerson's (1999) anchoring and adjustment model for the course design. This model aims to create awareness about the differences between the participants' ways of thinking about technical systems and pupils' ways of thinking and reactions in subsequent developmental phases. We estimated the effects of the course on the teachers' diagnostic ability by their relative-implicit (RI), absolute-implicit (AI), relative-explicit (RE) and absolute-explicit (AE) judgements and their self-efficacy beliefs about technology education at primary schools.

Most relative implicit (RI) judgements were already correct at the pre-test. At the post-test, all judgements about SMT pairs were correct, and overall there was a significant improvement with a moderate effect size.

The absolute implicit (AI) judgements were less accurate and showed large differences between teachers. At the pretest, about half of the teachers estimated the difference in knowledge application in line with the difference indicated by the Fischer scale. At post-test, about two-thirds of the teachers could infer the differences in understanding reflected by pupils' work products. Still, one-third of the teachers lacked accuracy for such estimates. The significant correlation with the AE post-test results suggests that those who could apply the Fischer scale could also make good estimates of the differences before determining the work products' position on the Fischer scale. However, the strong correlation between

AI_{pre} and AI_{post} opposes the suggestion that introducing the Fischer scale in the course explains the high AI_{post} scores of two-thirds of the participants.

The course had a significant positive effect on the teachers' ability to express the differences and similarities between the work products (RE) in terms of system properties (RE-sys), the use of technical terms (RE-tech) and their description of the possible influence of affordances (RE-af). At post-test, the teachers' diagnoses of the Fischer-scale level of the work products (AE) were in line with the levels calculated by the research team. There was a remarkably strong effect on the teachers' self-efficacy scores. Overall it may be concluded that the course supported most participants in developing their ability to diagnose pupils' understanding of technical systems. This supports the finding of Ostermann et al. (2018) that Nickerson's model (1999) might be an appropriate structure for a course to improve the diagnostic ability of pre-service teachers. More emphasis on the course as a way to become aware of how the unconscious application of one's thinking in instruction and feedback might sometimes be ineffective in bringing pupils' thinking forward might even attract more female prospective teachers to follow a course with scientific and technical content.

The results align with previous findings that teachers are generally better in their relative than their absolute estimates of performance (Lesterhuis et al., 2017; Südkamp et al., 2012). According to Schrader and Helmke (2001), an explanation of this difference is that a ranking better reflects the individual teacher's perspective on student performance. The significant correlations between the correct application of the scoring rules (AE) and the RI and AI judgements at post-test might indicate that those who understood the Fischer scale did apply that knowledge in their intuitive estimates. This might indicate that learning about and practising with the Fischer scale has improved the participants' ability to notice differences in pupils' technical ability.

The limited and sometimes negative correlations between RI, AI and AE, on the one hand, and the RE scores, on the other hand, imply that teachers who can indicate which work product is the better one, who can provide a good estimate of the magnitude of the differences and who can correctly apply scoring rules are not necessarily the teachers who can express the differences. That implies that a course to improve a teacher's diagnostic ability should not only focus on diagnostic accuracy but also on the ability to explain why something does not meet all the demands of a properly working system. Such an ability is crucial to provide tailored feedback that can help pupils' understanding forward (Van de Pol et al., 2010)

The average technological knowledge of the participants was about 60%, which differs from the average of 80% found by Rohaan (2012). Rohaan et al. noticed that their test was probably too easy for pre-service teachers, as their

questions were originally constructed for sixth-graders (ages 11-12). The current test seems better suited to revealing differences in technological knowledge between pre-service teachers. Sufficient subject matter knowledge has been identified as important for the quality of instruction (Hartell et al., 2015; Jones & Compton, 1998; Pleasants et al., 2020; Rohaan et al., 2009; Utey et al., 2019). For diagnostic ability, this study showed positive but non-significant correlations between the participant's scores on the knowledge test and their diagnostic ability. Therefore, it can be concluded that the participants' technical knowledge might influence their diagnostic capabilities but did not determine their scores on the diagnostic ability tests used in this study.

The strong positive effect of the course on the self-efficacy beliefs of the participants was an important side effect, as greater self-confidence is positively related to dedicating time to engineering education (Van Cleynenbreugel et al., 2011; Van der Molen, 2008). There are indications that experience improves the diagnostic ability of teachers (see Chapter 3). Strong self-efficacy beliefs might support novice teachers to get experienced in teaching science and technology.

Limitations

An important limitation of the present study is the limited number of participants combined with the elective nature of the course. Therefore, it cannot be assumed that the course would generate similar results with a random group of prospective or in-service teachers. The focus on two technical systems is another limitation. Engineering is a multi-faceted domain with specific skills (Mitcham, 1994; Pearson & Young, 2002). It is not likely that the trained teachers would be able to apply the knowledge about skill development to other types of systems or technical skills (Perkins & Salomon, 1992). Therefore, the conclusions about the effects are limited to the two technical systems used in the course.

According to Nickerson's (1999) model, insight into the thinking of others primarily arises from an awareness of how one's knowledge differs from that of others. This study did not explicitly examine the participants' or pupils' thinking about technical systems. Thus, the conclusion about the effectiveness of this model should be considered as hypotheses based on observations of the pupils' results and participants' considerations. Additionally, it should be emphasised that much technical knowledge is tacit and therefore offers limited opportunities for direct assessment.

Finally, strengthening diagnostic capacity does not necessarily equate to enabling teachers to adapt their instruction and feedback optimally to observed skill-level differences. To what extent that demands additional training in other aspects of teachers' pedagogical content knowledge requires further research.

Implications and Conclusion

Learning outcomes improve when teachers can adapt their instruction and feedback to differences in the proficiency levels of their pupils. This requires the ability to identify such levels through pupils' behaviour. It became clear that most teachers differed in their ability to compare pupils' results intuitively, analyse pupils' thinking, and interpret pupils' results using abstract scoring rules. This implies that the effect of a course on teachers' diagnostic ability should not be pinned down to a single variable. It can be concluded that a course based on Nickerson's model can positively affect the participants' diagnostic abilities and their self-efficacy beliefs about technology education.

CHAPTER 5

ADAPTING TASK DIFFICULTY TO PUPILS' PRIOR KNOWLEDGE ABOUT ELECTRIC CIRCUITS: EFFECTS ON PERCEIVED CHALLENGE ADEQUACY AND SKILL DEVELOPMENT

This chapter is based on:
Wammes, D., Slof, B., & Kester, L. (2022). Adapting task difficulty to pupils' prior knowledge about electric circuits: effects on perceived challenge adequacy and skill development. Learning and Instruction (submitted).

Acknowledgement of author contributions:
DW and LK designed the study, DW collected the data, planned the data analyses, analysed the data and drafted the manuscript; DW, LK and BS contributed to the critical revision of the manuscript; LK supervised the study.

Abstract

The challenge point framework presumes that the optimal challenge for skill development resides in the relationship between the difficulty of a task and pupils' prior knowledge. This relationship was explored in a lesson where pupils individually worked on practical tasks about electrical circuits with support that was limited to clarifying work-sheet texts and solving problems with the materials. It was hypothesised that pupils would appreciate the challenge of task sequences adapted to their level of prior knowledge and that this would be the optimal challenge for skill development. This hypothesis was confirmed for pupils' appreciation of the tasks but not for skill development. A challenge that transcended the level of prior knowledge was beneficial for low performers but hampered the skill development of high performers.

Introduction

Adaptive education provides better learning outcomes than education which ignores differences between pupils (Brühwiler & Blatchford, 2011; Hardy et al., 2019). One of those differences is students' prior knowledge about what is being taught. Learners can benefit from teaching that adapts instruction, task difficulty and feedback to their prior knowledge (Alsadoon, 2020; Atkinson, 1972, Bransford et al., 2000; Flores et al., 2012; Roschelle, 1997). The concept of prior knowledge must be interpreted broadly; it also concerns the mastery of metacognitive, procedural and motor skills (Davey et al., 2015; Glaser, 1984; Jonassen & Hung, 2006).

One of the domains in which primary school teachers find it difficult to estimate their students' prior knowledge is engineering (see Chapter 3). This limits the ability of the teachers to act adaptively. These difficulties have their origin in 1) a lack of technical knowledge of the teachers (Moreland & Jones, 2000), 2) the tacit character of the visual and procedural aspects of technical knowledge, which makes it harder to assess it (Mitcham, 1994), and 3) a lack of relevant assessment instruments to diagnose pupils' technical knowledge. Related to this last point, if a subject is frequently assessed, like reading comprehension and mathematics, teachers are quite accurate in estimating their pupils' prior knowledge, especially when they are asked to estimate a pupil's skill relative to other pupils (Südkamp et al., 2012).

The dissatisfaction of primary school teachers with their lack of insight into their pupils' technical knowledge did instigate the development of a diagnostic tool to determine the prior knowledge of pupils aged 9 to 12 years about technical systems (see Chapter 2). Research with this diagnostic tool has demonstrated that pupils' skills to construct technical systems differ considerably from their reading comprehension and mathematics abilities. As teachers tend to rely on those

abilities in their estimates of technical skills, this will result in incorrect estimates of the technical abilities of their students. This is especially affecting girls because their abilities are systematically underestimated (see Chapter 3).

The diagnostic tool has been developed for formative classroom use. Whereas adapting the difficulty of tasks to pupils' level of prior knowledge can be done outside class hours, adapting the task support is more difficult to accomplish, especially in hands-on lessons in which unforeseen practical problems are common and require direct attention (Doyle et al., 2019). The diagnostic tool is tailored to the hands-on lessons. Its results should enable teachers to adapt the challenge posed by the activities in lessons about a specific technical system to pupils' prior knowledge. The challenge can be adapted by offering tasks with a different difficulty level or by adapting the amount of support given for task completion (Ayres, 2006; Honomichl & Chen, 2012).

The difficulty of tasks can be determined by task characteristics (Jonassen, 2010; Savelsbergh et al., 2011; Schraagen, 2006; Sweller, 1994). The number of interacting elements determines the *nominal difficulty* of a task (Guadagnoli & Lee, 2004). In contrast, the *functional task difficulty* is the task difficulty relative to the learners' prior knowledge and expresses the challenge of a task for a particular learner. In their Challenge Point Framework (CPF), Guadagnoli and Lee predict that there is a challenge (i.e., a functional task difficulty) that optimally promotes learning. A low functional difficulty provides insufficient information for learning to occur, while a high functional difficulty provides too much information that exceeds the processing capacity of the learner, which also hampers learning.

The view that the optimal challenge for learning depends on the learners' prior knowledge is consistent with the Cognitive Load Theory (Sweller, 1994; Van Merriënboer & Sweller, 2005). The CLT states that the mental effort needed to solve a problem is affected by the amount of prior knowledge; that is, the more prior knowledge, the less mental effort is required to solve a particular task. However, both CPF and CLT do not indicate which amount of challenge or mental effort relative to prior knowledge is optimal for learning.

The general view in education is that learning should start at a level appropriate for each learner (Merrill, 2002). Approaches like mastery learning (Hattie, 2013) and the Increasing Difficulty (ID) strategy (Wickens, Hutchins, Carolan, & Cumming, 2013) all take pupils' prior knowledge as the starting point for learning and avoid a high cognitive load by presenting follow-up tasks that are slightly more difficult (Ahissar & Hochstein, 1997; Vanbecelaere et al., 2020). These approaches in which the complexity of follow-up tasks is based on the result of previous tasks are considered to be effective (Orvis et al., 2008; Salden et al., 2004), especially when there is shared control, in which the follow-up tasks are chosen in a dialogue between the pupil and the teacher (Corbalan et al., 2008).

Furthermore, learning also depends on pupils' *willingness* to invest cognitive effort (Feldon et al., 2019; Kuldas et al., 2015; Van Merriënboer & Sweller, 2005). Therefore, an optimal challenge will depend not only on the functional difficulty of the task and the mental effort needed to accomplish it, but also on how it is valued by the learner. A challenge that is valued as adequate is likely to support learning (Baird & Penna, 1996; Custodero, 2003). A non-adequate challenge may result in motivational problems and comprehension problems. Moreover, superfluous support from the teacher offers information that a more knowledgeable pupil already possesses, and processing this superfluous information reduces the cognitive capacity left for acquiring new skills or information and may induce frustration due to a lack of autonomy (Van de Pol et al., 2022).

Research question

The present study focuses on pupils' ability to construct electric circuits. This is a subject in the curriculum of primary schools in many countries (Mullis & Martin, 2017). In a lesson on this topic, task difficulty was manipulated while keeping support by the teacher equal. The main question addressed in this study is: What instructional strategy (an adaptive strategy, a non-adaptive strategy or a high-challenge strategy) leads to an optimal challenge for pupils' skill development about electric circuits? The challenge presented by an adaptive strategy is expected to contribute more to skill development than the challenge presented by a non-adaptive strategy.

Two experiments will be carried out to explore the effects of the adaptive strategy on pupils' perceived adequate challenge and skill development. The first experiment compares an adaptive strategy with task sequences that are tailored to pupils' level of prior knowledge with a non-adaptive strategy that randomly assigns the same task sequences to pupils. The second experiment compares the same adaptive strategy with a non-adaptive strategy in which only the most difficult tasks will be offered. Ethical approval for this study was provided by the faculty's ethical committee.

EXPERIMENT 1

Hypotheses Experiment 1

The first experiment uses a yoked design (Church, 1989; Kalyuga & Sweller, 2005) to compare an adaptive condition to a non-adaptive condition. Task sequences realised by pupils in the adaptive condition were randomly distributed among pupils in the yoked condition. We expect the highest level of skill development in the adaptive condition because it offers the most adequate challenge for learning.

For the yoked condition, we presume that the challenge for learning will be suboptimal (i.e., less adequate) as the task sequences they get are not adapted to their level of prior knowledge. We expect that challenge adequacy will be positively affected by pupils' ability to understand the procedural aspects of the tasks and by perceived teacher support. We also expect a positive relationship between mental effort and skill development (Paas et al., 2003), but we are uncertain about the effect of mental effort on challenge adequacy. Tasks that require low levels of mental effort might be experienced as adequate because they are easy to do, whereas at the same time, making tasks that require high levels of mental effort may also be perceived as a rewarding and adequate challenge (Inzlicht et al., 2018; Inzlicht & Campbell, 2022).

METHOD

Participants

The participants in this study were 104 female and 110 male pupils (mean age = 11.0, *sd* = .89). These were the pupils of eight classes of three primary schools in the Netherlands. Most pupils at these schools are from intermediate to high social-economic status families. Two schools had a specialised teacher and a special classroom for engineering activities. The experiment was embedded in their regular program. Parental consent was asked and granted for storing information about gender and age in a data repository.

Learning tasks

For this study, 45 different tasks were designed for the subsequent levels of skill development as described by the Fischer scale (Fischer & Bidell, 2007). No tasks were developed for the first level of this scale. A level one task would only ask pupils to pick up a part and move it to another position. Level seven was chosen as the maximum level for which tasks were designed. Wammes et al. (2022a) showed that only 5% of pupils of similar age understood an electric circuit at this level. The tasks and task copies allowed 15 pupils to work simultaneously and individually. Table 17 provides an overview of these tasks.

All tasks consisted of a worksheet with instructions and probe questions (Klahr et al., 2011) and a box with all necessary materials. Pictures were used that showed pupils how to start and change a configuration. Figure 20 shows for task 4.1 how the battery, wires, lamp, and material (A) should be connected to start. Low-level tasks were designed to be completed relatively quickly, allowing pupils to progress to higher levels within the available time.

Figure 20

Example of a worksheet picture used in task no 1 at level 4

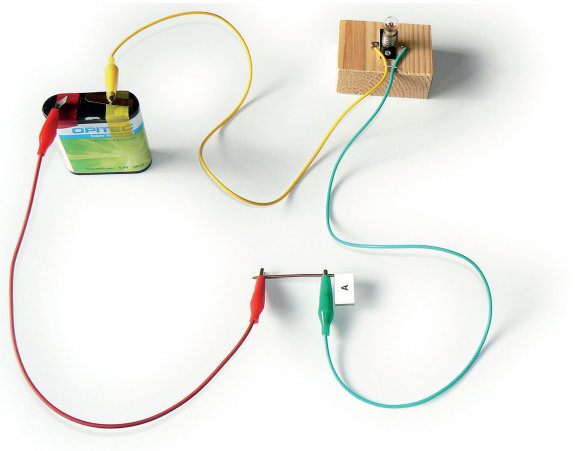


Table 17

Overview of tasks designed for the various levels of the Fischer scale

Level	N	Tot	Content
2	4	8	Matching lamps with their fitting. Connecting sockets and plugs.
3	6	12	Connecting sockets, battery and switches in different ways and with different materials. Discovering the benefits of a fitting and battery holder compared to fixed (soldered) connections. Test a solar panel with a Voltmeter in different light conditions
4	9	20	Conductivity. Battery as a source of power. A device (lamp) will only work when both poles are connected. A switch also has two poles.
5	8	20	Combinations of two level-4 concepts:
6	9	20	Electric circuits with multiple components.
7	9	22	Voltage, Amperage, resistance, serial and parallel circuits, logic circuits
	45	102	

Note. Level=Fischer scale level, N=number of different tasks, Tot=total number of tasks, including copies

Measurements

Challenge adequacy. The perceived challenge adequacy was measured with the Questionnaire on Teacher Support Adaptivity (QTSA; Van de Pol et al., 2022). This questionnaire has been developed for secondary school students and has 27 items about the perceived adaptivity of a teacher. The items can be answered on a five-point Likert scale ranging from totally disagree to totally agree. Van de Pol et al. have demonstrated that these items can be used to identify two aspects of a teacher's adaptive behaviour: challenge and support. Challenge can be raised by providing difficult tasks or by limiting teacher support. Providing difficult tasks can be considered adequate for pupils with sufficient prior knowledge, which is

asked for by five items, e.g., 'When I know how to do it, I get a more difficult exercise.' Providing difficult tasks can also be perceived as non-adequate when pupils lack sufficient prior knowledge, which is asked for by four items, e.g., 'When I do not yet understand the task, the teacher makes it more difficult for me.'

Regarding limiting support to raise challenge, the questionnaire has three items referring to a situation in which such behaviour is considered adequate, e.g., 'When I understand a task well, the teacher lets me do it on my own', and three items referring to situations in which refraining from additional support is perceived as non-adequate, e.g., 'The teacher tells me to do it on my own, even though I am unable to continue.'

Teacher support (raising challenge items excluded) can be perceived as adequate when it helps pupils to accomplish their tasks. The questionnaire has six items about such a situation, e.g., 'When I get completely stuck with an exercise, this teacher shows me how to do it.' Teacher support can also be perceived as non-adequate when pupils perceive it as superfluous. Six items refer to such a situation, e.g., 'This teacher helps me with things I already understand'.

Understanding and mental effort. At the end of each task, the pupils answered two five-point Likert scale items. The first item asked pupils how well they understood the instruction, varying from 'I did not understand' (non-adequate) to 'I fully understood' (adequate) what I had to do'. The second item asked about the effort needed for task completion, which ranged from little to extensive (Paas, 1992).

Skill development. Pupils' skill in constructing electric circuits was measured with the Buzz-Wire (BW) task. Scores are based on the Fischer scale (see Supplementary Materials Chapter 2, Scoring rules) and were calculated with an algorithm based on 15 variables, each representing the status of a connection (e.g. battery connection can be absent on one pole or on two poles). Wammes et al. (2022a) reported an interrater reliability of the algorithm with independent judgements of $ICC_{(3,1)}$=.875, p<.001, a test-retest reliability of $ICC_{(3,1)}$=.813, p=.002 and a good fit for the one dimensional Rasch model (LR-test: X^2=7.117, df=3, p=.71). Pupils cannot check whether their pre-test result is correct and do not receive information about their task performance. This allows the task to be used for post-testing.

Design and Procedure

The first experiment used a yoked pre-post design in which the adaptive condition was compared with a yoked condition. At pre-test, a PowerPoint was used to introduce the BW task and the QTSA. All pupils started with the BW task. When they had finished the task, they answered the QTSA questions while waiting for a picture to be taken of the results (i.e., work product) of the BW task. The QTSA questions

were answered about the preceding lesson. The questionnaire was used at the pre-test to familiarise pupils with answering questions on a 5-point Likert scale. The introduction, the BW task and answering the QTSA questions were done in half an hour in a whole classroom setting. Results were reported to the teacher of the class and not used in this study. The pictures taken of the pupils' work products were used to establish their level of prior knowledge. Pupils were not informed about their pre-test results or the condition that they were randomly assigned. Pupils who were absent at the pre-test were assigned to the yoked condition. Their pre-test level was established at the start of the lesson.

To limit differences between classes and teacher support, all lessons were given by the same teacher (i.e., the first author). In the lesson, pupils worked individually on the tasks, using the instructions provided on the worksheets. Per class, first the pupils assigned to the adaptive condition attended the lesson. They all started with two tasks that matched their pre-test level. The level of the follow-up tasks was determined by shared control and could be one level higher, at the same level or one level lower compared to their pre-test level. Pupils assigned to the yoked condition followed their lesson within a week after the lesson of the adaptive group. They were randomly matched with one of the pupils in the adaptive condition and got the task sequence made by that pupil. Post-testing was done at the end of the lesson. Pupils repeated the BW task and filled in the QTSA questionnaire about the lesson. The delayed post-test was administered at two schools (n=152) after six months without any engineering lessons due to COVID-19 restrictions. The data collection at the third school had to be postponed because of these restrictions. A delayed post-test could not be performed at this school as most pupils went to secondary education a few months after the lesson.

Scoring and Analyses

Challenge adequacy. The QTSA items on adequate challenge and the reversed scores of the items on non-adequate challenge due to task difficulty were used to calculate an average challenge adequacy score. The Cronbach's α for challenge adequacy items was .75. The results of the items on challenge resulting from limiting support were not used because challenge was only manipulated by differences in task difficulty. The *teacher support* adequacy was calculated by combining the scores on the items referring to adequate teacher support with the reversed scores of the items referring to non-adequate teacher support. The Cronbach's α for these items was .78. Individual scores were considered missing and excluded from the analyses when multiple items were unanswered. For understanding and effort, the average score was calculated from the responses on the accomplished tasks.

Skill development. Pictures of the completed BW tasks were used to identify the status of each of the 15 variables. The codes representing the status of the

variables were entered into a database from which the algorithm calculated the Fischer scale score. Test scores of zero were excluded from the pre-post analyses, as a lack of action does not provide any clue about pupils' understanding of electric circuits and is more likely to originate from motivational problems or fear of failure at test-taking.

Task difficulty. The nominal functionality of a task is its designed level (see Table 17). The functional difficulty of a task for a pupil was calculated as the difference between the tasks' nominal difficulty and the pupils' pre-test level. For example, a task at level 3 has a functional difficulty of minus one for a pupil with pre-test level 4.

Analyses. The SPSS mixed models MLmed macro (Rockwood, 2017) was used for multilevel mediation analysis. The condition was used as the predictor, challenge adequacy as the mediator and BW post-test scores as the outcome variable. Teacher support, understanding and mental effort were added as covariates. Data were clustered per class.

RESULTS

Descriptive statistics

The yoked design resulted in a similar number of completed tasks and a similar average nominal difficulty of the tasks per condition but in a larger variability in functional task difficulty resulting from the fact that the levels of the tasks that pupils started with were not adapted to their level of prior knowledge (see Table 18).

Table 18

Descriptive statistics per condition

	Adaptive (n=103)		Yoked (n=111)	
	male	*female*	*male*	*Female*
Gender	47	56	63	48
	mean	*sd*	*mean*	*sd*
Age	11.0	.88	11.0	.90
Tasks completed	4.48	1.65	4.79	2.06
Nominal task difficulty	5.32	1.18	5.23	1.26
Functional task difficulty	0.41	.53	0.32	2.25
BW score pre-test	4.87	1.42	4.92	1.33
BW score post-test	5.52	1.18	5.54	1.09
BW score delayed	5.41	1.36	5.57	1.23
Teacher support	4.04	.54	3.90	.59
Challenge adequacy	3.85	.67	3.35	.60
Understanding	4.19	.64	4.18	.66
Mental effort	2.67	.78	2.50	.86

Hypothesis testing

In line with our hypothesis, condition, teacher support, and task understanding had a significant positive effect on the mediator challenge adequacy (see Table 19), but there was no effect of challenge adequacy on skill development. Therefore, contrary to our expectations, challenge adequacy did not mediate skill development (BW scores). The differences between classes were limited. The ICC indicates that for challenge adequacy, only 5.4% of the variance relates to differences between classes. For the BW scores, this was 2.5%.

Table 19

Parameter estimates for multilevel mediation model, experiment 1

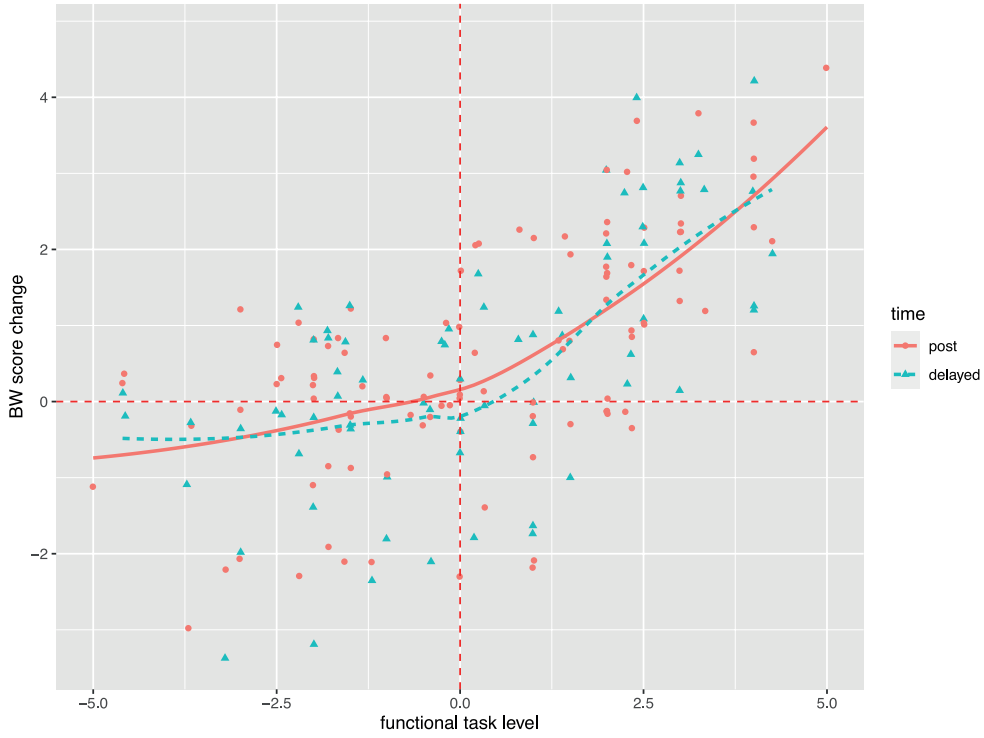
Predictors	Challenge adequacy (M)				BW post-test scores (Y)			
	b	SE	*t*	CI95%	*b*	SE	*t*	CI95%
Within classes (Level 1)								
Constant	-2.941	5.447	-.540	[-20.585; 14.702]	5.090	7.619	.668	[-25.958; 36.136
Condition (X)	.497	.057	8.713***	[.385; .610]	.056	.144	.385	[-.228; .339]
Challenge adequacy (M)					-.170	.114	-1.498	[-.393; .053]
Teacher support (Cov1)	.275	.058	4.762***	[.161; .388]	-.079	.133	-.594	[-.340; .182]
Understanding (Cov2)	.278	.047	5.891***	[.185; .371]	.173	.118	1.471	[-.058; .405]
Mental effort (Cov3)	-.045	.037	-1.220	[-.117; .027]	-.024	.084	-.281	[-.189; .142]
Between classes(Level 2)								
Condition (X)	2.573	4.089	.629	[-10.736; 15.883]	-.003	5.936	-.001	[-22.408; 22.401]
Challenge adequacy (M)					1.222	.803	1.523	[-1.956; 4.401]
Teacher support	.444	.264	1.683	[-.432; 1.320]	-.499	.532	-.937	[-2.523; 1.525]
Understanding	.685	.636	1.094	[-1.376; 2.765]	-.499	.979	-.509	[-4.418; 3.421]
Mental effort	.248	.591	.422	[-1,642; 2.138]	.047	.815	.057	[-3.293; 3.386]
Variance of random components								
Random intercept			.027				.031	
Residual variance			.319				1.426	
Direct effect, between-level					-.003	[-22.408; 22.401]		
Direct effect, within level					.056	[-.228; .339]		
Indirect effect, between-level					3.146	[-7.802; 17.790]		
Indirect effect, within-level					-.085	[-.203; -.026]		
AIC					1819.659			
BIC					1838.037			
-2LL					1811.659			
ICC			.054				.025	

* $p<.05$, ** $p<.01$, *** $p<.001$

The absence of an effect of condition on skill development was contrary to our hypothesis. We expected a negative effect on skill development in the yoked condition based on two assumptions of the CPF. The first assumption is that tasks that are less difficult than pupils' pre-test level will lack sufficient new information to enable skill development. This applies to pupils in the yoked condition whose average task level was below their pre-test level. The second assumption is that tasks more difficult than the pre-test level will offer an amount of new information that exceeds pupils' processing capability, which also will hamper skill development. The absence of a difference between the conditions implies that the results are not in accordance with both assumptions. To explore how our data deviated from these assumptions, we visualized the relationship between task difficulty and skill development.

Figure 21

Relation between average functional task level and BW score change compared to pre-test in the yoked condition.



When both assumptions are met, the plot should show the inverted U-shaped relationship as predicted by the CPF. The pattern in Figure 21 confirms the assumption that a lack of new information, which is expected for functional task levels below zero, does not further skill development. Contrary to our expectations, no negative effect of a high functional difficulty was found. Instead, skill development increased continuously with functional task difficulty. The data of the delayed post-test show that this pattern was still noticeable after six months.

Conclusion experiment 1

In the first experiment, a yoked design was used to explore our hypothesis that challenge adequacy mediates pupils' level of skill development with regard to electric circuits. Although, as expected, pupils' in the adaptive condition reported a higher challenge adequacy compared to those in the yoked condition, our hypothesis that challenge adequacy mediates skill development was not confirmed by the results. The major reason for the absence of more skill development in

our adaptive condition seems to be an unexpectedly strong positive and long-term effect in the yoked condition for pupils who had to start with difficult tasks compared to their prior knowledge.

EXPERIMENT 2

Hypotheses Experiment 2

The hypotheses of experiment 2 are similar to those of experiment 1. We expect that skill development will be mediated by challenge adequacy. Challenge adequacy, in turn, will be positively affected by perceived teacher support and self-reported understanding of the tasks. Based on the results of the first experiment, we expect no relationship between mental effort on skill development or perceived challenge adequacy.

METHOD

Participants

A total of 230 pupils participated in this experiment (104 male and 126 female; mean age = 10.4, *sd* = .99). The participants were pupils from a school with a population similar to that in the first experiment. Eight classes participated. Each class had pupils from the three highest grades.

Measurements

The measurement for *challenge adequacy* was changed as compared to experiment 1 for two reasons. The QTSA questionnaire was developed for secondary education. In the first experiment, pupils with a limited reading ability needed much time to read and answer all items. The second problem was that several items about challenge referred to questions posed by the teacher. This confused pupils because questioning was not allowed to keep teacher support similar in both conditions.

We revised the questionnaire to address these problems. To limit the time needed to complete the questionnaire, we left out the items about raising challenge by refraining from additional support. These items were also ignored in the analyses in experiment 1 because the differences in challenge between conditions only relate to task difficulty. Moreover, the number of items about teacher support was limited to six as this was considered sufficient to check upon unintended differences between the conditions. To take away the confusion raised by the items about questions posed by the teacher, the wording of these items was changed. For example, the phrase 'When I am doing well, this teacher lets me do a difficult exercise' was replaced with 'When I am doing well, I will get a more difficult task'. The shortened

version consisted of 16 items. For this version, Cronbach's α was .69 for the items about challenge. For the items referring to teacher support, the alpha was .71. The two questions at the end of each task about understanding and mental effort were identical to those of experiment 1.

Design and procedure

The second experiment was set up like the first one, with the same tasks and restrictions on teacher support. The differences were that the yoked condition was replaced with a high-challenge condition, and a shortened version of the QTSA was used. No delayed post-test was carried out.

Scoring and Analyses

The scoring and analyses were similar to those of the first experiment.

RESULTS

Descriptive statistics

For an overview of the descriptive statistics, see Table 20. The average age of the participants was somewhat younger than that in the first experiment, which relates to the fact that experiment 2 was done in the first months of the school year, while experiment 1 was done halfway (six classes) and at the end (two classes) of the school year. Pupils in the challenge condition completed fewer tasks than those in the adaptive condition. This stems from the design of the tasks. With increasing difficulty, more time was needed to complete a task. The different conditions are reflected in the average nominal and functional difficulty of the tasks.

Table 20

Descriptive statistics per condition experiment 2

	Adaptive (=114)		High challenge (n=116)	
	male	*female*	*male*	*female*
Gender	50	64	54	62
	mean	*sd*	*mean*	*sd*
Age (*sd*)	10.50	0.98	10.28	.91
Tasks completed	4.13	1.41	1.55	.65
Nominal task difficulty	5.04	1.03	7	NA
Functional task difficulty	0.68	.50	2.61	1.33
BW score pre-test	4.36	1.26	4.38	1.32
BW score post-test	5.18	1.22	5.15	1.16
Teacher support	3.46	.89	3.44	.76
Challenge adequacy	3.46	.71	3.38	.72
Understanding	4.03	.61	3.99	.89
Mental effort	2.34	.78	2.49	1.18

Hypothesis testing

In line with our hypothesis, teacher support had a significant positive effect on challenge adequacy (see Table 21), but as in experiment 1, challenge adequacy did not, as hypothesized, mediate BW scores. Contrary to our expectations and findings in experiment 1, condition and understanding did not relate to challenge adequacy. Unlike in the first experiment, there was a significant positive effect of mental effort on challenge adequacy, which indicates that higher reported average levels of mental effort did coincide with higher levels of perceived challenge adequacy. No effects were found between BW post-test scores and the other predictors. As in the first experiment, between-class differences were limited. The ICC indicates that for challenge adequacy, 1.5% did relate to differences between classes. For the BW scores, this was 2.7%.

Table 21

Parameter estimates for multilevel mediation model, experiment 2

Predictors	Challenge adequacy (M)				BW post-test score (Y)			
	b	SE	*t*	CI95%	*b*	SE	*t*	CI95%
Within classes (Level 1)								
Constant	7.691	4.389	1.752	[-6.288; 21.670]	18.908	10.373	1.823	[-22.570; 60.387]
Condition (X)	.109	.088	1.234	[-.065; .282]	-.014	.162	-.086	[-.305; .333]
Challenge adequacy (M)					.006	.129	.050	[-.247; .260]
Teacher support (Cov1)	.258	.055	4.684***	[.150; .367]	.089	.106	.841	[-.120; .299]
Understanding (Cov2)	-.054	.064	-.845	[-.181; .073]	.035	.118	.765	[-.198; .268]
Mental effort (Cov3)	.259	.051	5.043***	[.158; .360]	-.139	.100	-1.389	[-.335; .058]
Between classes(Level 2)								
Condition (X)	1.696	5.570	.305	[-16.288; 19.680]	4.851	9.950	.488	[-35.360, 45.063]
Challenge adequacy (M)					-1.287	.957	-1.344	[-5.037; 2.464]
Teacher support	-.388	.511	-.760	[-2.053; 1.277]	-.353	.961	-.367	[-4.294; 3.589]
Understanding	-.569	.520	-1.094	[-2.253; 1.115]	-1.480	1.050	-1.409	[-5.785; 2.799]
Mental effort	-.614	.642	-.957	[-2,700; 1.471]	-1.919	1.251	-1.534	[-7.038; 3.200]
Variance of random components								
Random intercept			.021				.061	
Residual variance			.410				1.377	
Direct effect, between-level					4.851	[-35.360; 45.063]		
Direct effect, within level					-.014	[-.305; .333]		
Indirect effect, between-level					-2.182	[-23.584; 14.283]		
Indirect effect, within-level					-.001	[-.038; .040]		
AIC					1114.905			
BIC					1130.969			
-2LL					1106.905			
ICC			.044				.013	

* $p<.05$, ** $p<.01$, *** $p<.001$

Conclusion experiment 2

The first experiment indicated that challenge had a positive impact on skill development. Contrary to our expectations, we found no negative effects of a high functional task difficulty. However, the number of pupils assigned to start with difficult tasks compared to their prior knowledge was small in the first experiment.

In the second experiment, we retested the optimal challenge point hypothesis but only on the predicted negative effects of a high functional challenge.

The results confirmed the findings of the first experiment. There was no effect of condition on skill development and perceived challenge adequacy did not mediate skill development. Unlike in the first experiment, there was no effect of condition on challenge adequacy. This might relate to less variance in the functional difficulty of tasks in the second experiment as no tasks below pupils' level of prior knowledge were provided. The positive effect of mental effort on challenge adequacy, not found in the first experiment, might have become manifest with the larger amount of data related to a high challenge. This positive effect might support the assumption that high levels of mental effort may be perceived as a rewarding and adequate challenge when tasks are considered meaningful (Inzlicht & Campbell, 2022).

Post-hoc analyses

In the first experiment, the adaptive condition had a positive effect on perceived challenge adequacy, but that did not result in a positive effect on skill development. The similar skill development in the yoked condition compared to the adaptive condition resulted from two opposing effects in the yoked condition. On the one hand, there was a negative effect on skill development for pupils who got tasks less difficult than their level of prior knowledge, but on the other hand, there was an above-average skill development for pupils who got the most difficult tasks. The latter effect was contrary to our expectation that the highest challenge levels would have a restrictive effect on skill development. The second experiment also revealed no restrictive effects of the most challenging tasks compared to the adaptive condition. To get more insight into the patterns that underlie these results, we visualised the relationship of task difficulty with reported challenge adequacy and skill development for each pre-test level separately.

Method

The data of all participants in both experiments were used (214 male, 230 female; mean age 10.7, *sd* = 1,0). Only 18 pupils (4%) had a level one or two result on the pre-test. That also affected the number of level-two tasks that were made. Therefore, we aggregated pre-test levels two and three as one category. This new category indicated as level three in the subsequent graphics includes all levels of the sensorimotor tier of the Fischer scale (Fischer & Bidell, 2007). GGplot was used to visualise the relationship between task and post-test levels for each pre-test level separately. We did not design our experiments to realise a proportional distribution of all data over the different combinations (see Table 22). The patterns shown in Figure 22 and Figure 23 thus should be considered indicative.

Results

The effect of task difficulty on pupils' perceived challenge adequacy is displayed in Figure 22. The highest scores on challenge adequacy for most pre-test levels were found at the points where task difficulty matches pupils' pre-test level (i.e., at a zero functional task difficulty). This situation was created in the adaptive condition and accounts for the significant relationship between condition and challenge adequacy found in the first experiment. Remarkable is that pupils with the lowest pre-test level perceived the challenge offered by the lowest-level tasks as well as the challenge offered by the most difficult tasks as most adequate. As the latter situation was created in the challenge condition, that may explain why the significant relationship between condition and challenge adequacy no longer occurred in the second experiment.

Figure 22

Relationship between initial task-level and reported challenge adequacy per pre-test category

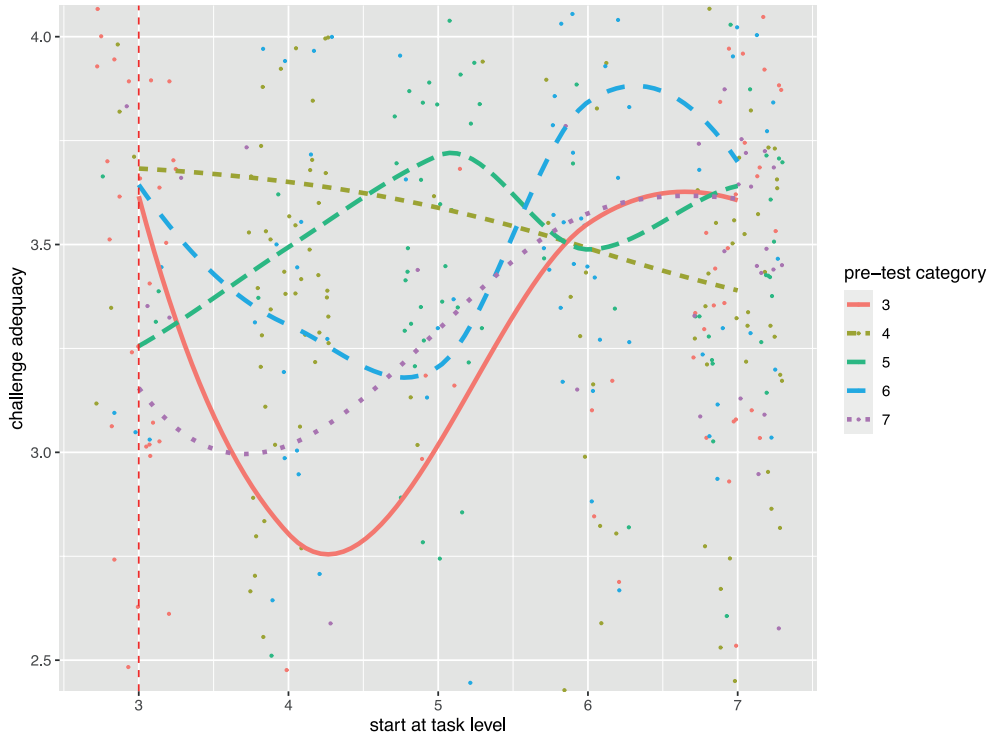


Table 22 summarises the mean post-test scores of participants with different pre-test scores and initial task levels. These data are displayed in Figure 23.

Table 22

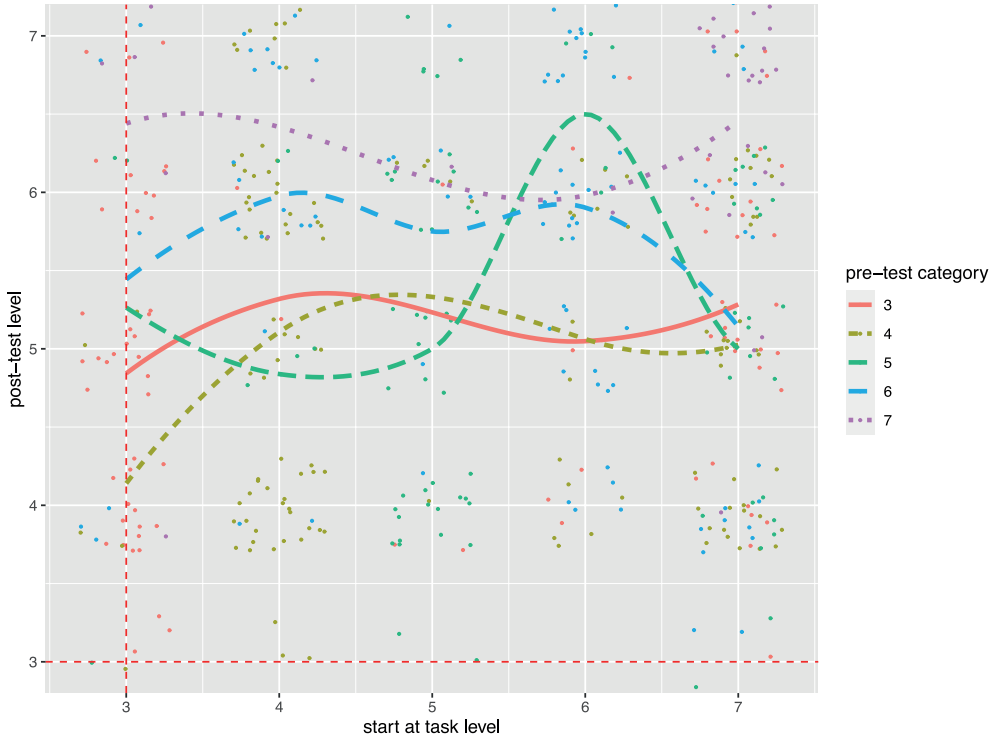
Average post-test scores per pre-test level and start-task level

		Task level started with														
Pre		3[a]			4			5			6			7		
score	n(tot)	*m*	*sd*	*n*	*m*	*sd*	*n*	*m*	*sd*	*n*	*m*	*sd*	*n*	*m*	*sd*	*n*
3[a]	97	**4.89**	*1.11*	*45*	**5.00**	*1.41*	*2*	**5.0**	*0*	*2*	**5.22**	*1.20*	*9*	**5.30**	*1.15*	*37*
4	134	**4.33**	*.58*	*3*	**5.06**	*1.20*	*68*	**5.57**	*.79*	*7*	**5.13**	*.92*	*15*	**5.00**	*1.00*	*41*
5	82	**5.00**	*1,73*	*3*	**5.25**	*.50*	*4*	**5.05**	*1.13*	*43*	**6.50**	*.55*	*6*	**4.92**	*1.04*	*25*
6	96	**5.33**	*1.51*	*6*	**5.87**	*.99*	*15*	**6.00**	*.93*	*8*	**5.96**	*1.03*	*46*	**5.14**	*1.39*	*21*
7	37	**6.43**	*1.13*	*7*	**6,67**	*.58*	*3*			*0*	**6**	*0.0*	*1*	**6.42**	*.81*	*26*

Note. [a] Including levels 1 and 2.

Figure 23

Relationship between initial task-level and average post-test result per pre-test category



Overall, the pattern shown in Figure 23 shows that starting with tasks of a level that matched the pre-test (i.e., our adaptive condition) did not result in the highest post-test levels.

Pre-test level seven. As the BW task cannot indicate skill development above level seven, the mean post-test level can be only affected by lower post-test results. The post-test result of 6.4 of pupils with a level seven pre-test result thus

might be affected by this ceiling effect.

Pre-test level six. The post-test results of pupils with pre-test level six were not affected by a ceiling effect as there was an opportunity to progress to level seven. However, this group lacked any progression, and their post-test results were even negatively affected when they had to start with level seven tasks compared to starting with level six tasks. This negative effect was significant, $t(30.5)= -2.400$, $p=.023$, $d=-.705$. These pupils also showed lower challenge adequacy scores when starting with level seven tasks than starting at their pre-test level six (see Figure 22), but this difference was not significant, $t(41.3)=0.641$, $p=.525$.

Pre-test level five. The post-test results of pupils with pre-test level five show the inverted U shape relationship with tasks-difficulty that the CPF predicts. Furthermore, for this group, we expected the optimal challenge when starting with tasks at level five, but the post-test results of pupils who started with level six tasks were significantly better, $t(8.6)=5.801$, $p<.001$, $d=1.41$. However, it must be emphasised that only six pupils with a level five pre-test started with level six tasks.

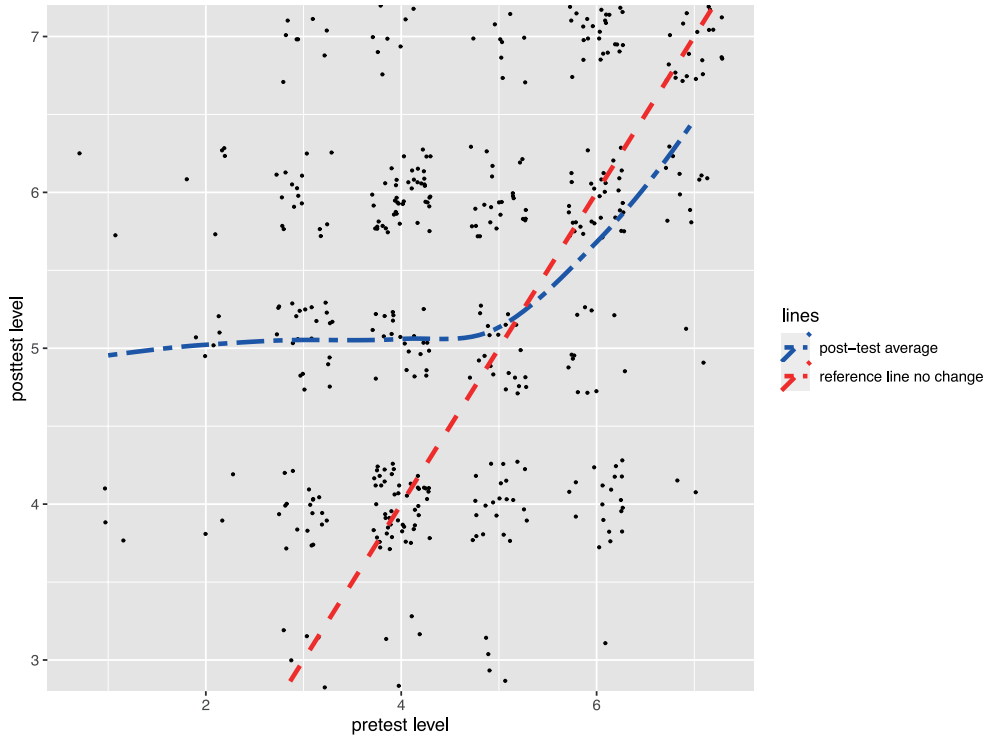
Pre-test level four. Pupils with pre-test level four show the same pattern as those with pre-test level five, with a less outspoken inverted U-shaped relationship between task difficulty and skill development and the best post-test results of pupils (n=6) that started with tasks one level beyond their pre-test level.

Pre-test levels two and three. Pupils with the lowest pre-test levels (two and three) showed the most intriguing relationship between task difficulty and skill development. This group progressed towards an average level five post-test result regardless of the difficulty of the tasks they had to start with. They even showed the best post-test results when they had to start with the most difficult tasks. Interestingly, this coincides with high scores for challenge adequacy (see Figure 22).

The overall effect of the lesson on post-test results per pre-test level is shown in Figure 24. It shows that the hands-on lesson had a strong convergent effect with a progression of pupils with low pre-test level to a level five post-test result and no progression of pupils with the highest pre-test level.

Figure 22

Relationship between initial task-level and reported challenge adequacy per pre-test category



GENERAL DISCUSSION

Conclusions

The research question of this study was "What instructional strategy (an adaptive strategy, a non-adaptive strategy or a high-challenge strategy) leads to an optimal challenge for pupils' skill development about electric circuits?". Four conclusions can be drawn from our experiments. The first conclusion is that the highest scores on challenge adequacy were not found when pupils had to start with tasks below their level of prior knowledge. The second conclusion is that a hands-on lesson without content-related teacher support is only beneficial for pupils with low levels of prior knowledge. The third conclusion is that adaptive task selection can facilitate skill development but that offering the optimal challenge by task selection is complicated by a changing relationship between prior knowledge and the optimal challenge point. The fourth conclusion is that adaptive task selection based on the ID strategy positively affects pupils' appreciation of the challenge offered by the tasks.

Minimum level

The first conclusion that the minimum level of the tasks should be adapted to pupils' level of prior knowledge might be obvious as teaching what is already known is generally considered ineffective. However, teachers in primary education have a tendency to underestimate the technical abilities of their pupils, leading to a lower estimation of their prior knowledge (see Chapter 3), which may negatively affect skill development when the tasks a teacher selects for the pupils are too easy for them (De Boer et al., 2010). The contribution of the diagnostic tool in the adaptation process is that it offers a more objective evaluation of a pupil's skill development, and its use is especially important for girls as the underestimation of their technical capabilities is common (Hoffmann, 2002).

Teacher support and skill development

The second conclusion was that our hands-on adaptive lesson without content-related teacher support was only beneficial for pupils with low levels of prior knowledge. This is remarkable as it is often claimed that pupils with low levels of prior knowledge need more instructional guidance and support than pupils with higher levels of prior knowledge (Dalgarno et al., 2014; Kuldas et al., 2014; Matlen & Klahr, 2013; Renkl et al., 2011; Wulfeck, 2009). In contrast, in our experiments, this group progressed without such guidance and support, whereas pupils with pre-test level six did not show any skill development in any condition. We presume that teacher support may be especially beneficial for these pupils. Where tacit knowledge may be sufficient to construct a basic electric circuit at level five (i.e., the highest level reached by most low prior-knowledge pupils), we presume that declarative knowledge is necessary to construct the BW task at level seven (Danish et al., 2017; Hattie, 2013; Sutherland, 2002). Instruction and feedback may help pupils to transform their tacit knowledge into the declarative knowledge they need to complete the tasks at the seventh level.

Task selection and skill development

The third conclusion was that offering the optimal challenge by task selection to improve skill development is complicated. Such a conclusion was also drawn in other studies which explored the CPF (Balali et al., 2019; Bootsma et al., 2018; Hodges & Lohse, 2020; Onla-Or & Winstein, 2008; Pesce et al., 2013). One of the complicating aspects was the effect of high challenge. In contrast to our expectations, highly challenging tasks positively affected the skill development of pupils with limited prior knowledge. At first sight, this seems to contradict the CLT, which states that a large amount of new information is likely to exceed mental processing capabilities, which limits learning. The lack of such an effect may relate to two features that characterise the design of the high-level tasks. The first feature is that these tasks started with a stepwise instruction to construct a basic circuit guided by pictures, after which pupils

had to discover the effect of certain manipulations. We assume that the initial stepwise instruction with pictures served as a worked-out example which enabled pupils with low prior knowledge to apply the basic visual characteristics of an electric circuit at the post-test. This seems to be in line with the finding of Ahissar and Hochstein (1997) that practising simple visual tasks can lead to a substantial improvement in performance and with the well-documented effects of worked-out examples for novice learners (Atkinson et al., 2000; Chen et al., 2019; Cooper & Sweller, 1987; Sweller & Cooper, 1985; Van Peppen et al., 2021). As worked-out examples effectively reduce cognitive load (Van Merriënboer et al., 2003), this might also explain why pupils with low prior knowledge did not report higher levels of mental effort on the most complex tasks compared to tasks designed for their own level of prior knowledge. They even reported similar high challenge adequacy scores on tasks adapted to their level as on the most challenging tasks. This lack of a negative effect on mental effort and challenge adequacy points to a second feature of the task design that might have contributed to the unexpected positive effect of highly challenging tasks on the skill development of pupils with the lowest pre-test scores. Pupils could not infer whether their responses to the questions about the effect of resistance were correct or incorrect. This 'goal-free' character of the tasks is also known to reduce cognitive load (Van Merriënboer, 2013). We assume that the goal-free character of our tasks allowed pupils to reduce their cognitive load by selecting only the information they could process, resulting in a high satisfaction with the guided (i.e., by pictures) opportunity to construct functioning electrical circuits.

Another example of the complex relationship between task selection in relation to prior knowledge is the reverse effect of high challenge on the post-test results of pupils with the second highest level of prior knowledge. We suppose that most pupils constructing a level six electric circuit at the pre-test envision an electric circuit as a single circle, starting and ending with a battery. This construction has been described as the Closed Circuit Model (Shipstone et al., 1988). Often-used analogies like a bicycle chain or a water-flow model may have strengthened this conception (Chiu & Lin, 2005; Safadi & Yerushalmi, 2014). Most of the level seven tasks used parallel circuits. The experiences with these more complex circuits may have modified the view of an electric circuit as a single circle. This may have resulted in not using the initial tacit knowledge-based strategy in the post-test while, at the same time, these pupils could not apply a new, more adequate problem-solving strategy. In such a situation, it is common for pupils to apply a less adequate strategy. This scalloping effect (i.e., lower performance resulting from learning) is often observed in the transition towards higher levels of understanding (Fischer & Bidell, 2007; Granott et al., 2002; Levy, 2012; Van der Steen, 2014).

Task selection and challenge adequacy

Our fourth conclusion that adaptive task selection based on the ID strategy positively affects pupils' appreciation of the challenge offered by the tasks is interesting in combination with our finding that there was no relation between this appreciation and skill development. This implies that when the main objective of a lesson about electric circuits is providing positive experiences with the topic, then pupils should start with tasks at their pre-test level as these are valued most. When the lesson's main objective is to promote skill development about electric circuits, pupils with the lowest pre-test levels should start with tasks of one scale above their pre-test level, and pupils with high pre-test levels should get tasks of their pre-test level and presumably, additional teacher support to help them progress from tacit knowledge to declarative knowledge.

Limitations

The current study's main limitation is that the effect of challenge was only explored in the context of skill development about electrical circuits. Moreover, as we used a performance-based diagnostic task that allows pupils to apply their visual and procedural knowledge, our results are difficult to compare with other studies about pupils' understanding of electrical circuits, which assessed pupils' declarative knowledge. (Bumbacher et al., 2018; Greiff et al., 2014; Safadi & Yerushalmi, 2014; Solomonidou & Kakana, 2000; Zacharia, 2007).

In addition, the same scale categorised the difficulty of the tasks we used in our lesson and pupils' prior knowledge. This scale was validated to measure prior knowledge but not for the classification of the tasks used in the lesson. This classification was based on the theoretical structure of the Fischer scale and has not been validated by empirical data. The satisfaction of pupils with the challenge posed by the tasks categorised at their level of prior knowledge was only an indirect indication that the categorisation might be valid.

Furthermore, the study focused on the challenge offered by task difficulty in relation to prior knowledge. The effects could be attributed to task difficulty by excluding other adaptive support, like instruction, feedback, and cooperation. The only help offered was clarifying written instructions and solving problems like not properly mounted lamps or defective components. It is known that adaptive measures other than task difficulty have a substantial effect on pupils' learning (Hattie, 2013; Kirschner et al., 2006; Sweller et al., 2007). Therefore, this study can only be considered a first step to unravelling the effects of adapted challenge on pupils' skill development related to material-based systems.

The post hoc analysis revealed a relationship pattern between prior knowledge, task level and skill development. However, this pattern should be interpreted with caution because the data distribution over the combinations of pre-test and task

level was not homogeneous, as the experiments were not designed to create such a distribution. An explorative study to collect data points for all combinations of prior knowledge and task levels would be advisable to verify whether the patterns shown in this study are robust.

Suggestions for further research

The present study was the first to explore the relationship between prior knowledge, adaptive teaching and skill development in material-based systems. Our approach can be used for follow-up studies that combine the effect of the task-selection strategy with other adaptive measures, like small-group work, direct instruction, and feedback. It would be interesting to explore how these adaptive measures interact with the task selection strategy at different levels of prior knowledge. Our approach also offers an opportunity to explore whether hands-on experiences are necessary to develop an adequate understanding of material-based systems. It is a basic assumption of the dynamic systems theory that sensorimotor experiences are the primary building blocks for subsequent learning (Fischer, 1980; Thelen & Smith, 1994), a stance also advocated by Dewey (1986). From that point of view, learning about material-based systems should start with providing opportunities to explore a system like in our lesson. However, Matlen and Klahr (2013) have demonstrated that two lessons with only a teacher-guided exploratory approach were more effective in enabling pupils to master the Control of Variables Strategy than a hands-on lesson followed by a lesson that explained the strategy. Using their experimental design in the context of electric circuits would be interesting.

Key findings

The present study has shown that adaptive task selection based on the ID strategy is beneficial for pupils' appreciation of hands-on tasks, which is important as raising interest in engineering is an important objective of teaching the subject in primary education. We also showed that pupils' appreciation of the challenge offered by the tasks does not indicate how much they learn. Our results indicate that the relationship between functional task difficulty and skill development varies per level of prior knowledge. This complicates task selection with the objective of offering optimal learning conditions. However, we demonstrated that selecting tasks below the level of prior knowledge indicated by the diagnostic tool did not enhance skill development. This finding is important as teachers in primary education tend to underestimate their pupils' technical abilities, which may result in offering too easy tasks and therefore stresses the importance of assessing that ability.

CHAPTER 6

DIAGNOSING ENGINEERING SKILLS IN PRIMARY CLASSROOMS

Introduction

Engineering in primary education intends to familiarise pupils with the technologies that surround them and to raise their interest in engineering as an opportunity for further study and career (ITEEA, 2007; Pearson & Young, 2002). The latter is important as many challenges related to sustainable development require technological solutions, which in turn depend on people that are able to create, realise and maintain them (Lewis, 2019; Wiesner, 2014). The role of primary education cannot be underestimated. It is known that without positive experiences with the subject at a young age, it is more likely that pupils exclude engineering as an opportunity for further study and career (Van Tuijl & Walma van der Molen, 2016). Offering those positive experiences at primary school is important as pupils seldom encounter and participate in engineering-related activities in daily life. This dissertation focuses on the difficulties teachers experience concerning teaching engineering in primary education.

Teachers in primary education find it difficult to estimate their pupils' technical skills. Insight into these skills and their development prior to their lessons would offer them opportunities to adapt their teaching to differences in prior knowledge (e.g., skill level), evaluate the effectiveness of their lessons, inform pupils and parents about their skills and progress in this domain and communicate with colleagues about successful approaches (Gumaelius et al. 2019; Moreland & Jones, 2000; Scharten & Kat-de Jong, 2012). This difficulty and the potential benefits of increasing teachers' ability to diagnose technical skills resulted in our main research question: "Which support can a diagnostic tool designed for classroom use offer teachers to infer and promote pupils' engineering skills?" This question was addressed in four studies. First, we designed and validated the diagnostic tool; second, we explored the feasibility and value of its use in primary classrooms; third, we trained pre-service teachers in using the tool to improve their understanding of engineering skill development and fourth, we explored the effect of the tool's formative use.

Main findings

Pupils' prior knowledge about technological systems: design and validation of a diagnostic tool for primary school teachers

The first step to answering our main research question was to design and validate the diagnostic tool. **Chapter two** describes how we used the four layers of the evidence-centred design (ECD) framework (Mislevy & Riconscente, 2005) to develop a valid tool, feasible for classroom use. From the ECD layers 'domain analysis' and 'product requirements', it became clear that there were three major design challenges that should be solved when developing a valid and feasible tool.

The first challenge was to include the tacit aspects of prior knowledge. This

was addressed by a design focus on pupils' actions as opposed to verbalizations (Davey et al., 2015). Opportunities to find a correct solution by trial-and-error behaviour were limited by restricting the possibilities to evaluate the appropriateness of actions (Klahr & Robinson, 1981) to ensure that the work product is constructed by actions based on prior knowledge only.

The second challenge was to reconcile feasibility and validity. Feasibility was created by designing a tool with single tasks and a focus on work products as the unit of analysis. The validity of single-task use was strengthened by optimising the opportunities to make correct combinations after incorrect actions. The extended Rasch model, as described by Hessen (2011), showed a good fit, which indicates that the combined score of the variables that indicate the position and connection of the work products' parts is a sufficient statistic to reflect a pupil's skill to restore the system. Moreover, the test-retest reliability was high.

The third challenge was to develop a tool that could compare differences in skill development for various systems. Following the suggestion of Seeney and Sterman (2007) and the work of Van der Steen (2014), we used the Fischer scale (Fischer, 1980) as the cornerstone for our diagnostic approach. Researchers familiar with Fischer's work were consulted to construct task-specific scoring rules for this scale. Independent comparative judgements of experts in primary technology education confirmed the validity of this approach.

Addressing these three major challenges with ECD resulted in a blueprint that describes ten requirements for single tasks that can be used to determine pupils' understanding of material-based systems. With two tasks designed by this blueprint, we demonstrated that the tool identifies skills that are not reflected by reading comprehension and mathematics skills. Moreover, our findings were in line with those of Molnár et al. (2013), who demonstrated that there are considerable differences in a pupil's understanding of the various material-based systems used in primary school engineering lessons.

Teacher judgement accuracy of technical abilities in primary education

The product requirements that guided the design of the diagnostic tool were based on assumptions about what should be suitable for classroom use. The two tasks that were developed in accordance with the blueprint made it possible to check these assumptions. In the second study described in **Chapter 3**, we asked teachers to use the tool in their classrooms and to reflect on the feasibility of its use and the resulting findings. The teachers with pupils of nine years or older were positive about the limited time needed for the introduction of the tasks and test-taking and the possibility to organise this flexibly at a convenient time (Heitink et al., 2016; Sach, 2015). Teachers were also positive about the benefits of the tool. In almost every class, they noticed that there were some pupils who performed much better than

they had expected. Sometimes the results confirmed the teachers' intuition about a pupil's capabilities, which standard tests did not reveal. Moreover, the teachers recognised the potential of the tool to improve learning and instruction. Using the tool also identified that especially novice teachers tend to understimate their pupils' technical skills and all teachers tend to underestimate the technical skills of girls.

Fostering pre-service primary school teachers' ability to recognise differences in pupils' understanding of technical systems

In the study in Chapter 3, the teachers did not analyse their pupils' work products, as they were not familiar with the Fischer scale and its background. For independent use, teachers should be able to reliably analyse these work products and understand the underlying developmental scale and associated theory to diagnose their pupils' comprehension of technical systems. In **Chapter four**, we have described the results of a short training based on Nickerson's anchoring and adjustment model and Fischer's model of dynamic skill development. With this course, we tried to strengthen the ability of the participating prospective teachers to diagnose pupils' comprehension of technical systems.

The course used the diagnostic tasks to create awareness about the development of pupils' thinking. This developmental perspective was taken to make the course attractive for primary school teachers, who are often not very interested in engineering (De Vries et al., 2011), but very motivated to extend their capability to maximise their pupils' developmental opportunities.

The main result of the course was that the participants were better able to articulate pupils' thinking about technological systems. Most of them could also correctly apply the scoring rules of the diagnostic tool. These results align with other studies about training programs that aim to improve teachers' diagnostic abilities (Desimone et al., 2002; Wilson, 2013).

A remarkable finding was that the diagnostic abilities of the teachers did not relate to their technical knowledge, as research has repeatedly emphasised the importance of sufficient subject-matter knowledge (Hartell, Gumaelius, & Svärdh, 2015; Moreland & Jones, 2000; Pleasants & Olson, 2019; Rohaan et al., 2012; Utley et al., 2019). It may be that our comparative approach and the focus on pupils' thinking made the diagnoses less sensitive to differences in technical knowledge.

Another interesting finding was the course's considerable impact on the teachers' self-efficacy beliefs about technology education. Self-efficacy beliefs are an important predictor of a teacher's attitude toward technology education (Rohaan et al., 2012; Utley et al., 2019).

With the results of the studies in Chapter 3 and Chapter 4, we expect that the diagnostic tool will support the diagnostic ability of teachers. To which extent that ability may contribute to skill development was explored in chapter five.

Adapting task difficulty to pupils' prior knowledge about electric circuits: effects on perceived challenge adequacy and skill development

The study described in **chapter 5** compared an adaptive condition in which pupils started with tasks which difficulty matched their prior knowledge as measured by the diagnostic tool with (1) a non-adaptive yoked condition in which the task sequences resulting from the adaptive condition were randomly assigned, and (2) a high-challenge condition in which pupils got only the most difficult tasks, regardless their level of prior knowledge. The results were mixed. Starting with tasks at pupils' prior knowledge level was perceived by the pupils as an adequate challenge. Therefore, such an approach may be suitable to promote interest in engineering, which is an important objective of teaching engineering topics in primary education (Ozogul et al., 2017; Rohaan et al., 2010; Tuijl, Walma-van der Molen, & Grol, 2014). The effect of task selection on pupils' learning about electric circuits was less straightforward than expected.

Offering tasks of a lower level than pupils' prior knowledge level did not contribute to skill development. That may be stating the obvious, but this finding must be considered in the light of the systematic underestimation of pupils' technical abilities, especially for girls by especially novice teachers, as indicated by our study in Chapter 3. Such an underestimation could result in offering too easy tasks. For pupils with lower levels of prior knowledge, the results indicate that starting with tasks of the subsequent difficulty level seemed to provide the optimal challenge for learning, whereas this did not seem to be beneficial for pupils with a high level of prior knowledge.

General discussion

In this section, we will reflect on the main research question and discuss the value of using the Fischer scale in engineering education.

Using the tool for engineering activities in primary schools.

Our main research question was: "Which support can a diagnostic tool designed for classroom use offer teachers to infer and promote pupils' engineering skills?". The first study described the design process of the tool. The characteristics that made the tool valid, reliable and suitable for classroom use were: a) the tool is performance-based, b) a task is the reconstruction of a material-based system, c) each task allows for various combinations of appropriate and inappropriate actions, resulting in a wide range of work products, d) the possibilities to evaluate trial and error behaviour are restricted, e) the scoring rules indicate the various developmental phases of the Fischer scale and f) the tasks can be accomplished independently by pupils aged nine to twelve in a few minutes.

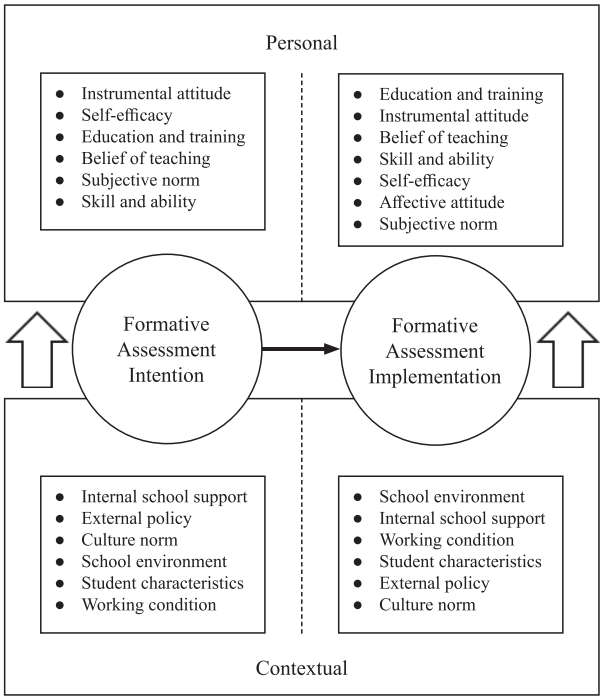
We have demonstrated that the tool can contribute to teachers' judgements of technical skills, for which underestimation seems to be more common than

overestimation, especially for girls. Its formative use demonstrated that the tool enables the selection of tasks that are most appreciated by pupils. However, it is known that despite the clear advantages (Förster et al., 2018; Hondrich et al., 2016; Lee et al., 2020; Shute et al., 2007; Wiliam et al., 2004), the formative use of assessments in education is limited (Ahmedi, 2019; Hui et al., 2017; Sach, 2012).

Zi Yan et al. (2021) provided an overview of the factors that influence the use of formative assessment in classrooms. Their model (see Figure 25) identifies contextual and personal factors that affect teachers' intentions and implementation of formative assessment. We will use this model to analyse our tool's potential for use in primary education.

Figure 25

Factors influencing formative assessment



Note: The arrows in the model indicate that, according to the author, the contextual factors have an impact on the personal factors and that the Formative Assessment Intention has an impact on the Formative Assessment Implementation. For each box, the factors are ordered from top to bottom according to their estimated importance. From "A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment" by Zi Yan, Ziqi Li, Ernesto Panadero, Min Yang, Lan Yang & Hongling Lao, 2021, Assessment in Education: Principles, Policy & Practice, 28:3, p. 250. Reprinted with permission.

Contextual factors. Our study did not explore contextual factors, but they were mentioned in the interviews with teachers that used the tool in their classroom when asked about their opinion on future use. It, thus, can be concluded that contextual factors will have an impact on the tool's use. Zi Yan et al. (2021) identified school environment and internal school support as the two most important contextual factors. Therefore, it can be expected that our tool is most likely to be used at schools in which formative assessment is encouraged by school leaders and discussed by teachers (Moss et al., 2013). The likelihood that the tool will be used at such schools is enhanced when there is internal school support for the role of engineering in the curriculum (Hammack & Ivey, 2019). The schools that initiated this dissertation are examples of such schools. At schools that lack a supportive environment for formative assessment and engineering, the use of the tool will depend on a teacher's personal interest in engineering and formative assessment.

External educational policies exert a general influence on engineering in primary schools. On the one hand, there are initiatives, often supported by industry, to promote engineering activities to generate interest in engineering as a possibility for study and profession (Male & King, 2019; Masson et al., 2016; Valentine et al., 2022). On the other hand, external policies pressure schools to improve their mathematics and reading comprehension results, often with detrimental effects on time and effort spent on other subjects like engineering (Brill et al., 2018).

Teachers' intentions (indicated as 'personal' in Figure 25) to use formative assessment were found to be related to their *instrumental attitude*, self-efficacy and education and training. This dissertation exemplifies the prominent role that Yi Zan et al. attribute to instrumental attitude, which indicates the teachers' perceptions about the effectiveness of formative assessment. This instrumental attitude characterized the teachers that motivated the present study. They expected that the availability of assessment opportunities would enable them to improve the effectiveness of their education about technical phenomena. This instrumental attitude is not only important in the context of formative assessment. The teachers that used the tool in their classroom considered it especially valuable from a pedagogical perspective. It helped them acknowledge and communicate pupils' technical skills that were not identified with standard tests.

Education and training, and especially *self-efficacy*, are the other two most important factors influencing teachers' intentions to use formative assessment (Karaman & Şahin, 2017; Yan & Cheng, 2015). As demonstrated in the third study, the tool can also be used in the context of education and training to broaden teachers' perspectives on cognitive development. The model proposed by Yan et al. (2021) presents education and training and self-efficacy beliefs as two separate

factors. Our results suggest that it may be very difficult to distinguish between the impact of those factors as our training had a large effect on the participants' self-efficacy beliefs.

The factors most frequently reported as personal predictors for actual use are 'education and training, 'instrumental attitude', 'belief in teaching' and 'skill and ability'. *Belief in teaching* refers to a more general attitude toward formative assessment, which was not explored in our study. We have shown that *education and training* can indirectly promote the use of the tool by changing the teachers' self-efficacy beliefs about teaching engineering and can enhance the *skill and ability* of most teachers to use formative assessment. Teachers' diagnostic skills are considered to depend on their subject matter knowledge (Lyon et al., 2019; Shulman, 1987). It was interesting to notice that in our training, focusing on becoming aware of the pupils' perspective, the ability to use our tool properly was not related to subject matter knowledge. For formative use, limited technical knowledge might hinder teachers from linking the tool's outcomes with corresponding adaptive support (Fox-Turnbull, 2006; Ramaligela, 2021; Rohaan et al., 2012; Sanjosé & Otero, 2021). This, in turn, could affect the formative use of the tool. However, as indicated by the factor instrumental attitude, it might be that teachers will use the tool anyway as it provides additional insight into pupils' technical skills.

From an implementation perspective, it can be concluded that the characteristics of our tool further its application in primary education. It can be used in classrooms, and most teachers are capable of analysing the work products after a short course. The teachers value the objective data that the tool offers as it either confirms their presumptions or makes them aware of biases in their estimation of pupils' technical skills.

Using the Fischer scale for work products

This study made a choice to use the Fisher scale to estimate pupils' knowledge of technical systems. Although the Fischer scale has been used in the context of non-verbal actions (Parziale, 2002; Sun et al., 2016), its application has so far been limited to research on microgenetical variability based on verbal utterances. This study is probably the first that uses the scale in the context of formative assessment and primary education, where cognitive development is usually described using Bloom's taxonomy (Krathwohl, 2002). We will use the results of our studies to reflect on this choice.

The main advantage of using the Fischer scale in the context of engineering is its emphasis on interaction. Interaction also defines the function of material-based systems. There is an interesting parallel with the concept of intrinsic cognitive load, which is determined by the number of elements that must be processed

simultaneously, which in turn depends on the extent of element interactivity of the materials or tasks that must be learned (Van Merriënboer & Sweller, 2005). The central notion of element interactivity, combined with the source of information (e.g. observed or remembered), makes the Fischer scale suitable to determine the nominal difficulty of tasks as well as pupils' level of understanding. This combination makes it possible to estimate the functional difficulty of tasks, that is, the difficulty from the learner's perspective. This makes the Fischer scale suitable for exploring the effect of challenge on learning.

The Fischer scale has mainly been used to demonstrate the short-term variability in understanding, which is, according to the dynamic systems approach, a phenomenon that characterises the phase transitions in skill development (Meindertsma et al., 2014; Siegler, 1994; Thelen & Smith, 1994). Being aware of this short-term variability, it could be argued that a work product is no more than a random snapshot of a pupil's ability. The position taken in this dissertation is different. We consider a work product as the representation of a series of choices. The construction of the tasks allows these choices to vary. An important difference is that most other studies use the Fischer scale to code verbal utterances. Here each answer, each 'choice', can be coded by its structure and meaning. A single constructive action cannot be coded likewise as it relates to previous or subsequent actions. Another difference is that a constructive action impacts the environment differently than an explanation. As a result of an action, the environment changes physically, and it is known from research on learning dynamics that such a change can change pupils' notion of the situation, which will have an impact on the next choice (Lickliter, 2000; Thelen & Smith, 1994). Therefore, the level of a work product cannot be considered a coincidental snapshot but should be considered a reflection of a pupil's ability resulting from a series of choices. Unpublished analyses of video recordings support this position. When in-between constructions were scaled on a timeline, the levels showed a, sometimes variable, increase towards the final product, which for almost all pupils represented the highest level of the sequence (see Supplementary materials, Chapter 6). Such a pattern was also shown by pupils building bridges (Parziale, 2002).

A drawback of using the Fischer scale in the context of education is that its use has been limited to developmental studies. Fischer's work and the associated theories about dynamic skill development are unknown among teachers and most teacher-trainers. This might hinder the use of the diagnostic tool in schools as its application requires additional training. On the other hand, we have shown that learning about the Fischer scale and its underlying theory about the dynamics of learning helps pre-service teachers to understand their pupils' behaviour from a developmental perspective and boosts their self-efficacy beliefs about teaching engineering.

Limitations and future directions

The idea for this dissertation originated in teachers' remarks about their dissatisfaction with the lack of possibilities to assess their pupils' progress in science and technology. The solution provided by the diagnostic tool described in this dissertation only solves a small part of this problem. First, the tool only applies to the engineering subdomain and, within this subdomain, to material-based systems. Also, the tool does not provide insight into pupils' problem-solving skills due to the decision to restrict feedback which was necessary to use the work product as an indicator of pupils' prior knowledge (Csapó & Funke, 2017; Jonassen, 2014).

Furthermore, the main purpose of the dissertation was to design a valid diagnostic tool that supports teachers in primary education in estimating pupils' prior knowledge about technical systems to enhance adaptive learning. After designing the tool, the follow-up studies explored the different aspects of this main purpose in isolation. Therefore, no conclusions can be drawn about teachers' independent use of the tool. That would require a study in which teachers use the tool, analyse the work products, use the results for adaptive teaching and evaluate the effect.

The diagnostic approach was explored by two tasks, each task for a different type of system. For more certainty about the validity of the approach, more tasks should be developed based on the set of requirements in the conceptual assessment framework. Preferably more than one task per system should be constructed to compare the reliability of the results for a specific type of system. On the other hand, it is known from developmental research on dynamic skill development that variability is an inherent aspect of skill development, especially in transitional phases (Thelen & Smith, 1994). Development should therefore be considered as a change in bandwidth and not as a transfer from one to the next level. This implies that a level inferred from the tasks should not be interpreted as a fact. Such an interpretation would deny the dynamic aspect of skill development, and the diagnostic tasks' results should be considered within that context.

A major limitation of the study in Chapter 5 was that adaptivity measures were restricted to task selection. Therefore, the full scope of benefits that may arise from using the tool by teachers (e.g., providing adapted instruction and feedback) is still unclear. It would be interesting to explore the effects on pupils' learning when task selection based on the diagnostic tool, which would reduce the bias indicated in Chapter 3, is combined with adaptive feedback. The results from the diagnostic tool can also be used for homogeneous ability grouping, as working with hands-on tasks is usually done in groups in educational practice and pupils seem to benefit most from homogenous grouping (Lou et al., 1996; So & Agbayewa, 2011)

The diagnostic tool offers new opportunities for research on the effect of hands-on versus instruction-based teaching. For instance, replication of Matlen and Klahr's research on sequential effects of high and low instructional guidance could place their findings about the effectiveness of instruction versus experimentation in a broader perspective. The tool may also contribute to research about the effect of school policy with respect to a subject like engineering and the development of pupils' understanding of material-based systems. Finally, the tool could be used to compare learning from computer simulations with learning from material-based systems.

Implications

The main implication of this dissertation is the availability of a tool that can be used in upper-grade primary classrooms to estimate pupils' technical skills. We demonstrated that those skills might differ considerably from the skills pupils demonstrate on the usual verbal tests. In most classes of the second study, some pupils with low performance on reading comprehension and mathematics were among the best performers on the diagnostic tasks in our studies. Such insight might positively impact the teachers' expectations and these pupils' self-efficacy beliefs.

It has been argued that career decisions may have their roots in early childhood and that a lack of knowledge about technical abilities may result in excluding technology from career options (van Tuijl & Walma van der Molen, 2016). Acknowledging their technical ability by the teacher may encourage pupils to maintain technology or engineering as an option for future study and careers. And as the second study has shown, information about pupils' technical abilities will help teachers become aware of these abilities, especially regarding girls.

An intriguing finding of the fourth study was that offering an optimal challenge as experienced by pupils did not cohere with maximal skill development. That might implicate that other adaptive approaches have to be considered when the primary goal is to motivate pupils than when the primary goal is to develop their engineering skills.

General Conclusion

The studies in this dissertation have created and explored new opportunities to assess pupils' understanding of material-based systems in the upper grades of primary education. Our diagnostic tool can support primary school teachers in the upper grades of primary education in their ability to infer and promote pupils' engineering skills. The tool also opens up new opportunities to study the effect of challenge on learning in the context of material-based systems.

REFERENCES

Admissiontests PABO. https://www.cito.nl/onderwijs/hoger-onderwijs/pabo-toelatingstoetsen.

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature, 387*, 401-406. https://doi.org/10.1038/387401a0

Ahmedi, V. (2019). Teachers' attitudes and practices towards formative assessment in primary schools. *Journal of Social Studies Education Research, 10*(3), 161-175. https://www.learntechlib.org/p/216460

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1-18. https://doi.org/10.1037/a0021017; 10.1037/a0021017.supp

Alsadoon, E. (2020). The impact of an adaptive e-course on students' achievements based on the students' prior knowledge. *Education and Information Technologies, 25*(5), 3541-3551. https://doi.org/10.1007/s10639-020-10125-3

Altrichter, H. (2006). Curriculum implementation–limiting and facilitating. *Making it Relevant: Context Based Learning of Science,* 35.

Arcia, G., Macdonald, K., Patrinos, H. A., & Porta, E. (2011). *School autonomy and accountability. SABER.* World Bank. https://openknowledge.worldbank.org/handle/10986/21546

Assaraf, O. B., & Orion, N. (2010). System thinking skills at the elementary school level. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching,* 47(5), 540-563. https://doi.org/10.1002/tea.20351

Atkinson, R. C. (1972). Ingredients for a theory of instruction. *American Psychologist,* 27(10), 921. https://psycnet.apa.org/doi/10.1037/h0033572

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research,* 70(2), 181-214. https://doi.org/10.3102/00346543070002181

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). *Educational psychology: A cognitive view* Holt, Rinehart and Winston.

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition,* 20(3), 287-298. https://doi.org/10.1002/acp.1245

Baird, J. R., & Penna, C. (1996). Challenge in learning and teaching science. *Research in Science Education,* 26, 257-269. https://doi.org/10.1007/BF02356938

Balali, M., VaezMousavi, M., Ghasemi, A., & Parvinpour, S. (2019). Effects of challenging games on manipulative motor skills of 4–6 years old children: An application of challenge point framework. *Early Child Development and Care,* 189 (5), 697-706. https://doi.org/10.1080/03004430.2017.1339276

Barlex, D., Givens, N., & Steeg, T. (2017). The curriculum. *A design and technology perspective on the ofsted curriculum survey*. https://dandtfordandt.files.wordpress.com/2017/08/wp-a-design-and-technology-perspective-on-the-ofsted-curriculum-survey2.pdf

Bassano, D., & Van Geert, P. (2007). Modeling continuity and discontinuity in utterance length: A quantitative approach to changes, transitions and intra-individual variability in early grammatical development. *Developmental Science, 10*(5), 588-612. https://doi.org/10.1111/j.1467-7687.2007.00629.x

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*(2), 177-187. https://doi.org/10.1080/01443410020043878

Baudson, T. G., Fischbach, A., & Preckel, F. (2016). Teacher judgments as measures of children's cognitive ability: A multilevel analysis. *Learning and Individual Differences, 52,* 148-156. https://doi.org/10.1016/j.lindif.2014.06.001

Baumert, J., Evans, R. H., & Geiser, H. (1998). Technical problem solving among 10-year-old students as related to science achievement, out-of-school experience, domain-specific control beliefs, and attribution patterns. *Journal of Research in Science Teaching, 35*(9), 987-1013. https://doi.org/10.1002/(SICI)1098-2736(199811)35:93.0.CO;2-P

Behrmann, L., & Souvignier, E. (2013). The relation between teachers' diagnostic sensitivity, their instructional activities, and their students' achievement gains in reading. *Zeitschrift Für Pädagogische Psychologie/German Journal of Educational Psychology, 27,* 283-293. https://doi.org/10.1024/1010-0652/a000112

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551-575. https://doi.org/10.1080/0969594X.2018.1441807

Bleicher, R. E. (2004). Revisiting the STEBI-B: Measuring self-efficacy in preservice elementary teachers. *School Science and Mathematics, 104*(8), 383-391. https://doi.org/10.1111/j.1949-8594.2004.tb18004.x

Bootsma, J. M., Hortobágyi, T., Rothwell, J. C., & Caljouw, S. R. (2018). The role of task difficulty in learning a visuomotor skill. *Medicine & Science in Sports & Exercise, 50*(9), 1842-1849. https://doi.org/10.1249/MSS.0000000000001635

Borrego, M., & Henderson, C. (2014). Increasing the use of evidence-based teaching in STEM higher education: A comparison of eight change strategies. *Journal of Engineering Education, 103*(2), 220-252. https://doi.org/10.1002/jee.20040

Box, C., Skoog, G., & Dabbs, J. M. (2015). A case study of teacher personal practice assessment theories and complexities of implementing formative assessment. *American Educational Research Journal, 52*(5), 956-983. https://doi.org/10.3102/0002831215587754

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). How people learn. National Academic Press.

Brill, F., Grayson, H., Kuhn, L., & O'Donnell, S. (2018). *What impact does accountability have on curriculum, standards and engagement in education? A literature review.* National Foundation for Educational Research.

Brophy, S., Klein, S., Portsmore, M., & Rogers, C. (2008). Advancing engineering education in P-12 classrooms. *Journal of Engineering Education, 97*(3), 369-387. doi:10.1002/j.2168-9830.2008.tb00985.x

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95-108. https://doi.org/10.1016/j.learninstruc.2009.11.004

Bumbacher, E., Salehi, S., Wieman, C., & Blikstein, P. (2018). Tools for science inquiry learning: Tool affordances, experimentation strategies, and conceptual understanding. *Journal of Science Education and Technology, 27*(3), 215-235. https://doi.org/10.1007/s10956-017-9719-8

Butler, R. (2012). Striving to connect: Extending an achievement goal approach to teacher motivation to include relational goals for teaching. *Journal of Educational Psychology, 104*(3), 726. https://psycnet.apa.org/doi/10.1037/a0028613

Carr, R. L., Bennett IV, L. D., & Strobel, J. (2012). Engineering in the K-12 STEM standards of the 50 US states: An analysis of presence and extent. *Journal of Engineering Education, 101*(3), 539-564. https://doi.org/10.1002/j.2168-9830.2012.tb00061.x

Catrysse, L., Gijbels, D., Donche, V., De Maeyer, S., Van den Bossche, P., & Gommers, L. (2016). Mapping processing strategies in learning from expository tekst: An exploratory eye tracking study followed by a cued recall. *Frontline Learning Research, 4*(1), 1-16. http://dx.doi.org/10.14786/flr.v4i1.192

Chandler, J., Fontenot, A. D., & Tate, D. (2011). Problems associated with a lack of cohesive policy in K-12 pre-college engineering. *Journal of Pre-College Engineering Education Research (J-PEER), 1*(1), 5. https://doi.org/10.7771/2157-9288.1029

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology, 15*(2), 181-195. https://doi.org/10.1207/S15326969EC01502_5

Chen, O., Retnowati, E., & Kalyuga, S. (2019). Effects of worked examples on step performance in solving complex problems. *Educational Psychology, 39*(2), 188-202. https://doi.org/10.1080/01443410.2018.1515891

Chien, S., Wu, H., & Hsu, Y. (2014). An investigation of teachers' beliefs and their use of technology-based assessments. *Computers in Human Behavior, 31,* 198-210. https://doi.org/10.1016/j.chb.2013.10.037

Chiu, M., & Lin, J. (2005). Promoting fourth graders' conceptual change of their understanding of electric current via multiple analogies. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 42*(4), 429-464. https://doi.org/10.1002/tea.20062

Church, R. M. (1989). The yoked control design. *Aversion, Avoidance, and Anxiety: Perspectives on Aversively Motivated Behavior,* 403-415. Routledge.

Cianciolo, A. T., Matthew, C., Stenberg, R. J., & Wagner, R. K. (2006). Tacit knowledge, practical intelligence, and expertise. *Handbook of expertise and expert performance* (pp. 613-632) https://psycnet.apa.org/doi/10.1017/CBO9780511816796.035

CITO. (2016). Natuur en techniek, technisch rapport over resultaten peil.onderwijs in 2015 [technical report on the results of the 2015 grade 6 survey on science and technology]. Retrieved from https://www.onderwijsinspectie.nl/onderwerpen/peil-onderwijs/documenten/rapporten/2017/05/31/peil-natuur-en-techniek-technisch-rapport-cito

Clarke-Midura, J., Silvis, D., Shumway, J. F., Lee, V. R., & Kozlowski, J. S. (2021). Developing a kindergarten computational thinking assessment using evidence-centered design: The case of algorithmic thinking. *Computer Science Education, 31:2,* 117-140. https://doi.org/10.1080/08993408.2021.1877988

Compton, V., & Harwood, C. (2005). Progression in technology education in new zealand: Components of practice as a way forward. *International Journal of Technology and Design Education, 15*(3), 253-287. http://dx.doi.org/10.1007%2Fs10798-004-5401-6

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*(4), 347. https://psycnet.apa.org/doi/10.1037/0022-0663.79.4.347

Corbalan, G., Kester, L., & Van Merriënboer, J. J. (2008). Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology, 33*(4), 733-756. https://doi.org/10.1016/j.cedpsych.2008.02.003

Csapó, B., & Funke, J. (2017). *The development and assessment of problem solving in 21st-century schools*. OECD Publishing. https://dx.doi.org/10.1787/9789264273955-en

Culver, D. E. (2012). *A qualitative assessment of preservice elementary teachers' formative perceptions regarding engineering and K-12 engineering education*. Doctoral dissertation Iowa State University

Cunningham, C. M., & Kelly, G. J. (2017). Epistemic practices of engineering for education. *Science Education, 101*(3), 486-505. https://doi.org/10.1002/sce.21271

Custodero, L. A. (2003). Perspectives on challenge: A longitudinal investigation of children's music learning. *Arts and Learning Research,* 23.

Dalgarno, B., Kennedy, G., & Bennett, S. (2014). The impact of students' exploration strategies on discovery learning using computer-based simulations. *Educational Media International, 51*(4), 310-329. https://doi.org/10.1080/09523987.2014.977009

Danish, J., Saleh, A., Andrade, A., & Bryan, B. (2017). Observing complex systems thinking in the zone of proximal development. *Instructional Science, 45*(1), 5-24. https://doi.org/10.1007/s11251-016-9391-z

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). Psychometric considerations for the next generation of performance assessment. *Center for K-12 Assessment & Performance Management, Educational Testing Service,* 1-100

De Boer, H., Bosker, R. J., & van der Werf, M. P. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*(1), 168. https://doi.org/10.1037/a0017289

De Grip, A., & Willems, E. (2003). Youngsters and technology. *Research Policy, 32*(10), 1771-1781. https://doi.org/10.1016/S0048-7333(03)00079-9

De Vries, M. J., Van Keulen, H., Peters, S., & Walma van der Molen, J. H. (Eds.). (2011). *Professional development for primary teachers in science and technology. the dutch VTB-pro project in an international perspective*. Sense.

De Vries, M. J. (2005). *Teaching about technology: An introduction to the philosophy of technology for non-philosophers*. Springer.

Defeyter, M. A., & German, T. P. (2003). Acquiring an understanding of design: Evidence from children's insight problem solving. *Cognition, 89*(2), 133-155. https://dx.doi.org.ezproxy.elib10.ub.unimaas.nl/10.1016/S0010-0277(03)00098-2

Department for Education. (2013). *The national curriculum in England*. Crown.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review, 1*(4), 346-357. https://doi.org/10.1207/s15327957pspr0104_5

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*(2), 81-112. https://doi.org/10.3102/01623737024002081

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31-54. https://doi.org/10.1016/j.edurev.2018.02.002

Dewey, J. (1986) Experience and education. In *The Educational Forum, 50*(3) 241-252. https://doi.org/10.1080/00131728609335764

Dochy, F., Moerkerke, G., & Martens, R. (1996). Integrating assessment, learning and instruction: Assessment of domain-specific and domaintranscending prior knowledge and progress. *Studies in educational evaluation, 22*(4), 309-339. https://doi-org.proxy.library.uu.nl/10.1016/0191-491X(96)00018-1

Dompnier, B., Pansu, P., & Bressoux, P. (2006). An integrative model of scholastic judgments: Pupils' characteristics, class context, halo effect and internal attributions. *European Journal of Psychology of Education, 21*(2), 119-133. https://doi.org/10.1007/BF03173572

Doyle, A., Seery, N., Gumaelius, L., Canty, D., & Hartell, E. (2019) Reconceptualising PCK research in D&T education: Proposing a methodological framework to investigate enacted practice. *International Journal of Technology and Design Education, 29*(3), 473-491. https://doi.org/10.1007/s10798-018-9456-1

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction, 24,* 58-61. https://doi.org/10.1016/j.learninstruc.2012.05.002

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence'. *Language Testing, 21*(3), 259-283. https://doi.org/I:10.1191/0265532204lt284oa

Edelman, G. M., & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination*. Basic books.

Fahrman, B., Norström, P., Gumaelius, L., & Skogh, I. (2020). Experienced technology teachers' teaching practices. *International Journal of Technology and Design Education, 30*(1), 163-186. https://doi.org/10.1007/s10798-019-09494-9

Feenstra, H., Kamphuis, F., Kleintjes, F., & Krom, R. (2010). *Scientific justification of reading comprehension tests for grades 3 to 6*. Arnhem: CITO. https://www.cito.nl/-/media/files/kennisbank/cito-bv/103-cito_lvs-rekenen_basisbewerkingen-gr-3-tm-8_wet-verantwoording.pdf?la=nl-NL

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*(1), 52. https://psycnet.apa.org/doi/10.1521/scpq.18.1.52.20876

Feldon, D. F., Callan, G., Juth, S., & Jeong, S. (2019). Cognitive load as motivational cost. *Educational Psychology Review, 31*(2), 319-337. https://doi.org/10.1007/s10648-019-09464-6

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477. https://psycnet.apa.org/doi/10.1037/0033-295X.87.6.477

Fischer, K. W., & Bidell, T. R. (2007). Dynamic development of action and thought. *Handbook of Child Psychology,* https://doi.org/10.1002/9780470147658.chpsy0107

Flores, R., Ari, F., Inan, F. A., & Arslan-Ari, I. (2012). The impact of adapting content for students with individual differences. *Journal of Educational Technology & Society, 15*(3), 251-261. https://www.jstor.org/stable/jeductechsoci.15.3.251

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*(3), 210-221. https://doi.org/10.5408/14-063.1

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short-and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learning and Instruction, 56,* 98-109. https://doi.org/10.1016/j.learninstruc.2018.04.009

Fox-Turnbull, W. (2006). The influences of teacher knowledge and authentic formative assessment on student learning in technology education. *International Journal of Technology and Design Education, 16*(1), 53-77. https://doi.org/10.1007/s10798-005-2109-1

Garmine, E., & Pearson, G. (Eds.). (2006). Tech tally; *approaches to assessing technological literacy*. National Acdemic Press.

Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2017). Assessment quality in tertiary education: An integrative literature review. *Studies in Educational Evaluation, 55,* 94-116. https://doi.org/10.1016/j.stueduc.2017.08.001

Gibson, J. J. (1977). The theory of affordances. *Hilldale,* USA, 1(2), 67-82

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box'differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055-1068. https://doi.org/10.1111/medu.12546

Ginns, I. S., Norton, S. J., & McRobbie, C. J. (2005). Adding value to the teaching and learning of design and technology. *International Journal of Technology and Design Education, 15*(1), 47-60. http://dx.doi.org/10.1007%2Fs10798-004-6193-4

Govaerts, M., Van de Wiel, M., Schuwirth, L., Van der Vleuten, C., & Muijtjens, A. (2013). Workplace-based assessment: Raters' performance theories and constructs. Ad*vances in Health Sciences Education, 18*(3), 375-396. https://doi.org/10.1007/s10459-012-9376-x

Graney, S. B. (2008). General education teacher judgments of their low-performing students' short-term reading progress. *Psychology in the Schools, 45*(6), 537-549. https://doi.org/10.1002/pits.20322

Granott, N., Fischer, K. W., & Parziale, J. (2002). Bridging to the unknown: A transition mechanism in learning and development. Microdevelopment: *Transition Processes in Development and Learning,* 131-156.

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*(3), 356-382. https://doi.org/10.1080/13546783.2014.989263

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review, 13,* 74-83. https://doi.org/10.1016/j.edurev.2014.10.002

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior, 36*(2), 212-224. https://doi.org/10.3200/JMBR.36.2.212-224

Gumaelius, L., Hartell, E., Svärdh, J., Skogh, I., & Buckley, J. (2019). Outcome analyses of educational interventions: A case study of the swedish "Boost of technology" intervention. *International Journal of Technology and Design Education, 29*(4), 739-758. https://doi.org/10.1007/s10798-018-9470-3

Guzey, S. S., Tank, K., Wang, H., Roehrig, G., & Moore, T. (2014). A high-quality professional development for teachers of grades 3–6 for implementing engineering into classrooms. *School Science and Mathematics, 114*(3), 139-149. https://doi.org/10.1111/ssm.12061

Hammack, R., & Ivey, T. (2017). Elementary teachers' perceptions of engineering and engineering design. *Journal of Research in STEM Education, 3*(1/2), 48-68. https://doi.org/10.51355/jstem.2017.29

Hammack, R., & Ivey, T. (2019). Elementary teachers' perceptions of K-5 engineering education and perceived barriers to implementation. *Journal of Engineering Education, 108*(4), 503-522. https://doi.org/10.1002/jee.20289

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online, 11*(2), 169-191. https://doi.org//10.25656/01:18004

Harlen, W. (2008). *Science as a key component of the primary curriculum: A rationale with policy implications.* (No. 1). www.wellcome.ac.uk/perspectives. (science primary education)

Harlen, W., Bell, D., Devés, R., Dyasi, H., de la Garza, G. F., Léna, P., & Yu, W. (2012). Developing policy, principles and practice in primary school science assessment. *The Nuffield Foundation*.

Hartell, E., Gumaelius, L., & Svärdh, J. (2015). Investigating technology teachers' self-efficacy on assessment. *International Journal of Technology and Design Education, 25*(3), 321-337. http://dx.doi.org/10.1007%2Fs10798-014-9285-9

Hast, M. (2020). "It is There but You Need to Dig a Little Deeper for It to Become Evident to Them": Tacit Knowledge Assessment in the Primary Science Classroom. In: Koh, C. (eds) *Diversifying Learner Experience*. Springer. https://doi.org/10.1007/978-981-15-9861-6_2

Hattie, J. (2013). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hedlund, J., Antonakis, J., & Sternberg, R. J. (2002). Tacit knowledge and practical intelligence: Understanding the lessons of experience. DTIC Document.

Heitink, M. C., Van der Kleij, Fabienne 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

Helmke, A., & Schrader, F. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3*(2), 91-98. https://doi.org/10.1016/0742-051X(87)90010-2

Hessen, D. J. (2011). Loglinear representations of multivariate bernoulli rasch models. *British Journal of Mathematical and Statistical Psychology, 64*(2), 337-354. https://doi.org/10.1348/2044-8317.002000

Hodges, N. J., & Lohse, K. R. (2020). Difficulty is a real challenge: A perspective on the role of cognitive effort in motor skill learning. *Journal of Applied Research in Memory and Cognition, 9*(4), 455-460. https://psycnet.apa.org/doi/10.1016/j.jarmac.2020.08.006

Hoffmann, L. (2002). Promoting girls' interest and achievement in physics classes for beginners. *Learning and Instruction, 12*(4), 447-465. https://doi.org/10.1016/S0959-4752(01)00010-X

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297-313. https://doi.org/10.3102/00346543059003297

Hondrich, A. L., Hertel, S., Adl-Amini, K., & Klieme, E. (2016). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assessment in Education: Principles, Policy & Practice, 23*(3), 353-376. https://doi.org/10.1080/0969594X.2015.1049113

Honey, M., Pearson, G., & Schweingruber, H. (2014). Committee on integrated STEM education, national academy of engineering, national research council. *STEM integration in K-12 education: Status, prospects, and an agenda for research*. National Academic Press.

Honomichl, R. D., & Chen, Z. (2012). The role of guidance in children's discovery learning. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(6), 615-622. https://doi.org/10.1002/wcs.1199

Hornstra, L., Bakx, A., Mathijssen, S., & Denissen, J. J. (2020). Motivating gifted and non-gifted students in regular primary schools: A self-determination perspective. *Learning and Individual Differences, 80,* 101871. https://doi.org/10.1016/j.lindif.2020.101871

Hourigan, M., O'Dwyer, A., Leavy, A. M., & Corry, E. (2022). Integrated STEM–a step too far in primary education contexts?. *Irish Educational Studies, 41*(4), 687-711. https://doi.org/10.1080/03323315.2021.1899027

Hsu, M., Purzer, S., & Cardella, M. E. (2011). Elementary teachers' views about teaching design, engineering, and technology. *Journal of Pre-College Engineering Education Research (J-PEER), 1*(2), 5. https://doi.org/10.5703/1288284314639

Hui, S. K. F., Brown, G. T., & Chan, S. W. M. (2017). Assessment for learning and for accountability in classrooms: The experience of four hong kong primary school curriculum leaders. *Asia Pacific Education Review, 18*(1), 41-51. https://doi.org/10.1007/s12564-017-9469-6

Inzlicht, M., & Campbell, A. V. (2022). Effort feels meaningful. *Trends in Cognitive Sciences,* https://doi.org/10.1016/j.tics.2022.09.016

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences, 22*(4), 337-349. https://doi.org/10.1016/j.tics.2018.01.007

ITEEA (2007). *Standards for technological literacy: Content for the study of technology* (Third Edition). International Technology and Engineering Educators Association.

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Scientific accounting of the LOVS mathematics tests for grades 3 through 8.* Arnhem: CITO. https://docplayer.nl/11108466-Wetenschappelijke-verantwoording-van-de-toetsen-lovs-rekenen-wiskunde-voor-groep-3-tot-en-met-8-j-janssen-n-verhelst-r-engelen-en-f.html

Johnson, S. D. (1995). Understanding troubleshooting styles to improve training methods.

Jonassen, D. H. (2010). *Learning to solve problems: A handbook for designing problem-solving learning environments.* Routledge.

Jonassen, D. H. (2014). Assessing problem solving. *Handbook of research on educational communications and technology* (pp. 269-288) Springer.

Jonassen, D. H., & Hung, W. (2006). Learning to troubleshoot: A new theory-based design architecture. *Educational Psychology Review, 18*(1), 77-114. https://doi.org/10.1007/s10648-006-9001-8

Jones, A., & Compton, V. (1998). Towards a model for teacher development in technology education: From research to practice. *International Journal of Technology and Design Education, 8*(1), 51-65. https://doi.org/10.1023/A:1008891628375

Jones, A., & Moreland, J. (2004). Enhancing practicing primary school teachers' pedagogical content knowledge in technology. *International Journal of Technology and Design Education, 14*(2), 121-140. https://doi.org/10.1023/B:ITDE.0000026513.48316.39

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum zusammenhang von intelligenz und urteilsgenauigkeit bei der beurteilung von schülerleistungen im simulierten klassenraum. *Zeitschrift Für Pädagogische Psychologie,* https://doi.org/10.1024/1010-0652/a000076

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28,* 73-84. https://psycnet.apa.org/doi/10.1016/j.learninstruc.2013.06.001

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development, 53*(3), 83-93. https://doi.org/10.1007/BF02504800

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170. https://doi-org.proxy.library.uu.nl/10.1207/s15366359mea0203_1

Karaman, P., & Şahin, Ç. (2017). Adaptation of teachers' conceptions and practices of formative assessment scale into turkish culture and a structural equation modeling. *International Electronic Journal of Elementary Education, 10*(2), 185-194. https://www.iejee.com/index.php/IEJEE/article/view/320

Kärkkäinen, K., & Vincent-Lancrin, S. (2013). *Sparking innovation in STEM education with technology and collaboration: A case study of the HP catalyst initiative*. OECD Publishing. doi://dx.doi.org/10.1787/5k480sj9k442-en

Kelley, T. R. (2009). Using engineering cases in technology education. *Technology Teacher, 68*(7), 5-9.

Kimbell, R. (1997). *Assessing technology: International trends in curriculum and assessment: UK, germany, USA, taiwan, australia* McGraw-Hill Education.

Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching, 50*(5), 530-560. https://doi.org/10.1002/tea.21086

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86. https://doi.org/10.1207/s15326985ep4102_1

Kirton, A., Hallam, S., Peffers, J., Robertson, P., & Stobart, G. (2007). Revolution, evolution or a trojan horse? piloting assessment for learning in some scottish primary schools. *British Educational Research Journal, 33*(4), 605-627. https://doi.org/10.1080/01411920701434136

Klahr, D., & Robinson, M. (1981). Formal assessment of problem-solving and planning processes in preschool children. *Cognitive Psychology, 13*(1), 113-148. https://doi.org/10.1016/0010-0285(81)90006-2

Kloke, J. D., & McKean, J. W. (2012). Rfit: Rank-based estimation for linear models. *R J., 4*(2), 57.

Knezek, G., Christensen, R., & Tyler-Wood, T. (2011). Contrasts in teacher and student perceptions of STEM content and careers. *Contemporary Issues in Technology and Teacher Education, 11*(1), 92-117. https://www.learntechlib.org/primary/p/35400/

Koski, M., & de Vries, M. J. (2013). An exploratory study on how primary pupils approach systems. *International Journal of Technology & Design Education, 23*(4), 835-848. https://doi.org/10.1007/s10798-013-9234-z

Kramer, M., Förtsch, C., Boone, W. J., Seidel, T., & Neuhaus, B. J. (2021). Investigating pre-service biology teachers' diagnostic competences: Relationships between professional knowledge, diagnostic activities, and diagnostic accuracy. *Education Sciences, 11*(3), 89. https://doi.org/10.3390/educsci11030089

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Kuldas, S., Hashim, S., Ismail, H. N., & Abu Bakar, Z. (2015). Reviewing the role of cognitive load, expertise level, motivation, and unconscious processing in working memory performance. *International Journal of Educational Psychology, 4*(2), 142-169. https://doi.org/10.17583/ijep.2015.832

Kuldas, S., Satyen, L., Ismail, H. N., & Hashim, S. (2014). Greater cognitive effort for better learning: Tailoring an instructional design for learners with different levels of knowledge and motivation. *Psychologica Belgica, 54*(4), 350-373. https://psycnet.apa.org/doi/10.5334/pb.aw

Lachapelle, C. P., & Cunningham, C. M. (2014). Engineering in elementary schools. In Engineering in Pre-College Settings: *Synthesizing Research, Policy, and Practices* (pp. 61-88). Purdue University Press.

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

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: A systematic review. *Applied Measurement in Education, 33*(2), 124-140. | https://doi.org/10.1080/08957347.2020.1732383

Lee, S. W., Min, S., & Mamerow, G. P. (2015). Pygmalion in the classroom and the home: Expectation's role in the pipeline to STEMM. *Teachers College Record, 117*(9), 1-40.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. *Innovative practices for higher education assessment and measurement* (pp. 119-138) IGI Global.

Levy, S. T. (2012). Young children's learning of water physics by constructing working systems. *International Journal of Technology and Design Education,* 1-30. https://doi.org/10.1007/s10798-012-9202-z

Lewis, P. A. (2019). Technicians and innovation: A literature review. https://dx.doi.org/10.2139/ssrn.3405406

Lickliter, R. (2000). An ecological approach to behavioral development: Insights from comparative psychology. *Ecological Psychology, 12*(4), 319-334. https://doi.org/10.1207/S15326969ECO1204_06

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiacoM). *Teaching and Teacher Education, 91,* 103059. https://doi.org/10.1016/j.tate.2020.103059

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of educational research, 66*(4), 423-458.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260-293. https://doi-org.proxy.library.uu.nl/10.3102%2F0162373719849044

Lyon, C. J., Nabors Oláh, L., & Caroline Wylie, E. (2019). Working toward integrated practice: Understanding the interaction among formative assessment strategies. *The Journal of Educational Research, 112*(3), 301-314. https://doi.org/10.1080/00220671.2018.1514359

Male, S. A., & King, R. (2019). Enhancing learning outcomes from industry engagement in australian engineering education. *Journal of Teaching and Learning for Graduate Employability, 10*(1), 101-117. https://search.informit.org/doi/10.3316/informit.580683621107131

Malik, X. (2014). *The future of europe is science*. Luxembourg: Publications Office of the European Union. https://doi.org/10.2796/28973

Masson, A., Klop, T., Osseweijer, P., & de Vries, M. J. (2016). The role of engineers in pre-university education: Success-factors and challenges. *Pre-university engineering education* (pp. 221-236). Brill.

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition, 78*(1), 1-26. https://doi.org/10.1016/S0010-0277(00)00094-9

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? *Instructional Science, 41*(3), 621-634. https://doi.org/10.1007/s11251-012-9248-z

McFadden, A., & Williams, K. E. (2020). Teachers as evaluators: *Results from a systematic literature review* https://doi-org.proxy.library.uu.nl/10.1016/j.stueduc.2019.100830

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psyc*hological Methods, 1*(1), 30-46. https://psycnet.apa.org/doi/10.1037/1082-989X.1.1.30

Meindertsma, H. B., Van Dijk, M. W., Steenbeek, H. W., & Van Geert, P. L. (2014a). Assessment of preschooler's scientific reasoning in Adult–Child interactions: What is the optimal context? *Research in Science Education, 44*(2), 215-237. https://doi.org/10.1007/s11165-013-9380-z

Meindertsma, H. B., Van Dijk, M. W., Steenbeek, H. W., & Van Geert, P. L. (2014b). Stability and variability in young children's understanding of floating and sinking during one Single-Task session. *Mind, Brain, and Education, 8*(3), 149-158. https://doi.org/10.1111/mbe.12049

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development, 50*(3), 43-59. https://doi.org/10.1007/BF02505024

Mislevy, R., & Riconscente, M. (2005). *Evidence-centered design* Menlo Park, CA: SRI International. https://search.credoreference.com/content/entry/sageermae/evidence_centered_design/0

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62. https://doi.org/10.1207/S15366359MEA0101_02

Mitcham, C. (1994). *Thinking through technology: The path between engineering and philosophy* University of Chicago Press.

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity, 9,* 35-45. https://doi.org/10.1016/j.tsc.2013.03.002

Moreland, J., & Jones, A. (2000). Emerging assessment practices in an emergent curriculum: Implications for technology. *International Journal of Technology and Design Education, 10*(3), 283-305. https://doi.org/10.1023/A:1008990307060

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364-386. https://psycnet.apa.org/doi/10.1177/1094428106291059

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. https://doi.org/10.1080/08957347.2013.793186

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 assessment frameworks* TIMMS & PIRLS International Study Center. Lynch School of Education, Boston College.

Nadelson, L. S., Callahan, J., Pyke, P., Hay, A., Dance, M., & Pfiester, J. (2013). Teacher STEM perception and preparation: Inquiry-based STEM professional development for elementary teachers. *The Journal of Educational Research, 106*(2), 157-168. https://doi.org/10.1080/00220671.2012.667014

National Assessment Governing Board. (2013). *Technology and engineering literacy framework for the 2014 national assessment of educational progress.* (No. ED08C00134). National Assessment Governing Board.

National Assessment of Educational Progress. (2018). NAEP technology & engineering literacy (TEL) report card. https://www.nationsreportcard.gov/tel

Nauta, M. M., Kahn, J. H., Angell, J. W., & Cantarelli, E. A. (2002). Identifying the antecedent in the relation between career interests and self-efficacy: Is it one, the other, or both? *Journal of Counseling Psychology, 49*(3), 290. https://psycnet.apa.org/doi/10.1037/0022-0167.49.3.290

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*(6), 737. https://psycnet.apa.org/doi/10.1037/0033-2909.125.6.737

Niiranen, S. (2021). Supporting the development of students' technological understanding in craft and technology education via the learning-by-doing approach. *International Journal of Technology and Design Education, 31*(1), 81-93. https://doi.org/10.1007/s10798-019-09546-0

Nitko, A. J. (1996). *Educational assessment of students* ERIC.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18. https://doi.org/10.1016/0022-2496(66)90002-2

O'Connell, B., de Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2016). Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment & Evaluation in Higher Education, 41*(3), 331-349. https://doi.org/10.1080/02602938.2015.1008398

OECD. (2013). *PISA 2012 assessment and analytical framework*. OECD Publishing. https://dx.doi.org/10.1787/9789264190511-en

OECD. (2014). *PISA 2012 technical report*. OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing,* 1-31. https://doi.org/10.1080/15305058.2018.1543308

Onla-Or, S., & Winstein, C. J. (2008). Determining the optimal challenge point for motor skill learning in adults with moderately severe parkinson's disease. *Neurorehabilitation and Neural Repair, 22*(4), 385-395. https://doi.org/10.1177/1545968307313508

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology, 62*(11), 1126-1137. https://doi.org/10.1016/j.jclinepi.2009.03.013

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior, 24*(5), 2415-2433. https://doi.org/10.1016/j.chb.2008.02.016

Ostermann, A., Leuders, T., & Nückles, M. (2018). Improving the judgment of task difficulties: Prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education, 21*(6), 579-605. https://doi.org/10.1007/s10857-017-9369-z

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76,* 214-226. https://doi.org/10.1016/j.tate.2018.02.007

Ozogul, G., Miller, C. F., & Reisslein, M. (2017). Latinx and caucasian elementary school children's knowledge of and interest in engineering activities. *Journal of Pre-College Engineering Education Research (J-PEER), 7*(2), 2. https://doi.org/10.7771/2157-9288.1122

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71. https://psycnet.apa.org/doi/10.1037/0022-0663.84.4.429

Parziale, J. (2002). Observing the dynamics of construction: Children building bridges and new ideas. *Microdevelopment: Transition Processes in Development and Learning,* 157-180. https://psycnet.apa.org/doi/10.1017/CBO9780511489709.007

Pearson, G., & Young, A. T. (2002). *Technically speaking: Why all americans need to know more about technology* National Academies Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81. https://doi.org/10.1080/00461520.2016.1145550

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. *International Encyclopedia of Education, 2,* 6452-6457.

Pesce, C., Crova, C., Marchetti, R., Struzzolino, I., Masci, I., Vannozzi, G., & Forte, R. (2013). Searching for cognitively optimal challenge point in physical activity for children with typical and atypical motor development. *Mental Health and Physical Activity, 6*(3), 172-180. https://doi.org/10.1016/j.mhpa.2013.07.001

Philpot, R., Ramalingam, D., Dossey, J. A., & Mccrae, B. (2017). *Factors that influence the difficulty of problem-solving items*. Paris: OECD Publishing. https://dx.doi.org/10.1787/9789264273955-en

Platform Bèta Techniek. (2013). *Advies verkenningscommissie wetenschap en technologie primair onderwijs [advice, exploratory committee science and technology primary education]*. Den Haag: Platform Bèta Techniek. ww.platformbetatechniek.nl

Pleasants, J., & Olson, J. K. (2019). Refining an instrument and studying elementary teachers' understanding of the scope of engineering. *Journal of Pre-College Engineering Education Research (J-PEER), 9*(2), 1. https://doi.org/10.7771/2157-9288.1207

Pleasants, J., Olson, J. K., & De La Cruz, I. (2020). Accuracy of elementary teachers' representations of the projects and processes of engineering: Results of a professional development program. *Journal of Science Teacher Education, 31*(4), 362-383. https://doi.org/10.1080/1046560X.2019.1709295

Plumm, K. M. (2008). Technology in the classroom: Burning the bridges to the gaps in gender-biased education? *Computers & Education, 50*(3), 1052-1068. https://doi.org/10.1016/j.compedu.2006.10.005

Potgieter, C. (2012). Linking learning activities and assessment activities to learning outcomes and assessment standards when teaching technology: A case study. *International Journal of Technology and Design Education, 22,* 1-18. https://doi.org/10.1007/s10798-012-9226-4

Praetorius, A., Koch, T., Scheunpflug, A., Zeinz, H., & Dresel, M. (2017). Identifying determinants of teachers' judgment (in) accuracy regarding students' school-related motivations using a bayesian cross-classified multi-level model. *Learning and Instruction, 52,* 148-160. https://doi.org/10.1016/j.learninstruc.2017.06.003

Priestley, M., & Philippou, S. (2018). Curriculum making as social practice: Complex webs of enactment. *The Curriculum Journal,* (29:2), 151-158. https://doi.org/10.1080/09585176.2018.1451096

Ramaligela, S. M. (2021). Exploring pre-service technology teachers' content and instructional knowledge to determine teaching readiness. *International Journal of Technology and Design Education, 31*(3), 531-544. https://doi.org/10.1007/s10798-020-09570-5

Rasinen, A., Virtanen, S., Endepohls-Ulpe, M., Ikonen, P., Ebach, J., & Stahl-von Zabern, J. (2009). Technology education for children in primary schools in finland and germany: Different school systems, similar problems and how to overcome them. *International Journal of Technology and Design Education, 19*(4), 367-379. https://doi.org/10.1007/s10798-009-9097-5

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335-360. https://doi.org/10.3102/0002831210374874

Reilly, D., Neumann, D. L., & Andrews, G. (2017). Gender differences in spatial ability: Implications for STEM education and approaches to reducing the gender gap for parents and educators. *Visual-spatial ability in STEM education* (pp. 195-224) Springer. https://doi.org/10.1007/978-3-319-44385-0_10

Renkl, A., Atkinson, R. K., & Große, C. S. (2004). How fading worked solution steps works–a cognitive load perspective. *Instructional Science, 32*(1), 59-82. https://doi.org/10.1023/B:TRUC.0000021815.74806.f6

Resh, N., & Benavot, A. (2009). Educational governance, school autonomy, and curriculum implementation: Diversity and uniformity in knowledge offerings to israeli pupils. *Journal of Curriculum Studies, 41*(1), 67-92. https://doi.org/10.1080/00220270802446826

Retnawati, H., Djidu, H., Kartianom, A., & Anazifa, R. D. (2018). Teachers' knowledge about higher-order thinking skills and its learning strategy. *Problems of Education in the 21st Century, 76*(2), 215. https://www.proquest.com/scholarly-journals/teachers-knowledge-about-higher-order-thinking/docview/2343794353/se-2.

Rey, G. D., & Buchwald, F. (2011). The expertise reversal effect: Cognitive load and motivational explanations. *Journal of Experimental Psychology: Applied, 17*(1), 33-48. https://psycnet.apa.org/doi/10.1037/a0022243

Riggs, I. M., & Enochs, L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education, 74*(6), 625-637.

Robinson, J., Myran, S., Strauss, R., & Reed, W. (2014). The impact of an alternative professional development model on teacher practices in formative assessment and student learning. *Teacher Development, 18*(2), 141-162. https://doi-org.proxy.library.uu.nl/10.1080/13664530.2014.900516

Rockwood, N. J. (2017). *Advancing the formulation and testing of multilevel mediation and moderated mediation models.* (Doctoral dissertation, The Ohio State University)

Roelofs, E. (2019). A framework for improving the accessibility of assessment tasks. *Theoretical and practical advances in computer-based educational measurement* (pp. 21-45) Springer. https://doi.org/10.1007/978-3-030-18480-3_2

Roelofs, E. C., Emons, W. H., & Verschoor, A. J. (2021). Exploring task features that predict psychometric quality of test items: The case for the dutch driving theory exam. *International Journal of Testing,* 1-25. https://doi.org/10.1080/15305058.2021.1916506

Rohaan, E. J., Taconis, R., & Jochems, W. M. (2009). Measuring teachers' pedagogical content knowledge in primary technology education. *Research in Science & Technological Education, 27*(3), 327-338. https://doi.org/10.1080/02635140903162652

Rohaan, E. J., Taconis, R., & Jochems, W. M. (2010). Reviewing the relations between teachers' knowledge and pupils' attitude in the field of primary technology education. *International Journal of Technology and Design Education, 20*(1), 15. https://doi.org/10.1007/s10798-008-9055-7

Rohaan, E. J., Taconis, R., & Jochems, W. M. (2012). Analysing teacher knowledge for technology education in primary schools. *International Journal of Technology and Design Education, 22*(3), 271-280. https://doi.org/10.1080/02635140903162652

Roschelle, J. (1997). Learning in interactive environments: *Prior knowledge and new experience* Citeseer.

Sach, E. (2012). Teachers and testing: An investigation into teachers' perceptions of formative assessment. *Educational Studies, 38*(3), 261-276. https://doi.org/10.1080/03055698.2011.598684

Sach, E. (2015). An exploration of teachers' narratives: What are the facilitators and constraints which promote or inhibit 'good'formative assessment practices in schools? *Education 3-13,* 43(3), 322-335. https://doi.org/10.1080/03004279.2013.813956

Safadi, R., & Yerushalmi, E. (2014). Problem solving vs. troubleshooting tasks: The case of sixth-grade students studying simple electric circuits. *International Journal of Science and Mathematics Education, 12*(6), 1341-1366. http://dx.doi.org/10.1007/2Fs10763-013-9461-5

Salden, R. J., Paas, F., Broers, N. J., & Van Merriënboer, J. J. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science, 32*(1), 153-172. https://doi.org/10.1023/B:TRUC.0000021814.03996.ff

Sanjosé, V., & Otero, J. (2021). Elementary pre-service teachers' conscious lack of knowledge about technical artefacts. *International Journal of Technology and Design Education,* 1-18. https://doi.org/10.1007/s10798-021-09696-0

Savelsbergh, E. R., de Jong, T., & Ferguson-Hessler, M. G. (2011). Choosing the right solution approach: The crucial role of situational knowledge in electricity and magnetism. *Physical Review Special Topics-Physics Education Research, 7*(1), 010103. https://doi.org/10.1103/PhysRevSTPER.7.010103

Scharten, R., & Kat-de Jong, M. (2012). *Koersvast en enthousiast. kritieke succesfactoren van gelderse vindplaatsen [enthusiastic and purposeful. what makes primary schools in gelderland succesful in their science and technology education]*. Nijmegen: Expertisecentrum Nederlands, Nijmegen.

Schipper, T., Goei, S. L., de Vries, S., & van Veen, K. (2018). Developing teachers' self-efficacy and adaptive teaching behaviour through lesson study. *International Journal of Educational Research, 88,* 109-120. https://doi.org/10.1016/j.ijer.2018.01.011

Schraagen, J. M. (2006). Task analysis. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. Hoffman R. (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 185-201). Cambridge University Press.

Schrader, F., & Helmke, A. (2001). Alltägliche leistungsbeurteilung durch lehrer. In F. E. Weinert (Ed.), *Leistungsmessungen in schulen,* 2, 45-58.

Schunk, D. H., Meece, J. R., & Pintrich, P. R. (2012). *Motivation in education: Theory, research, and applications*. Pearson Higher Ed.

Schwartz, M., & Fischer, K. W. (2004). Building general knowledge and skill: Cognition and microdevelopment in science learning. *Cognitive Developmental Change: Theories, Models, and Measurement,* 157-185.

Seiter, J. (2009). "Crafts and technology" and "technical education" in austria. *International Journal of Technology and Design Education, 19*(4), 419-429. https://doi.org/10.1007/s10798-009-9096-6

Shavelson, R. J. (1978). Teachers' estimates of student' states of mind and behavior. *Journal of Teacher Education, 29*(5), 37-40. https://doi.org/10.1177/002248717802900511

Shipstone, D. M., Rhöneck, C. v., Jung, W., Kärrqvist, C., Dupin, J., Johsua, S. E., & Licht, P. (1988). A study of students' understanding of electricity in five european countries. *International Journal of Science Education, 10*(3), 303-316. https://doi.org/10.1080/0950069880100306

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). An assessment for learning system called ACED: Designing for learning effectiveness and accessibility. *ETS Research Report Series, 2007*(2), i-45. https://doi.org/10.1002/j.2333-8504.2007.tb02068.x

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science, 3*(1), 1-5. https://psycnet.apa.org/doi/10.1111/1467-8721.ep10769817

Siuty, M. B., Leko, M. M., & Knackstedt, K. M. (2018). Unraveling the role of curriculum in teacher decision making. *Teacher Education and Special Education, 41*(1), 39-57. https://doi.org/10.1177/0888406416683230

Slangen, L., Van Keulen, H., & Gravemeijer, K. (2011). What pupils can learn from working with robotic direct manipulation environments. *International Journal of Technology and Design Education, 21*(4), 449-469. https://doi.org/10.1007/s10798-010-9130-8

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21. https://doi.org/10.3102/2F0013189X031007015

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology, 10,* 2366. https://doi.org/10.3389/fpsyg.2019.02366

So, A., & Agbayewa, J. O. (2011). Effect of homogenous and heterogeneous ability grouping class teaching on student's interest, attitude and achievement in integrated science. *International Journal of Psychology and Counselling, 3*(3), 48-54.

Solomonidou, C., & Kakana, D. (2000). Preschool children's conceptions about the electric current and the functioning of electric appliances. *European Early Childhood Education Research Journal, 8*(1), 95-111. https://doi.org/10.1080/13502930085208511

Stang, J. (2016). *Zur urteilsgenauigkeit von mathematiklehrkräften: Genauigkeitsbeeinflussende faktoren, stabilität und auswirkungen* (Doctoral dissertation, Universität Passau). https://nbn-resolving.org/urn:nbn:de:bvb:739-opus4-4646

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743. https://psycnet.apa.org/doi/10.1037/a0027627

Sun, H., Steinkrauss, R., Van Der Steen, S., Cox, R., & De Bot, K. (2016). Foreign language learning as a complex dynamic process: A microgenetic case study of a chinese child's english learning trajectory. *Learning and Individual Differences, 49,* 287-296. https://doi.org/10.1016/j.lindif.2016.05.010

Sutherland, L. (2002). Developing problem solving expertise: The impact of instruction in a question analysis strategy. *Learning and Instruction, 12*(2), 155-187. https://doi.org/10.1016/S0959-4752(01)00003-2

Svensson, M. (2018). Learning about systems. In M. J. De Vries (Ed.), *Handbook of technology education* (Handbook of Technology Education ed., pp. 447-462). Springer.

Svensson, M., Zetterqvist, A., & Ingerman, Å. (2012). On young people's experience of systems in technology. *Design & Technology Education, 17*(1). https://urn.kb.se/resolve?urn=urn%3Anbn%3Ase%3Aliu%3Adiva-63748

Swaak, J., & de Jong, T. (1996). Measuring intuitive knowledge in science: The development of the what-if test. *Studies in Educational Evaluation, 22*(4), 341-362.

Sweeney, L. B., & Sterman, J. D. (2007). Thinking about systems: Student and teacher conceptions of natural and social systems. *System Dynamics Review, 23*(2-3), 285-311. https://doi.org/10.1002/sdr.366

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*(1), 59-89. https://doi.org/10.1207/s1532690xci0201_3

Sweller, J., Kirschner, P. A., & Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist, 42*(2), 115-121. https://doi.org/10.1080/00461520701263426

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action* MIT press.

Thiede, K. W., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., & Osguthorpe, R. D. (2018). Improving the accuracy of teachers' judgments of student learning. *Teaching and Teacher Education, 76,* 106-115. https://doi.org/10.1016/j.tate.2018.08.004

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., . . . Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49,* 36-44. http://dx.doi.org/10.1016/j.tate.2015.01.012

Timmermans, A. C., Kuyper, H., & van der Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology, 85*(4), 459-478. https://doi.org/10.1111/bjep.12087

Timmermans, A. C., Rubie-Davies, C. M., & Rjosk, C. (2018). Pygmalion's 50th anniversary: The state of the art in teacher expectation research. *Educational Research and Evaluation, 24*(3-5), 91-98. https://doi.org/10.1080/13803611.2018.1548785

Tomesen, M., & Weekers, A. (2012). Aanvulling bij de wetenschappelijke verantwoording papieren toetsen begrijpend lezen voor groep 7 en 8: Digitale toetsen [supplement to the scientific justification for paper tests. reading comprehension for groups 7 and 8: Digital tests]. [Aanvulling bij de wetenschappelijke verantwoording papieren toetsen Begrijpend lezen voor groep 7 en 8: Digitale toetsen] *Arnhem:* CITO,

Turja, L., Endepohls-Ulpe, M., & Chatoney, M. (2009). A conceptual framework for developing the curriculum and delivery of technology education in early childhood. *International Journal of Technology and Design Education, 19*(4), 353-365. https://doi.org/10.1007/s10798-009-9093-9

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32,* 100374. https://doi.org/10.1016/j.edurev.2020.100374

Utley, J., Ivey, T., Hammack, R., & High, K. (2019). Enhancing engineering education in the elementary school. *School Science and Mathematics, 119*(4), 203-212. https://doi.org/10.1111/ssm.12332

Valentine, A., Marinelli, M., & Male, S. (2022). Successfully facilitating initiation of industry engagement in activities which involve students in engineering education, through social capital. *European Journal of Engineering Education, 47*(3), 413-428. https://doi.org/10.1080/03043797.2021.2010033

Van Cleynenbreugel, C., De Winter, V., Buyse, E., & Laevers, F. (2011). Understanding the physical world: Teacher and pupil attitudes towards science and technology. *Professional development for primary teachers in science and technology* (pp. 121-143) Springer.

Van de Pol, J., de Vries, N., Poorthuis, A. M., & Mainhard, T. (2022). The questionnaire on teacher support adaptivity (QTSA): Reliability and validity of student perceptions. *The Journal of Experimental Education,* 1-33. https://doi.org/10.1080/00220973.2022.2100732

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271-296. https://doi.org/10.1007/s10648-010-9127-6

Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2014). Teacher scaffolding in small-group work: An intervention study. *Journal of the Learning Sciences, 23*(4), 600-650. https://doi.org/10.1080/10508406.2013.805300

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal, 47*(2), 497-527. https://doi.org/10.3102%2F0002831209353594

Van der Schaaf, M., Slof, B., Boven, L., & De Jong, A. (2019). Evidence for measuring teachers' core practices. *European Journal of Teacher Education, 42*(5), 675-694. https://doi.org/10.1080/02619768.2019.1652903

Van der Steen, S. (2014). *"How does it work?": A longitudinal microgenetic study on the development of young children's understanding of scientific concepts.* (doctoral dissertation). http://hdl.handle.net/11370/408b8e4e-2be4-4312-a48a-8898995dc273

Van der Ven, S., Boom, J., Kroesbergen, E., & Leseman, P. (2012). Microgenetic patterns of children's multiplication learning: Confirming the overlapping waves model by latent growth modeling. *Journal of Experimental Child Psychology, 113*(1), 1-19. https://doi.org/10.1016/j.jecp.2012.02.001

Van Eijk, R., Evers, J., Haan, F., Klapwijk, T., Kooistra, B., Klink, I., Snoek, M. (2015). *Development prospects and careers of teachers: Advice from the critical friends of the teachers agenda to the minister and state secretary. the critical friends of the teachers' agenda.* Amsterdam University of Applied Sciences.

Van Merriënboer, J. J. (2013). Perspectives on problem solving and instruction. *Computers & Education, 64,* 153-160. https://doi.org/10.1016/j.compedu.2012.11.025

Van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist, 38*(1), 5-13. https://doi.org/10.1207/S15326985EP3801_2

Van Merriënboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147-177. https://doi.org/10.1007/s10648-005-3951-0

Van Niekerk, E., Ankiewicz, P., & de Swardt, E. (2010). A process-based assessment framework for technology education: A case study. *International Journal of Technology and Design Education, 20*(2), 191-215. https://doi.org/10.1007/s10798-008-9070-8

Van Peppen, L. M., Verkoeijen, P. P., Kolenbrander, S. V., Heijltjes, A. E., Janssen, E. M., & van Gog, T. (2021). Learning to avoid biased reasoning: Effects of interleaved practice and worked examples. *Journal of Cognitive Psychology, 33*(3), 304-326. https://doi.org/10.1080/20445911.2021.1890092

Van Tuijl, C., & Walma van der Molen, J. H. (2016). Study choice and career development in STEM fields: An overview and integration of the research. *International Journal of Technology and Design Education, 26*(2), 159-183. https://doi.org/10.1007/s10798-015-9308-1

Van Tuijl, C., Walma van der Molen, J., & Grol, M. (2014). Techniek? niks voor mij! vroege beroepsuitsluiting. *Jeugd School En Wereld, 99*(4), 12-15.

Vanbecelaere, S., Van den Berghe, K., Cornillie, F., Sasanguie, D., Reynvoet, B., & Depaepe, F. (2020). The effectiveness of adaptive versus non-adaptive learning with digital educational games. *Journal of Computer Assisted Learning, 36*(4), 502-513. https://doi.org/10.1111/jcal.12416

Verhavert, S. (2018). Beyond a mere rank order: *The method, the reliability and the efficiency of comparative judgment*. Doctoral dissertation, University of Antwerp.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice,* 1-22. https://doi.org/10.1080/0969594X.2019.1602027

Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2014). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science,* 1-15. http://dx.doi.org/10.1007/2Fs11251-014-9334-5

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*(2), 436. https://psycnet.apa.org/doi/10.1037/0022-3514.49.2.436

Walma van der Molen, J. (2008). De belangstelling voor wetenschap en techniek in het basisonderwijs. In D. Fourage, & A. de Grip (Eds.), *Technotopics III: Essays over onderwijs en arbeidsmarkt voor bètatechnici* (pp. 12-21). Den Haag: Platform Bèta Techniek. http://dare.uva.nl/record/306426

Wammes, D., Slof, B., Schot, W., & Kester, L. (2022a). Pupils' prior knowledge about technological systems: Design and validation of a diagnostic tool for primary school teachers. *International Journal of Technology and Design Education, 32*(5), 2577-2609. https://doi.org/10.1007/s10798-021-09697-z

Wammes, D., Slof, B., Schot, W., & Kester, L. (2022b). Fostering pe-service primary school teachers' ability to recognize differences in pupils' understanding of technical systems. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-022-09774-x

Wammes, D., Slof, B., Schot, W., & Kester, L. (2023). Teacher judgement accuracy of technical abilities in primary education. *International Journal of Technology and Design Education, 33:* 415-438. https://doi.org/10.1007/s10798-022-09734-5.

Wang, L. (2017). Various spatial skills, gender differences, and transferability of spatial skills. Visual-spatial Ability in STEM Education: *Transforming Research into Practice,* 85-105. https://doi.org/ 10.1007/978-3-319-44385-0_5

Wickens, C. D., Hutchins, S., Carolan, T., & Cumming, J. (2013). Effectiveness of part-task training and increasing-difficulty training strategies: A meta-analysis approach. *Human Factors, 55*(2), 461-470. https://doi.org/10.1177/0018720812451994

Wiesner, S. (2014). The development of technicians as a key factor for a sustainable development of renewable energies using an adapted education method based on the successful german dual education (duale ausbildung). *Energy Procedia, 57,* 1034-1036. https://doi.org/10.1016/j.egypro.2014.10.069

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49-65. https://doi.org/10.1080/0969594042000208994

Williams, P. J. (2013). Research in technology education: Looking back to move forward. *International Journal of Technology and Design Education, 23*(1), 1-9. https://doi.org/10.1007/s10798-011-9170-8

Wilson, S. M. (2013). Professional development for science teachers. *Science, 340*(6130), 310-313. https://doi.org/10.1126/science.1230725

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education, 19*(3), 409-427. https://doi.org/10.1007/s10459-013-9453-9

Wu, C. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14*(4), 1261-1295. https://doi.org:10.1214/aos/1176350142

Wulfeck, W. H.Adapting instruction. Paper presented at the *International Conference on Foundations of Augmented Cognition,* 687-695.

Yan, Z., & Cheng, E. C. K. (2015). Primary teachers' attitudes, intentions and practices regarding formative assessment. *Teaching and Teacher Education, 45,* 128-136. https://doi.org/10.1016/j.tate.2014.10.002

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice,* 1-33. https://doi.org/10.1080/0969594X.2021.1884042

Zacharia, Z. C. (2007). Comparing and combining real and virtual experimentation: An effort to enhance students' conceptual understanding of electric circuits. *Journal of Computer Assisted Learning, 23*(2), 120-132. https://doi.org/10.1111/j.1365-2729.2006.00215.x

Zhu, M., & Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education, 30*(1), 21-39. https://doi.org/10.1007/s10212-014-0225-6

Zhu, M., Urhahne, D., & Rubie-Davies, C. M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology, 38*(5), 648-668. https://doi.org/10.1080/01443410.2017.1412399

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20*(2) https://doi.org/10.1016/j.pse.2014.11.003

Zuzovsky, R. (1999). Performance assessment in science: Lessons from the practical assessment of 4th grade students in israel. *Studies in Educational Evaluation, 25*(3), 195-216. https://doi-org.proxy.library.uu.nl/10.1016/S0191-491X(99)00022-X

SUMMARY

Due to the importance of technology in contemporary society, engineering has been introduced into the curricula of primary schools in many countries. Besides familiarising pupils with technology, it aims to provide positive experiences with the subject. It is known that without such experiences, pupils, even in primary education, tend to exclude engineering as a direction for further study or professional careers. Although in primary education, a strong focus on basic language and numeracy skills often results in a marginal focus on engineering, there are schools that consider it a subject which offers pupils the opportunity to develop a broad spectrum of knowledge and skills. This dissertation is inspired by teachers of such schools who try to improve their teaching in this domain but face the problem that it is very difficult to establish pupils' skills and skill development in engineering.

Insight into technical skills and their development offers teachers opportunities to adapt teaching to differences in prior knowledge (e.g., skill level), evaluate the effectiveness of lessons, inform pupils and parents about skills and progress in this domain and communicate with colleagues about successful approaches. Primary school teachers, however, find it difficult to gain insight into those skills. Assuming that a diagnostic tool would contribute to such insight, the main research question of this dissertation was: 'Which support can a diagnostic tool designed for classroom use offer teachers to infer and promote pupils' engineering skills?' Four studies were carried out to answer this question.

Engineering skills often relate to systems, which function depends on interactions between its components. The first study identified the characteristics of skill development related to the construction and understanding of such systems and designed a blueprint for diagnostic tasks suitable to monitor that development. The second study explores the tool's suitability for classroom use with two tasks that comply with the blueprint defined in the first study. The third study evaluates the effect of a training based on the diagnostics developed in the first study on the development of teachers' understanding of technical skills. The fourth and final study used the diagnostic tool to explore the effect of adapting task difficulty on skill development.

The *first study* presented in **Chapter 2** used the Evidence Centred Design framework to create a blueprint, describing ten guidelines for compiling diagnostic tasks that are feasible and valid to assess pupils' prior knowledge about material-based systems often used in engineering activities at primary

schools (e.g., constructions, mechanical systems and electric circuits). The guidelines are: (1) the diagnosis should be based on *performance*, (2) diagnostic tasks should *represent a particular material-based system,* (3) skill development should be identified on a *one-dimensional scale,* (4) a single task should be sufficient to assess pupils' skill related to a specific system, (5) pupils should be able to make a task *independently within ten minutes,* (6) pupils' work-products (i.e., task results) should suffice to infer their skill level, (7) the *opportunities to demonstrate partial knowledge* should be maximised, (8) opportunities to *evaluate actions* should be *restricted,* (9) the *Fischer scale* should be used as a generic ruler, and (10) each task should have *specific scoring rules* derived from the Fischer scale. Two tasks were compiled based on these guidelines: the Buzz-Wire (BW) task, based on an electric circuit, and the Stairs Marble Track (SMT) task, which is a mechanical device with a camshaft.

The *second* study described in **Chapter 3** explores the tools' suitability for classroom use and its value for the teachers' insight into their pupils' technical skills. Eight teachers were asked to predict and substantiate their pupils' technical skills, after which they used the Buzz-Wire and Stairs Marble Track task in their classroom to get information about pupils' performance. With that knowledge, the teachers were interviewed.

Task performance did deviate from the teachers' estimates. Teachers' estimates of pupils' technical abilities seem biased by their knowledge about pupils' reading comprehension and math results. Biases were also found for pupils' general learning behaviour and gender. Moreover, the less experienced teachers systematically underestimated their pupils' technical abilities. This resulted in a low judgement accuracy on pupils' technical abilities of most participating teachers, their relative judgements (i.e., ranking pupils by presumed skill level) being, on average, slightly better than their absolute judgements (i.e., prediction skill levels).

The teachers with pupils aged nine or older were positive about the feasibility of the tool for classroom use. In the interviews, they expressed their surprise at pupils who performed much better than they expected. For some pupils, teachers indicated that the results confirmed their idea that they had more capabilities as demonstrated on the standardised tests.

Chapter 4 describes the *third* study, which evaluated the effects of a training based on Nickerson's anchoring and adjustment model. This model states that people are inclined to think that others think as they do and need to be made aware of the fact that others might think differently. This is especially relevant for teachers. The diagnostic tool and its underlying theory were used

to make teachers aware of the similarities and differences between their own understanding of technical systems and that of pupils at different stages of their comprehension of such systems.

The pre-posttest design of the study revealed that the teachers were, at the pre-test, already quite accurate in the comparative classification of pupils' results on the BW and SMT task. This accuracy improved slightly on post-test. The training had a positive effect on the teachers' ability to infer pupils' thinking from task results, and most teachers were able to identify skill levels from results on the diagnostic tasks. Moreover, the teachers' self-efficacy beliefs with regard to teaching technical subjects and its effect on pupils' skill development were positively affected by the training. This is meaningful as teachers' self-efficacy beliefs strongly predict their teaching of this domain in practice. (Rohaan et al., 2012; Utley et al., 2019).

Chapter 5 describes the results of the two experiments of the *fourth* study based on the Challenge Point Framework (CPF, Guadagnoli & Lee, 2004). The CPF states that there is an optimal challenge to promote skill development which relates to the learners' prior knowledge (i.e., skills). Challenges below that optimal challenge point offer insufficient information to induce learning. Challenges that exceed the optimal challenge point are likely to restrict learning as they offer more information than the learners' processing capacity can handle, a stance which is consistent with the Cognitive Load Theory (Sweller, 1988).

Based on the CPF, we hypothesised that skill development would be mediated by challenge adequacy. However, the results of both experiments did not support this hypothesis. Although the first experiment showed a significant positive effect of the adaptive condition on perceived challenge adequacy compared to a yoked condition, this did not result in differences in skill development between conditions. In the second experiment, which contrasted the adaptive condition with a condition in which only the most difficult tasks were offered, there was no significant effect of condition on perceived challenge adequacy and skill development. The absence of the relation with perceived challenge adequacy, as found in the first experiment, may relate to the absence of low levels of challenge in the contrasting condition of the second experiment.

Post hoc analyses revealed an unexpected positive effect of a high challenge for pupils with the lowest levels of prior knowledge. In contrast, a high challenge was not beneficial for pupils with the second highest pre-test level, for which it even had a significant negative effect. We assume that the positive effect of high challenge on low pre-test performers might be caused by the worked-out-example-like aspects of the most difficult tasks

(Atkinson et al., 2000; Chen et al., 2019; Van Peppen et al., 2021). The negative effect of high challenge for pupils with the second-highest pre-test level might relate to a scalloping effect (Granott et al., 2002; Van der Ven et al., 2012).

Chapter 6 discusses the answer to our main research question "Which support can a diagnostic tool designed for classroom use offer teachers to infer and promote pupils' engineering skills?" We answered this question through the model of Zi Yan et al. (2021), which describes the factors that influence the use of assessment in classrooms. The model states that intentions to use assessments are especially affected by instrumental attitude, training and self-efficacy beliefs. A positive instrumental attitude was expressed in the interviews. Teachers considered the tool valuable as it helped them to acknowledge and communicate skills that were not identified with existing tests. Training improved the teachers' diagnostic ability and had a strong positive effect on their self-efficacy beliefs about teaching engineering.

The application of the Fischer scale in the context of formative assessment has created new opportunities to study the relationship between prior knowledge, task difficulty and adaptive teaching. We could demonstrate that hands-on tasks were especially beneficial for pupils with low levels of prior knowledge and that a challenge, as perceived by pupils, did not always coincide with an optimal challenge for learning. Our tool offers opportunities for further exploration of, for instance, the effect of hands-on tasks compared to a teacher demonstration or computer-based simulations or the effect of adapted feedback.

NEDERLANDSE SAMENVATTING (SUMMARY IN DUTCH)

Techniek is alom aanwezig in de hedendaagse samenleving en wordt daarom in veel landen al op de basisschool onderwezen. Naast het vertrouwd maken van leerlingen met techniek, heeft dit tot doel leerlingen positieve ervaringen met het vak op te laten doen. Het is bekend dat zonder dergelijke ervaringen leerlingen al aan het eind van de basisschool geneigd zijn om techniek uit te sluiten als een richting voor verdere studie of beroepsloopbaan. Hoewel in het basisonderwijs een sterke focus op basisvaardigheden taal en rekenen vaak resulteert in beperkte aandacht voor techniek, zijn er scholen die het vak gebruiken om leerlingen de kans te geven een breed spectrum aan kennis en vaardigheden te ontwikkelen. Dit proefschrift is geïnspireerd door leerkrachten van dergelijke scholen die proberen hun onderwijs op dit gebied te verbeteren, maar geconfronteerd worden met het probleem dat het erg moeilijk is om de vaardigheden en ontwikkeling van leerlingen bij techniek vast te stellen.

Zicht op technische vaardigheden en hun ontwikkeling biedt leraren mogelijkheden om het onderwijs aan te passen aan verschillen tussen leerlingen, de effectiviteit van lessen te evalueren, leerlingen en ouders te informeren over vaardigheden en vorderingen op dit gebied en te communiceren met collega's over succesvolle benaderingen. Leerkrachten in het basisonderwijs hebben echter moeite om zicht te krijgen op die technische vaardigheden. Vanuit de veronderstelling dat een diagnostisch hulpmiddel hen zal helpen bij het verkrijgen van zicht op de technische vaardigheid van hun leerlingen, is de centrale vraag in dit proefschrift: 'Welke ondersteuning kan een diagnostisch hulpmiddel dat ontworpen is voor gebruik in de klas leraren bieden om de technische vaardigheden van leerlingen af te leiden en te bevorderen?' Om deze vraag te beantwoorden zijn vier onderzoeken uitgevoerd.

Technische vaardigheden hebben veelal betrekking op systemen die functioneren door interactie tussen de componenten. De eerste studie verkent de ontwikkeling van vaardigheden die leerlingen in staat stellen om technische systemen te begrijpen en te construeren. Op basis daarvan is een blauwdruk ontworpen voor diagnostische taken die geschikt zijn om die ontwikkeling te volgen. De tweede studie onderzoekt de geschiktheid van de in de eerste studie ontwikkelde diagnostiek voor gebruik in de klas. De derde studie evalueert het effect op de ontwikkeling van inzicht in technische vaardigheden van leerkrachten van een training die gebaseerd was op de in de eerste studie ontwikkelde diagnostiek. In de vierde en laatste studie is de diagnostiek gebruikt om opdrachten af te stemmen op de beginsituatie van de

leerlingen. Het effect daarvan op de vaardigheidheidsontwikkeling is vergeleken met het aanbieden van niet op voorkennis afgestemde taken.

De eerste studie beschreven in hoofdstuk 2, ontwikkelde aan de hand van het Evidence Centered Design-raamwerk een blauwdruk voor diagnostische taken om de vaardigheid van leerlingen om technische systemen te construeren op een valide en in de basisschool toepasbare wijze vast te kunnen stellen. De blauwdruk bestaat uit tien richtlijnen voor de constructie van dergelijke diagnostische taken. Deze richtlijnen zijn: (1) de vaardigheid moet handelend worden aangetoond, (2) elke taak moet gebaseerd zijn op een specifiek technisch systeem, (3) de vaardigheid moet op een ééndimensionale schaal worden geïdentificeerd, (4) een enkele taak moet voldoende zijn om de vaardigheden van leerlingen met betrekking tot een specifiek systeem te beoordelen, (5) leerlingen moeten in staat zijn om zelfstandig een taak binnen tien minuten uitvoeren, (6) het vaardigheidsniveau moet vastgesteld kunnen worden aan de hand van het werkproduct (d.w.z. het resultaat), (7) de taak moet de leering voldoende kans bieden om deelvaardigheden aan te tonen, (8) de mogelijkheid om uit te proberen moet beperkt zijn, (9) de Fischer-schaal moet worden gebruikt als referentiekader voor elke taak, en (10) elke taak moet specifieke scoreregels hebben die van de Fischer-schaal zijn afgeleid. Op basis van deze richtlijnen zijn twee taken gecontrueerd: de zenuwspiraaltaak (Buzz-Wire; BW), gebaseerd op een electrisch circuit en de trapkogelbaantaak (Stairs Marble Track, SMT), gebaseerd op een mechanisch systeem met een nokkenas.

De tweede studie beschreven in hoofdstuk 3 verkent of op de bovengenoemde richtlijnen gebaseerde taken toepasbaar zijn in de klas en of de uitkomsten bijdragen aan het zicht van leerkrachten op de technische vaardigheid van hun leerlingen. Acht leraren is gevraagd om de technische vaardigheden van hun leerlingen te voorspellen en de voorspellingen te onderbouwen. Daarna gebruikten ze de zenuwspiraal en trapkogelbaan taak in hun klas. Tenslotte werd elk leerkracht geïnterviewd.

De resultaten op de taken weken af van de inschattingen. Leerkrachten lijken bij die inschattingen de voorspellende waarde van prestaties op het gebied van begrijpend lezen en rekenen voor technische vaardigheden te overschatten. Daarnaast lijken de leerkrachten de technische vaardigheid van meisjes en van leerlingen met taak-werkhouding problemen lager in te schatten dan de prestaties op de taken lieten zien. Een systematische onderschatting van de technische vaardigheid van leerlingen was vooral merkbaar bij leerkrachten met minder dan vijf jaar ervaring. De op rangorde gebaseerde inschattingen waren beter dan inschattingen van vaardigheidsniveaus.

De leerkrachten met leerlingen van negen jaar of ouder waren positief over de praktische toepasbaarheid van de taken in hun klas. In de interviews spraken ze hun verbazing uit over leerlingen die veel beter presteerden dan ze hadden verwacht. Voor sommige leerlingen bevestigden de resultaten de vermoedens van hun leerkracht dat zij over meer capaciteiten beschikten dan zichtbaar bij de reguliere toetsen.

Hoofdstuk 4 beschrijft de derde studie, die de effecten evalueerde van een training gebaseerd op het verankerings- en aanpassingsmodel van Nickerson. Dit model stelt dat mensen geneigd zijn te denken dat anderen denken zoals zij denken en bewust gemaakt moeten worden van het feit dat anderen anders kunnen denken. Dit is vooral relevant voor docenten. Het diagnostisch hulpmiddel en de onderliggende theorie werden gebruikt om leraren bewust te maken van de overeenkomsten en verschillen tussen hun eigen begrip van technische systemen en dat van leerlingen in verschillende stadia van de ontwikkeling van hun inzicht in de werking van dergelijke systemen.

Leerkrachten bleken voorafgaand aan de training al vrij nauwkeurig in hun inschatting van verschil in vaardigheidsontwikkeling als ze twee resultaten konden vergelijken. Deze inschatting was nog wat beter na de training. De training had een positief effect op het vermogen van de leerkrachten om zich een voorstelling te maken van het soort kennis dat leerlingen hadden gebruikt om de taak uit te voeren. De meeste leraren konden na de training het vaardigheidsniveau vrij nauwkeurig afleiden uit de taakresultaten. De training had een sterk effect op de opvattingen van de deelnemers over hun vermogen om techniekonderwijs te geven en op de bijdrage van techniekonderwijs aan de ontwikkeling van technische vaardigheden van leerlingen. Dit is betekenisvol omdat overtuigingen over zelfeffectiviteit een sterke voorspeller zijn van de aandacht die leerkrachten besteden aan techniekonderwijs. (Rohaan et al., 2012; Utley et al., 2019).

Hoofdstuk 5 beschrijft de resultaten van de twee experimenten gebaseerd op het Challenge Point Framework (CPF, Guadagnoli & Lee, 2004). Het CPF stelt dat er een optimale uitdaging is om de ontwikkeling van vaardigheden te bevorderen. Wat dat optimum is hangt af van de mate waarin leerlingen de vereiste vaardigheid al beheersen. Een sub-optimale uitdaging biedt de leerlingen te weinig informatie om hun vaardigheid verder te ontwikkelen. Te grote uitdagingen kunnen de vaardigheidsontwikkeling beperken omdat de hoeveelheid informatie groter is dan de leerling kan verwerken, hetgeen ook bekend is vanuit de Cognitive Load Theory (Sweller, 1988).

Op basis van het CPF, veronderstelden we dat de ontwikkeling van vaardigheden zou worden gemedieerd door een als optimaal ervaren uitdaging.

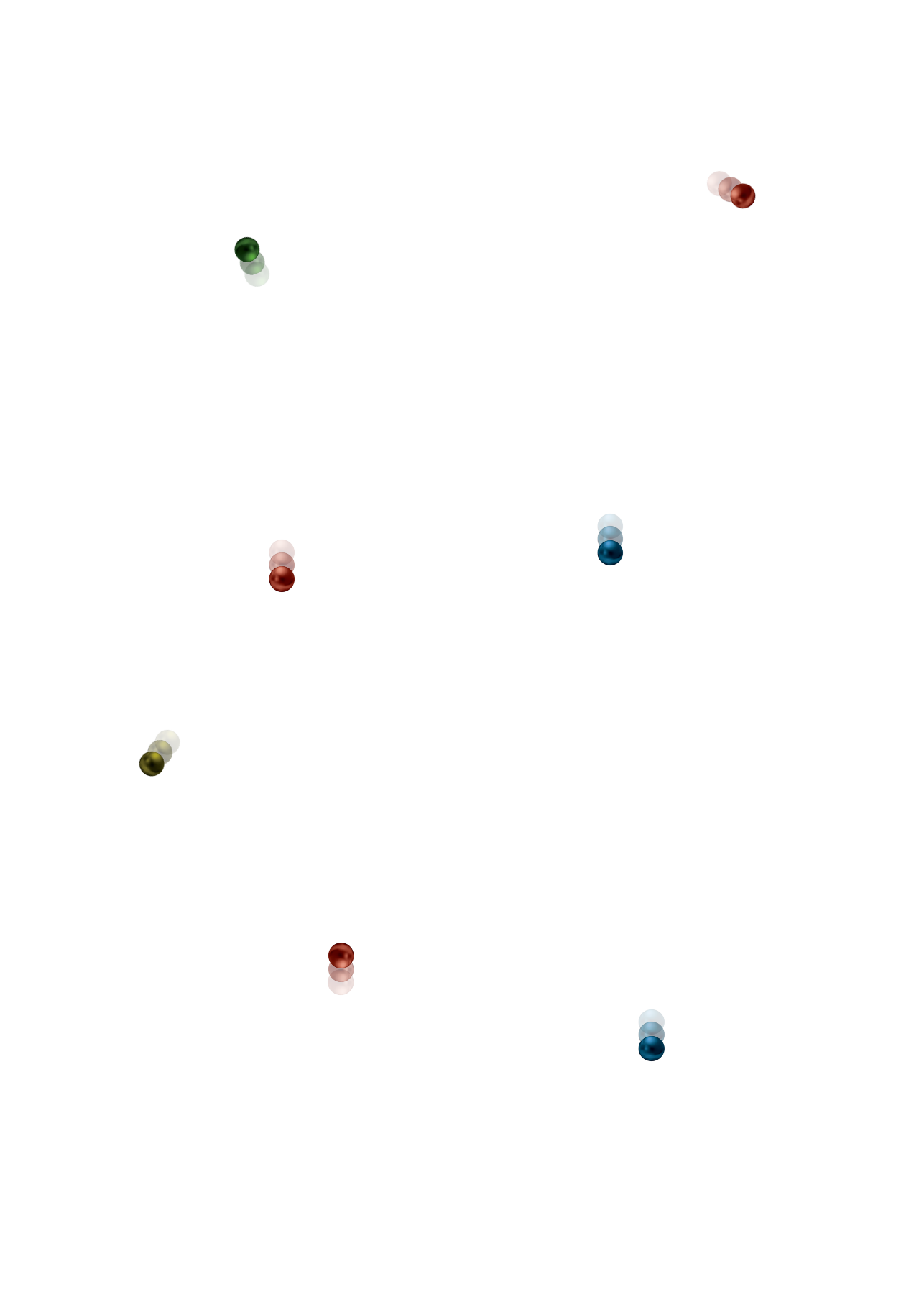
De resultaten van beide experimenten kwamen niet overeen met deze hypothese. Bij het eerste experiment was er weliswaar een significant positief effect op de tevredenheid van de leerlingen over als de taken waren afgestemd op hun voorkennis, maar dit veroorzaakte geen verschil in vaardigheidsontwikkeling in vergelijking met leerlingen waarbij de taken niet op hun voorkennis waren afgestemd. In het tweede experiment waarbij leerlingen in de niet op voorkennis afgestemde conditie uitsluitend de moeilijkste taken kregen was er geen significant verschil in tevredenheid over de geboden uitdaging en in vaardigheidsontwikkeling. De afwezigheid van het eerder wél aanwezige positieve effect op de tevredenheid van op voorkennis afgestemde taken hangt mogelijk samen met het ontbreken van weinig uitdaging in het tweede experiment.

Post hoc analyses lieten een onverwacht positief effect zien van een hoge uitdaging voor de minst vaardige leerlingen. Daarentegen had een hoge uitdaging een significant negatief effect bij leerlingen met het op één na hoogste vaardigheidsniveau. We veronderstellen dat het positieve effect van een hoge uitdaging bij de minst vaardige leerlingen samenhangt met de structuur van de moeilijkste taken. Bij deze taken construeren en veranderen de leerlingen elektrische circuits aan de hand van foto's, waarna leerlingen wordt gevraagd om met informatie over weerstand de waargenomen veranderingen te verklaren. De gefotografeerde voorbeelden bieden ook de minst vaardige leerlingen de mogelijkheid om werkende circuits te construeren hetgeen zij als een positieve en haalbare uitdaging ervaren (Atkinson et al., 2000; Chen et al., 2019; Van Peppen et al., 2021). Het negatieve effect van een hoge uitdaging voor leerlingen die voorafgaand aan de les het op één na hoogste vaardigheidsniveau lieten zien kan mogelijk een gevolg zijn van een scalloping-effect (Granott et al., 2002; Van der Ven et al., 2012).

Hoofdstuk 6 bespreekt het antwoord op onze hoofdonderzoeksvraag "Welke ondersteuning kan een diagnostisch hulpmiddel dat ontworpen is voor gebruik in de klas leraren bieden om de technische vaardigheden van leerlingen af te leiden en te bevorderen?" Bij het beantwoorden van deze vraag hebben we gebruik gemaakt van het model van Zi Yan et al. (2021), waarin de factoren worden beschreven die van invloed zijn op het gebruik van toetsen in de klas. Dit model stelt dat intenties om assessments te gebruiken vooral worden beïnvloed door instrumentele houding (i.e. de waarde die men aan diagnostiek hecht) , training en overtuigingen over de eigen bekwaamheid. Een positieve instrumentele houding kwam naar voren in de interviews. Leerkrachten vonden dat de diagnostische taken waardevol omdat ze vaardigheden identificeren die niet blijken uit de reguliere toetsing. Training verbeterde het diagnostisch vermogen van de leraren en had een sterk positief effect op hun

zelfeffectiviteitsovertuigingen ten aanzien van het onderwijzen van techniek.

Het toepassen van de Fischer-schaal in de context van formatieve beoordeling heeft nieuwe mogelijkheden gecreëerd om de relatie tussen voorkennis, taakmoeilijkheid en adaptief onderwijs te bestuderen. We konden hiermee al zichtbaar maken dat vooral leerlingen met lage vaardigheidsniveaus lijken te profiteren van praktische taken en dat een optimale uitdaging, zoals die door leerlingen wordt ervaren, geen sterke voorspellende waarde lijkt te hebben voor hun vaardigheidsontwikkeling. De door ons ontwikkelde diagnostiek biedt mogelijkheden voor verder onderzoek naar het effect op vaardigheidsontwikkeling van praktische taken in vergelijking met andere onderwijsvormen zoals demonstratie, computer simulaties of het effect van adaptieve feedback.



SUPPLEMENTARY MATERIALS

CHAPTER 2: PUPILS' PRIOR KNOWLEDGE ABOUT TECHNOLOGICAL SYSTEMS: DESIGN AND VALIDATION OF A DIAGNOSTIC TOOL FOR PRIMARY SCHOOL TEACHERS

dataverse: doi:10.34894/PUQDCX

1. General data
   1.1. Ethical approval and information for parents
2. Data collection
   2.1. Coding Files: description of the coding of work products
   2.2. D-PAC: data of comparisons of BW and SMT work products by technology-education experts from the D-PAC platform
   2.3. Fischer Experts: reactions of Fischer experts on initial BW and SMT coding, interview transcripts (Dutch)
   2.4. Interrater: coding of pupils' actions by two independent raters.
   2.5. SchoolData: coding of actions and final result of pupils
   2.6. Task Explanation: explanation of tasks to pupils
3. Data analyses: explanation of data collection and analyses
   3.1. SPSS: SPSS scripts, data- and output files
4. Article: final article and figures

Initial scoring rules

Level	Buzz-Wire	Stairs Marble Track
Single abstractions (Rp4/Sa1) An abstraction is used to realise the solution. Here only the BW-task needs abstract thinking to realise a correct solution: the proper connection of the loop and spiral as an off/on switch within an electric circuit.		
(7) Abs1/Rp4 Single abstraction	The correct solution, use of all representations *If not: go to Rp3*	Not included.
Representational system (Rp3) The relations between all elements of the system are incorporated into the solution.		
(6) Rp3 Representational system	One element of the correct solution is missing, representation 1 or 2 (see Rp1) *If not: go to Rp2*	Correct solution.
Representational mappings (Rp2) Causal relationships with an intermediate step, linking single causal relations.		
(5) Rp2 Representational mapping	The solution contains the representations 1 and 2 (see Rp1) *or* there is a circuit in which the lamp and buzzer will function *or* the spiral and loop will function as a switch *If not: go to Rp1*	Combination of two representations. 1. Slanted tops upward and in the correct direction (but not all bars at the correct position). 2. Bars in the correct position and slanted tops enabling a marble to move to the next bar (but not all slanted tops in the correct direction).
Single representations (Sm4/Rp1) A single representation (mental coordination of two or more sensory-motor systems) is part of the action. Single causal relationships.		
(4) Rp1/Sm4 Single representation	All connections are on metal, *or* the battery is used as a source of pone pole to the other by another part. *If not: go to Sm3*	Use of one representation. 1. The orientation of the slanted top prevents marble from rolling sideways *or* 2. Bars in the correct order
Sensorimotor – system (Sm3) Observable causal relationships. A manipulation is linked to an observable consequence.		
(3) Sm3 Sensorimotor system	At least four objects are connected, considering the spiral and loop as different objects. *If not: go to Sm2*	The frame contains at least 5 bars. Within that condition, there are several possibilities. 1. The slides are upside down, positioned around the profile of the eccentric wheel. 2. Slides are placed upward, but not in the correct order and not preventing a marble to roll-off sideways. 3. The bars are skewed in the frame.
Sensorimotor mapping (Sm2) Combining features of two objects.		
(2) Sm2 Sensorimotor mappings	At least two objects are connected. *If not: go to Sm1*	A combination of two parts (two bars or bar and frame), based on their length or shape.
Sensorimotor actions (Sm1) A single feature of object or task. Observable.		
(1) Sm1 Sensorimotor actions	Remaining solutions.	Remaining solutions. For instance, observation or inserting a single bar into the frame.

Scoring rules BW and SMT work products

Level	Buzz-Wire	Stairs Marble Track
Single abstractions (Rp4/Sa1) An abstraction is used to realise the solution.		
7	Correct solution. Loop and spiral connected as a switch *and* correct connection of the battery *and* all connections on metal *and* all components will function when the circuit closes (no short circuit). If not: go to Rp3	Not used. The SMT task does not require abstract knowledge.
Representational system (Rp3) Relationships have been established between all components of the system.		
6	The work product demands the combination of multiple representations • An electric circuit in which both lamp and buzzer function by default *and* all connections on metal OR • Loop and spiral connected as a switch that sets a lamp or buzzer on/off, *disregarding* whether all connections are on metal If not: go to Rp2	The correct configuration. It demands the combination of all representations. • All bars are ordered according to their length. • The orientation of the slanted bar tops potentially allows a marble to roll onto the slanted top of an adjacent bar. • The slope of all slanted tops will cause the marble to roll down in the direction of the high roll-off point. • A correct estimate of the effect of a turning camshaft on the movement and height of adjacent bars.
Representational mappings (Rp2) Causal relationships with an intermediate step, linking single causal relations.		
5	The outcome contains a mapping: • A connected lamp or buzzer will function in an electric circuit, disregarding whether all connections are on metal. OR • All connections on metal, including both connection points of the lamp *and* the spiral and loop, are linked through a connection by one or more other components. If not: go to Rp1	The outcome contains a mapping: • All bars are in the correct order, *and* the direction of all slides at the bar top allows a marble to role upon the slide of an adjacent bar (any correct combination of correct or 180°-rotated slide positions) OR • The slope of at least five tops will cause the marble to roll down in the direction of the high roll-off point (but bars not in the correct order or one bar incorrect).

Level	Buzz-Wire	Stairs Marble Track
Single representations (Sm4/Rp1) A single representation (mental coordination of two or more sensory-motor systems) is part of the action. Single causal relationships.		
4	Use of a single representation. • There is a connection from one pole of the battery to the other pole by at least one other component (lamp, buzzer, loop, spiral). • Both poles of the lamp or buzzer connected to the battery. • All connections should be on metal, conducting electricity. Both poles of the lamp should be connected in this way • The ring and spiral are linked. Not directly but via at least one other component. If not: go to Sm3	Use of a single representation • All bars in the frame are ordered by their length at their correct position in the frame. OR • The direction of all slides in the frame potentially allows a marble to roll upon the slide of an adjacent bar. (any combination of correct or 180°-rotated slide positions)
Sensorimotor – system (Sm3) Observable causal relationships. A manipulation is linked to an observable consequence.		
3	All components are connected, treating the loop and spiral as a single component. If not: go to Sm2	Bars positioned to fill the gap in the frame between the roll-on and roll-off point, • There are bars vertically positioned in the frame with the slanted tops upward (but bars missing or at least one slanted top rotated by 90° or 270°). OR • There are at least five bars in the frame (filling up the space between the low roll-on and high roll-off point) but not with the slanted top upward (vertically top-down or horizontally)
Sensorimotor mapping (Sm2) Combining features of two objects.		
2	Any connection between two components with a wire. If not: go to Sm1	Combinations of single properties of bars and frame (single or repeated) • At least one combination [mapping] of a single property (e.g. length or top-shape) of two or more bars OR • At least one combination of a single property of a bar and a property of the frame (wheel-support, length of the gap between roll-on and roll-off point)
Single sensorimotor actions (Sm1) Use of a single feature of object or task. Observable.		
1	The work product has a connection but does not include a connection between two components by a wire. If not: no action = 0	Something has been changed, but the work product does not include a combination of bar or bar and frame features.

Note. Text in grey represents the general rule (adapted from Van der Steen, 2014).

CHAPTER 3: TEACHER JUDGEMENT ACCURACY OF TECHNICAL ABILITIES IN PRIMARY EDUCATION

dataverse: doi:10.34894/YWIVC4

- General data:
  - readme_dataset.pdf: explanation of structure of dataset
  - Dataset Teacher judgement accuracy: explanation of variables of all datafiles
  - Scoring rules work products: BW and SMT scoring rules

1. Article: article published
2. Interviews: Interview transcripts and notes; teachers' arguments for pupils' rank
3. Knowledge test
4. SPSS: syntax, datafiles and output
5. Work poduct pictures

CHAPTER 4: FOSTERING PRE-SERVICE PRIMARY SCHOOL TEACHERS' ABILITY TO RECOGNISE DIFFERENCES IN PUPILS' UNDERSTANDING OF TECHNICAL SYSTEMS

Dataverse: doi:10.34894/NEII5T

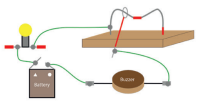
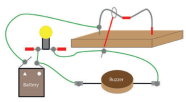
1. Article: article published
2. Instruments: knowledge test, STEBI-adapted and Diagnostic test
3. Stats: R and SPSS syntax, datafiles and spss output files
4. Training: PowerPoints of training sessions , BW and SMT scoring rules (Dutch)

Scoring rules Dutch; BW task

Zenuwspiraal niveau. Begin bovenaan, voldoet het niet aan de regel(s), ga dan een rij naar beneden. Niveau= voldoen aan één regel.

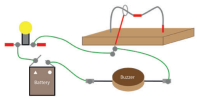
Kennis over schakeling, weerstand en stroomkring gecombineerd. Niveau 7

Correcte oplossing. Stroomkring aan de ene kant aangesloten op de spiraal en aan de andere kant op de ring. Aanraking ring op metaal activeert zowel lamp als zoemer.

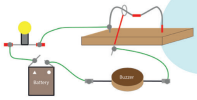



Meerdere representaties gebruikt. Niveau 6

R1R2R3) Er is een circuit waarin de lamp en zoemer permanent aan zijn. Alle verbindingen op metaal

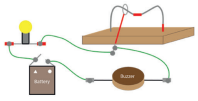


R1 R2 R4 De spiraal en de ring zijn schakelaar in een circuit voor de lamp en/of de zoemer. Verbindingen op metaal of kunststof

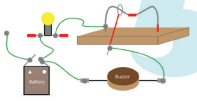


Twee representaties gecombineerd. Niveau 5

R1 R2 Er is een circuit met daarin de lamp of zoemer. Verbindingen op metaal of kunststof



R3 R4 De ring is (alleen) via een andere component verbonden met de spiraal. EN Alle verbindingen op metaal, inclusief beide lamp-aansluitingen.



Eén van de onderstaande vier representaties gebruikt. Niveau 4

R1) De ene pool van de batterij is via een andere component verbonden met de andere pool

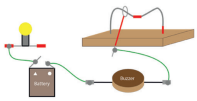
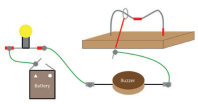
R3) Alle verbindingen op metaal, waaronder beide aansluitpunten van de lamp

R2) Beide polen van de zoemer of lamp verbonden met de batterij (rechtstreeks of via andere component)

R4) De ring is via een andere component verbonden met de spiraal.

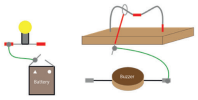
Alle verbindingscombinaties. Niveau 3

Alle componenten zijn met elkaar verbonden. Zien hierbij de ring en spiraal als één component

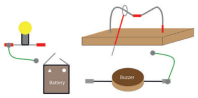
Twee verbindingscombinaties. Niveau 2

Minimaal twee componenten zijn met een snoer aan elkaar verbonden



Eén verbindingscombinatie. Niveau 1

Er is minimaal één snoer aangesloten



Geen acties. Niveau 0

Scoring rules Dutch: SMT task

Trapkogelbaan niveau. Begin in de linker kolom, voldoet het niet aan de regel(s), ga dan een kolom naar rechts. Niveau= voldoen aan één regel.

	6	5	4	3	2	1	0
REGEL	*Alle representaties gebruikt Representational system niveau 6*	*Twee representaties gecombineerd. Representational mapping, niveau 5*	*Eén representatie gebruikt. Single representation, niveau 4*	*Alle zichtbare eigenschappen gecombineerd. Sensorimotor system, niveau 3*	*Twee zichtbare eigenschappen gecombineerd. Sensorimotor mapping, niveau 2*	*Eén zichtbare eigenschap gebruikt. Single sensorimotor action, niveau 1*	*Geen enkele actie ondernomen. niveau 0*
VOORBEELD	1) Alle staafjes zijn geordend op lengte, met de kleinste bij het laagste punt. EN 2) Alle uithollingen aan de bovenkant in de hoog-laag richting. EN 3) Alle uitsparingen lopen omlaag in de richting van het hoogste punt. (correcte oplossing)	1) De staafjes (minimaal 5) op **lengte-volgorde**, met de kleinste bij het laagste punt EN 2) Alle uithollingen aan de bovenkant in de hoog-laag richting OF 2) Alle uithollingen aan de bovenkant in de hoog-laag richting (minimaal 5 staafjes). EN 3) Alle uitsparingen lopen omlaag in de richting van het hoogste punt.(lijkt op correcte oplossing, maar staafjes niet in goede volgorde)	1) Alle staafjes in het frame op lengte-volgorde, met de kleinste bij het laagste punt (maar ondersteboven) OF 2) Alle uithollingen aan de bovenkant in de hoog-laag richting. (staafjes niet in de goede volgorde, minimaal 2 staafjes geplaatst)	Het frame is opgevuld met vijf of zes staafjes, horizontaal, verticaal op de kop (niet op volgorde) of verticaal rechtop. Bij verticaal rechtop zijn er uithollingen die aflopen naar de zijkant.	Vier of minder staafjes, horizontaal of verticaal op de kop in het frame OF Staafjes niet in het frame, maar geordend op tafel	Er is tenminste één blokje verplaatst op de tafel.	

CHAPTER 5: ADAPTING TASK DIFFICULTY TO PUPILS' PRIOR KNOWLEDGE ABOUT ELECTRIC CIRCUITS: EFFECTS ON PERCEIVED CHALLENGE ADEQUACY AND SKILL DEVELOPMENT

Dataverse: doi:10.34894/NEII5T

1. Article: article published
2. Instruments: knowledge test, STEBI-adapted and Diagnostic test
3. Stats: R and SPSS syntax, datafiles and spss output files
4. Training: PowerPoints of training sessions , BW and SMT scoring rules (Dutch)

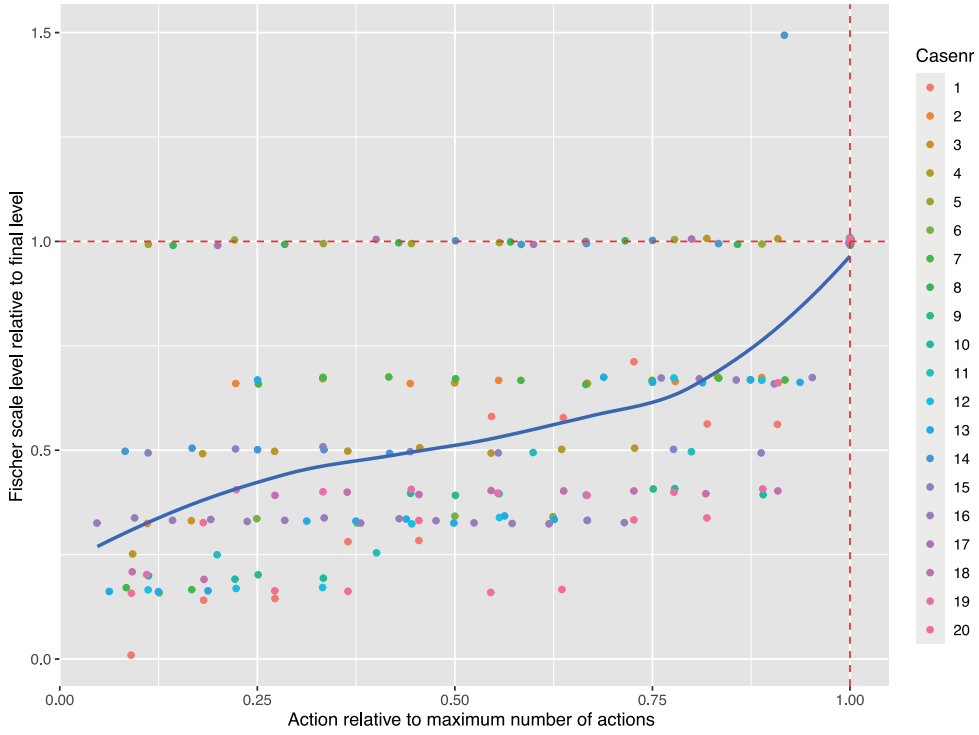
CHAPTER 6:
DIAGNOSING ENGINEERING SKILLS IN PRIMARY CLASSROOMS

Fischer scale level at subsequent actions

The level of the workproduct of the Buzz-Wire (BW) and Stairs Marble Track (SMT) task usually increases towards the level at the point where pupils indicate that they have finished the task. Figure 1 shows the relationship between subsequent actions and the Fischer scale level of 20 pupils on the BW task. As the Fischer scale level differs per pupil, the Y-axis represents the relative Fischer level. The relative Fischer level is calculated by dividing the Fischer level at the n^{th} action by the Fischer level after the last action. This results in 1 as indication of the final level, represented by a dashed red horizontal line. As the number of actions also varies per pupil, the X scale value is the nth action of each pupil divided by the total number of actions of that pupil. The final level is therefore 1 and represented by a dashed red vertical line.

Figure 1

BW task, Fischer level relative to final level at subsequent actions (n=20)

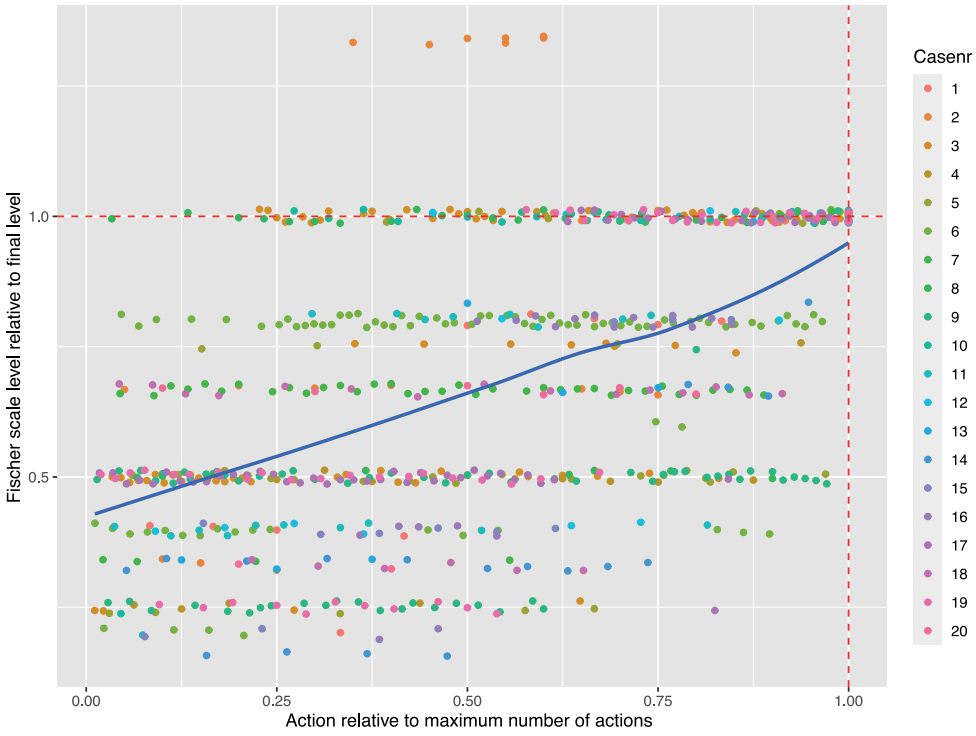


For the BW task, there was one pupil with an intermediate higher level. This pupil disconnected a wire before ending the task, which made the result drop from 3 at the second last action to 2 at the last action.

The results of 20 pupils on the SMT task are represented in Figure 2. The relative Fischer level on the Y-axis and the relative number of actions represented on the X-axis were calculated as in the BW task.

Figure 2

SMT task, Fischer level relative to final level at subsequent actions (n=20)



Here one pupil (case 2) changed the position of the first bar, which has a somewhat different shape at the top to allow marbles to roll on that bar, at the 13th of 20 actions. This caused a drop from level 4 to level 2 as a result of that action, level 3 being the final level as the upper side of the first bar did not align with the other five bars.

ABOUT THE AUTHOR

Dannie Wammes (1956) was born in 1956 in Utrecht, The Netherlands. He received a Master degree in Biology in 1980 at Utrecht University. After being as a researcher at the State Forestry from 1980 to 1982, he worked at the Environmental Education centre in Wageningen. There he organised many different educational projects for primary schools and initiated projects to improve the quality of educational materials at EE centres in the Netherlands. From 2006 onwards, he also became a teacher at the HAN teacher training academy for primary education. Here he trained in-service and pre-service teachers on science and technology. With German training institutes, he realised the website 'Learning by curiosity'. This website contains video recordings and comments on the lessons of German and Dutch teachers in primary and pre-primary education in which they use their pupils' curiosity to enhance learning. During his work with in-service teachers, the question arose how to they could evaluate their teaching approaches in science and technology. This resulted in the application of an NWO grant, which was awarded and made it possible to start a PhD project for two days a week at Utrecht University. The project focussed on engineering skills and was done at schools related to the HAN teacher training academy and with students. Currently, Dannie still works at the Wageningen EE Centre for one day a week and for four days as a teacher at the HAN. At the HAN, his focus is on bachelor research. He also contributes to studies of the Dutch Inspectorate of Education about the contribution of primary schools to pupils' knowledge and skills related to science and technology.

LIST OF PUBLICATIONS AND PROFESSIONAL CONTRIBUTIONS

Peer-Reviewed Publications

Roelofs, E., Wammes, D., Emons, W., & Raijmakers, M. (2022). Dynamisch toetsen van onderzoeksvaardigheden op het terrein van Natuur en Techniek bij leerlingen van groep 8 van het basisonderwijs. Pedagogische Studien, 98(5), 369-387.

Wammes, D., Slof, B., Schot, W., & Kester, L. (2022a). Pupils' prior knowledge about technological systems: Design and validation of a diagnostic tool for primary school teachers. *International Journal of Technology and Design Education, 32*(5), 2577-2609.

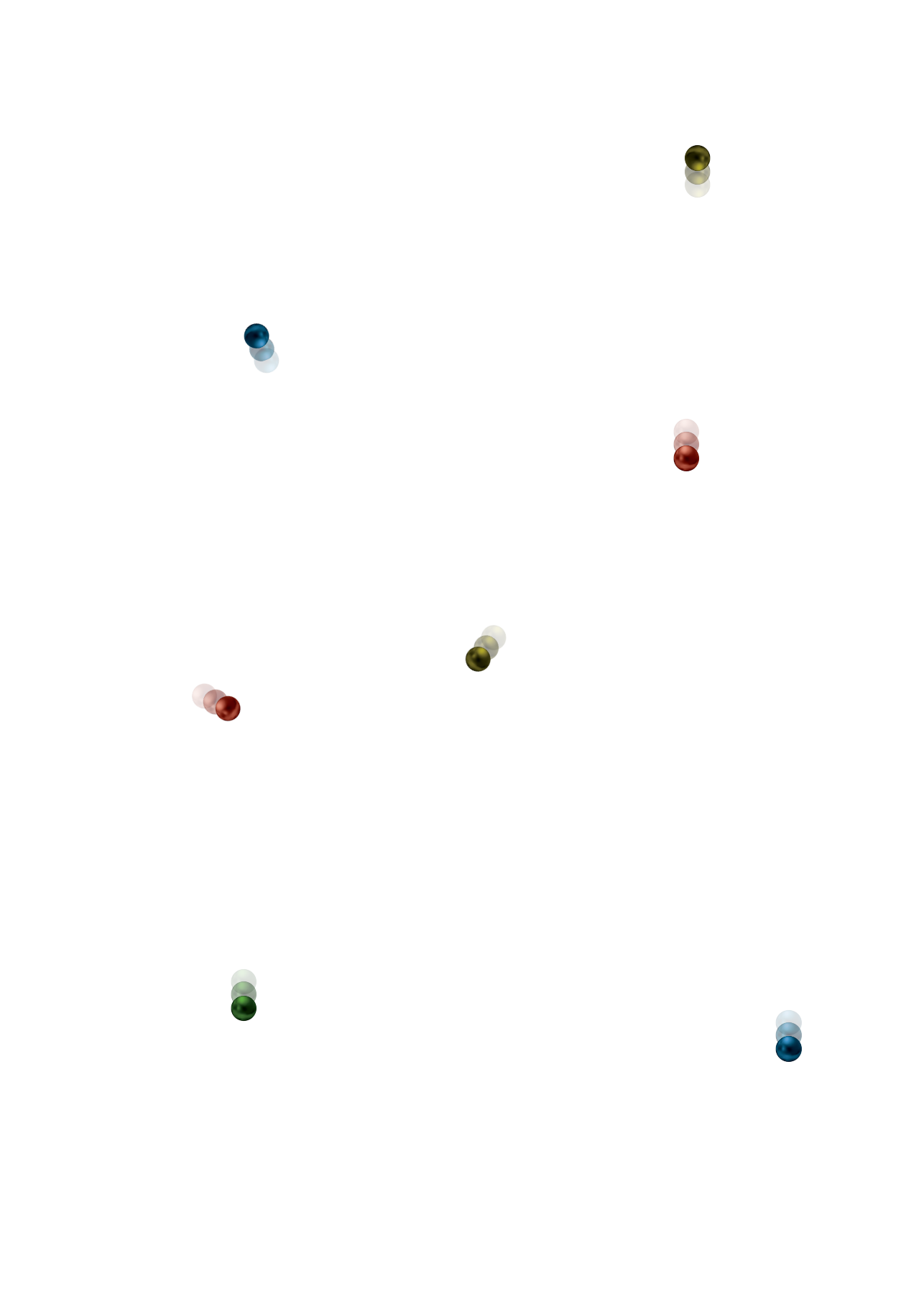
Wammes, D., Slof, B., Schot, W., & Kester, L. (2023). Teacher judgement accuracy of technical abilities in primary education. *International Journal of Technology and Design Education,* 1-24.

Wammes, D., Slof, B., Schot, W., & Kester, L. (2022b). Fostering pre-service primary school teacher's ability to recognise differences in pupils' understanding of technical systems. *International Journal of Technology and Design Education*.

Professional Contributions

https://DoorNieuwsgierigheidLeren.eu (2008).

Verheijen, S., van Koppen, C. S. A., & Wammes, D. F. (2010). Naar een kern voor leerlijnen natuur-en milieueducatie: analyse van bestaande leerlijnen en synthese van een kern-leerlijn NME. Universiteit Utrecht.



DANKWOORD (ACKNOWLEDGEMENTS)

Dit proefschrift is het gevolg van het streven van directeuren zoals Carla van den Bosch ('t Talent – Lent), Nelleke Remerie (Holthuus – Huissen) en Hans Thissen (Lanteerne – Nijmegen) naar een goede invulling van onderwijs over wetenschap en technologie op hun basisschool. Zij stimuleerden mij om onderzoek te gaan doen naar de opbrengst van dat onderwijs op hun scholen. Een promotietraject bood die mogelijkheid. Hanno van Keulen en Jan van Tartwijk (Universiteit Utrecht) hielpen mij bij het opstellen van een onderzoeksvoorstel. Dit voorstel werd goedgekeurd door de HAN en resulteerde vervolgens in een promotiebeurs voor leraren.

Met Liesbeth Kester als promotor en Bert Slof en Willemijn Schot als dagelijks begeleiders startte ik met goede moed in de verwachting dat ik het onderzoek in een sneltreinvaart zou kunnen voltooien. Dat liep anders. Willemijn en Bert wisten mij telkens op subtiele en soms minder subtiele wijze te stimuleren om grondiger na te denken over mijn onderzoek. Van hun begeleiding, maar ook van al het onderzoek dat er werd gedaan en gedeeld op LV3 heb ik veel geleerd. Daardoor is niet alleen mijn onderzoek beter geworden, het heeft ook veel bijgedragen aan mijn kwaliteit als docent op de PABO.

Al was ik een 'buitenbeentje' ik voelde mij altijd heel erg welkom op de donderdagen dat ik in Utrecht een plekje vond in F3.01. Het was altijd mogelijk om vragen te stellen en ervaringen te delen. Met Mare, Sophie, David, Mei, Xiaojing, Jonne, Eva, Jaël, Renée, Sophia, Yuanyuan, Christa, Gesa, Michaela was het er altijd gezellig. Jammer dat dit in de laatste paar jaar niet meer mogelijk was door COVID. Nadat Bert vertrok uit Utrecht en ook Willemijn andere taken kreeg, heeft Liesbeth bij de laatste studie ook de dagelijkse begeleiding op zich genomen. Dat was erg prettig voor de continuïteit.

Mijn dank gaat ook uit naar alle scholen waar ik terecht kon voor het uitvoeren van de verschillende studies en naar de studenten die daarbij hebben geholpen. Op de HAN was het fijn om met mijn mede-promovendus Harry Stokhof het lief en leed van een promotietraject te delen en om te merken dat er bij de collega's altijd veel belangstelling was voor mijn onderzoek en kennis die ik uit Utrecht meenam. Steffie van der Steen wil ik bedanken voor het delen van haar kennis over het werk van Kurt Fischer. Haar werk werd een belangrijke bouwsteen van mijn promotieonderzoek.

Tenslotte ben ik mijn partner Baukje dankbaar dat ze het promotietraject zo lang heeft gedoogd.