International Journal of Technology and Design Education (2022) 32:2577–2609
https://doi.org/10.1007/s10798-021-09697-z

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

Dannie Wammes[1] · Bert Slof[2] · Willemijn Schot[3] · Liesbeth Kester[2]

Accepted: 10 August 2021 / Published online: 20 August 2021

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 test 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.

Keywords Technology education · Primary education · Formative assessment · Evidenced Centered Design · Prior knowledge

✉ Dannie Wammes
dannie.wammes@han.nl

[1] HAN University of Applied Sciences, Nijmegen, The Netherlands

[2] Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands

[3] Educational Consultancy & Professional Development, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands

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 to provide 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 how to assess and foster 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) and 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 for developing a valid diagnostic tool

The evidence-centered design approach

Evidence-Centred Design (ECD) was developed to facilitate a systematic design of large-scale assessments (Mislevy, Steinberg, & Almond, 2003; Roelofs, Emons, & Verschoor, 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 Clarke-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 Fig. 1). The design decisions made in preceding layers serve as input for the decisions made in subsequent layers. 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?

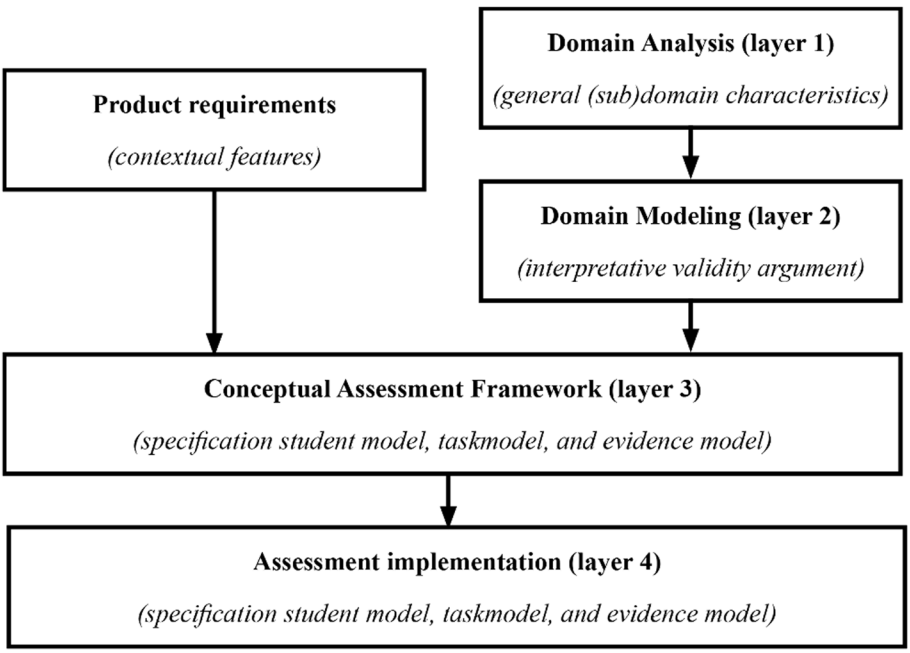

Fig. 1 Evidence-Centered Design (based on Riconscente, Mislevy, & Hamel, 2005)

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 technological systems

Based on the ECD approach, a diagnostic tool aimed at gaining insight into pupils' prior knowledge of technological 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 concrete materials 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 predescribed 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 technological 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" (ITEA, 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 about technological systems is often limited to the ones they are already familiar with (Baumert et al., 1998; CITO, 2015; 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 Fig. 2). 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 why it is reasonable to assume that the diagnostic tool assesses construct-relevant (i.e., understanding of technological systems) instead of construct-irrelevant (e.g., math or reading skills) pupil characteristics. The underlying assumptions should be explicated in the design decisions, which, in turn, should be substantiated by theoretical and, preferable 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.

Decision 1: Enable pupils to make use of their tacit knowledge The domain analysis revealed that pupils' understanding of technological systems is partially 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 by 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

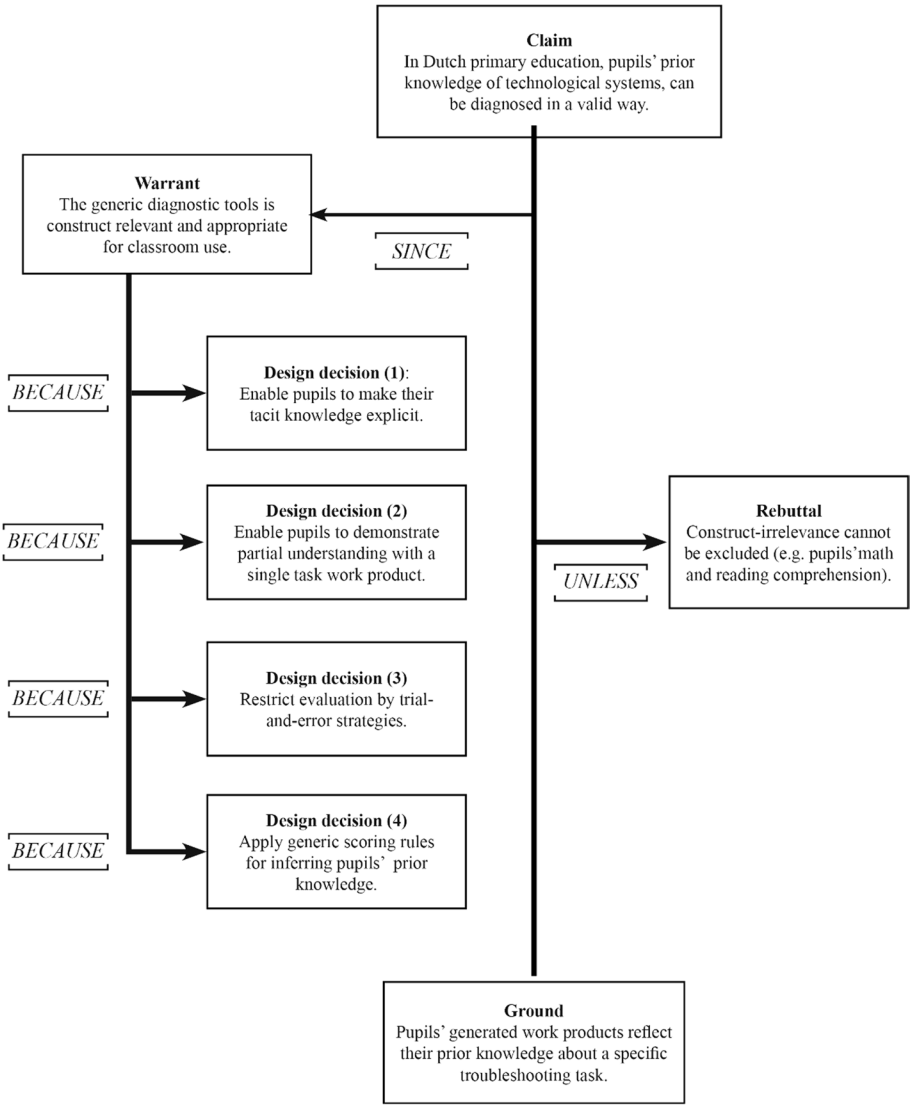

Fig. 2 Overview of interpretative validity argumentation (based on Oliveri et al., 2019)

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 backings. From a theoretical viewpoint, Fisher's (1980) framework for describing the development of dynamic skills might offer relevant guidelines for developing the generic scoring rules (see Fig. 3).

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 action. In the

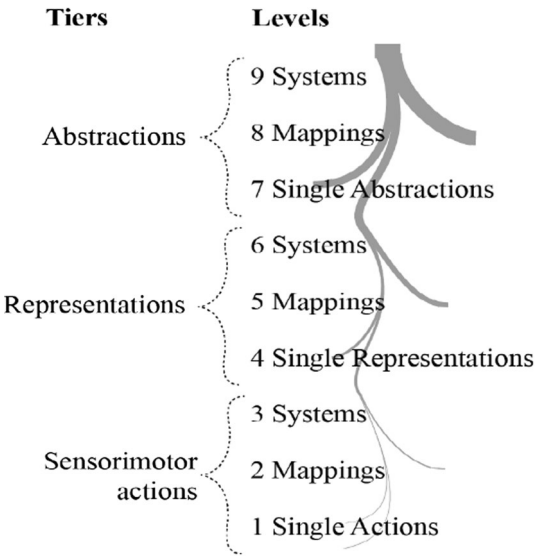

Fig. 3 Dynamic skill development (based on Fischer & Bidell, 2007)

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 systems' components which cannot be observed at 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' 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 is 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 for 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 ability 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. As indicated above, 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' understanding of technological systems. Pupils' levels of understanding manifest itself by the behaviour that they exhibited 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 the tool; (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, 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 for determining 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 all seven targeted levels of understanding should be described. 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 systems that pupils' can construct from scratch. The Buzz-Wire (BW) device (see Fig. 4) is an electrical circuit that has been used in a

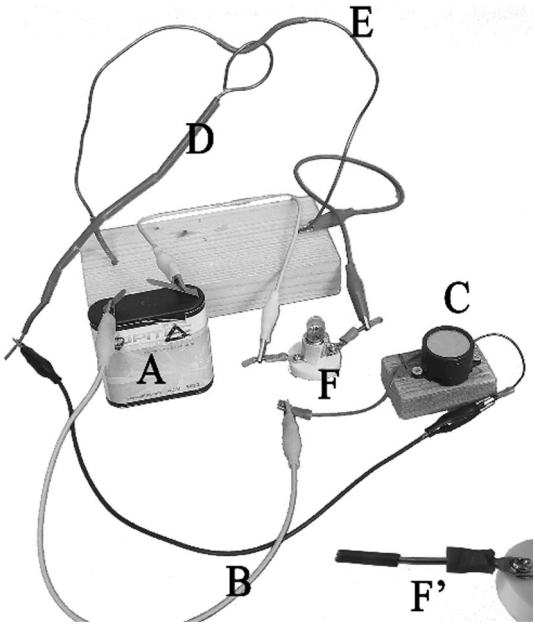

Fig. 4 Buzz-Wire device (A = battery, B = wire with crocodile clamps, C = buzzer, D = loop, E = spiral, and F = lamp with copper-wire connectors, F' = outer points protected by a piece of cable-mantle)

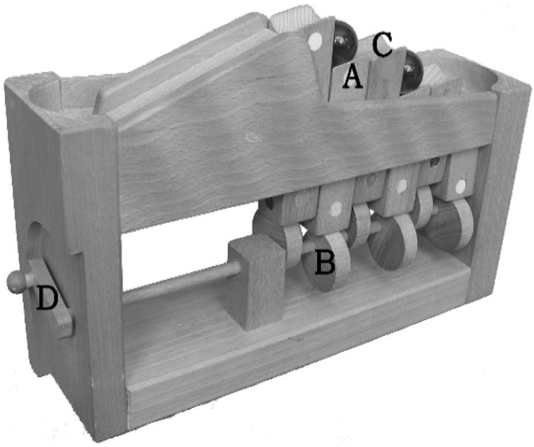

Fig. 5 Stairs Marble Track device (A = marble, B = camshaft eccentric wheel, C = bar with a slanted top, and D = handle to turn the camshaft)

Dutch national study on the quality of science and technology education in primary education (CITO, 2015). 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 Fig. 5), 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, so they were not confused when they did not see the 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 technological 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 for pupils' understanding of electrical and mechanical technological systems. In case math and reading scores are strong predictors for pupils' understanding of technological 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 technological 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, the limitations, and implications for educational practices and research. Ethical approval for this study was provided by the faculties 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. 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 assessments 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 Fig. 5). 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.

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

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	*Circuit* − 1 = no wire connected to any component 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

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 0.1 interval were 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

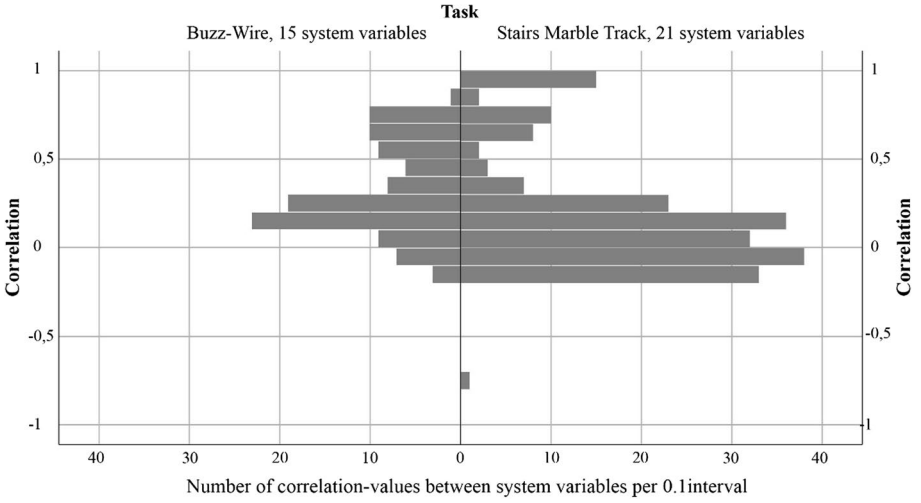

Fig. 6 Left side: Distribution of correlations between the 15 BW variables described in Table 1. Right side: Distribution of correlations between the 21 SMT variables described in Table 2

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 6 shows that 99% of the 86 BW variable correlations were below $r=0.8$ and 65% below $r=0.5$, indicating that pupils do combine the components in various ways. From the 209 correlations between the SMT variables, 92% was below $r=0.8$ and 82% below $r=0.5$. BW correlations above $r=0.50$ 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 to 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.

Conclusion

The results show that both assessment tasks (BW and SMT device) facilitated pupils to combine the device components in various ways and, consequently, in generating a wide variety of work product. 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 on 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 Fig. 7).

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.

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 min)

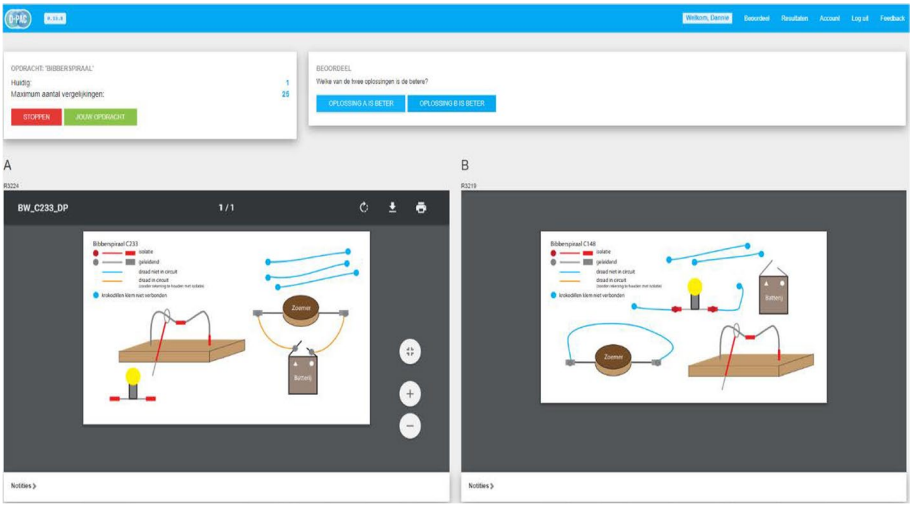

Fig. 7 Screenshot D-PAC tool (Verhavert, 2018)

was held 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 task) was selected by the first author based on 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 technological 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 restest 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 Table 1 and 2). For both sets of items, the SPSS aggregate function was used to create an extended Rasch model table (see supplementary material), of which the subsequent higher-order interactions were the coefficients of the covariates given by $y_r = t(t-1)\ldots(t-r+1)/r!$, for $r = 2,..,a$, where a is the order of the highest constant interaction. Y_2 being the first higher-order interaction, y_3 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.

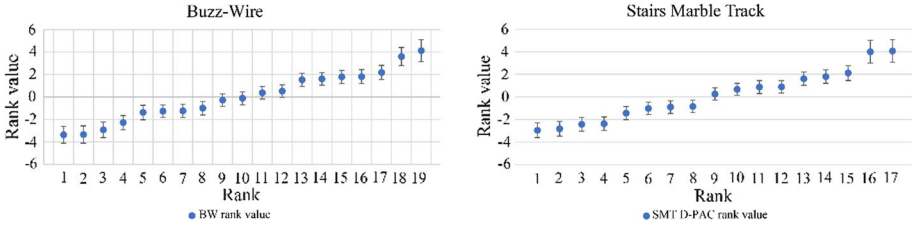

Fig. 8 D-PAC ranking of BW and SMT work products

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 Fig. 8). 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 the 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 position 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).

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	NA
184	2	2	3	2
147	2	2	2	NA
008	3	4	5	2
080	4	4	5	5
172	4	5	5	NA
100	5	5	5	4
127	5	5	5	NA
033	6	7	7	4

Work-product levels by the Fischer experts

The comparison between the initial Fischer scale levels, as labelled by the first author using the scoring rules from Table 1, 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).

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.

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 = 0.875, $p < 0.001$, 95%CI[0.704,0.950]) and 17 SMT work products (ICC = 0.843, $p < 0.001$, 95%CI[0.618,0.940]).

Table 4 Scoring rules for inferring pupils' understanding of technological 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 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 slide 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)

Table 4 (continued)

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.	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 low roll-on and high roll-off point) but not with the slanted top upward (vertically top-down or horizontally)
	If not: go to Sm2	
Sensorimotor mapping (Sm2) Combining features of two objects.		
2	Any connection between two components with a wire.	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)
	If not: go to Sm1	
Single sensorimotor actions (Sm1) Use of 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.

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 0.813, $p = 0.002$ for the BW task and 0.920, $p < 0.001$ 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: $\chi^2 = 7.117$, $df = 10$, $p = 0.71$). For the SMT task, a good model fit was found when three constant higher-order interactions were added to the model ($\chi^2 = 6.127$, $df = 3$, $p = 0.11$). (see supplementary material).

Table 5 Distribution of Fischer-scale levels on the BW and SMT task

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	–	–	–

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 BW and SMT 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.

Table 6 Reading comprehension and mathematics test scores

Assessment	Test grade	N	Reference*	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

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

Table 7 Overview of multiple regression analysis

	Predictors	B[1)]	*b*[1)]	t[1)]	Sig[1)]	Fchange[2)]	Sig. Fchange[2)]	Adj $R^2_{2)}$
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

*Correlation is significant at the 0.05 level (2-tailed)

**Correlation is significant at the 0.01 level (2-tailed)

1) Values of the final model with three predictors, with BCa95%CI

2) Value after predictor was added to the model

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 of 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 1000 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 underlying technological 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 match and reading ability are represented in Table 6 and show that their average scores are slightly higher than the national reference values.

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.

Conclusion

The results show a low predictive value of both pupil's 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 for 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 with 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 technological systems from their generated work products. Since pupils' prior knowledge may differ considerably per device (CITO, 2015; Molnár et al., 2013), 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 Table 1). 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 the 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 task specific 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 technological 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 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, Zetterqvist, & Ingerman, 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 technological 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 and 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 systems' structure through interaction. See, for instance, 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 pupils' prior knowledge of different systems. The level resulting from a work-product only indicates prior knowledge about the technical 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—was 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 to the BW device. Although, after the 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 technological systems. It might, for example, also be feasible to assume that a pupils' motivation affects his/her task engagement and academic achievement (Hornstra et al., 2020; Schunk, Meece, & Pintrich, 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 technological 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, 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).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10798-021-09697-z.

Funding This study was supported by a grant (023.007.027) from the Netherlands Organisation for ScientificResearch (NWO) and with the cooperation of the HAN University of Applied Sciences.

References

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

Arcia, G., Macdonald, K., Patrinos, H. A., & Porta, E. (2011). *School autonomy and accountability.* SABER. World Bank. Retrieved from 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

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

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

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.

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. https://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

Cianciolo, A. T., Matthew, C., Stenberg, R. J., & Wagner, R. K. (2006). *Tacit knowledge, practical intelligence, and expertise. Handbook of expertise and expert performance*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.035

CITO. (2015). 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.

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.
Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). Psychometric considerations for the next generation of performance assessment. *Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service*
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. (2005). *Teaching about technology: An introduction to the philosophy of technology for non-philosophers*. Dordrecht: 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.
Department for Education. (2013). *The national curriculum in England*. London: Crown.
Dochy, F., Moerkerke, G., & Martens, R. (1996). Integrating assessment, learning and instruction: Assessment of domain-specific and domain transcending prior knowledge and progress. *Studies in Educational Evaluation, 22*(4), 309–339.
Feenstra, H., Kleintjes, F., Kamphuis, F., & Krom, R. (2010). *Wetenschappelijke verantwoording begrijpend lezen groep 3 t/m 6 scientific account for the reading comprehension tests grade 1 to 4*. Arnhem, the Netherlands: Cito.
Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 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
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
Garmine, E., & Pearson, G. (Eds.). (2006). *Tech tally; approaches to assessing technological literacy*. Washington D.C: National Academic 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
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.
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
Harlen, W. (2008). *Science as a key component of the primary curriculum: A rationale with policy implications.* (No. 1). Retrieved from www.wellcome.ac.uk/perspectives. (science primary education)
Harlen, W. (2012). *Developing policy, principles and practice in primary school science assessment*. London: 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.
Hedlund, J., Antonakis, J., & Sternberg, R. J. (2002). *Tacit knowledge and practical intelligence: Understanding the lessons of experience*. Alexandria: DTIC Document.
Heitink, M. C., van der Kleij, F. M., Veldkamp, B. P., Schildkamp, P., & 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
Hessen, D. J. (2011). Loglinear representations of multivariate Bernoulli Rasch models. *British Journal of Mathematical and Statistical Psychology, 64*(2), 337–354.
Honey, M., Pearson, G., & Schweingruber, H. A. (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. Washington DC: National Academies Press.
Hornstra, T. E., Bakx, A., Mathijssen, S., & Denissen, J. J. (2020). Motivating gifted and non-gifted students in regular primary schools. *Learning and Individual Differences., 80*, 101871.
ITEA (2007). *Standards for technological literacy: Content for the study of technology* (3rd Ed.) International Technology Education Association.
Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenen-wiskunde voor groep 3 tot en met 8 scientific justification of the mathematics test for grade 1 until grade 6*. Arnhem, Netherlands: Cito.
Johnson, S. D. (1995). *Understanding troubleshooting styles to improve training methods Paper presented at the American Vocational Association Convention*. Denver CO: ERIC.

Jonassen, D. H. (2010). *Learning to solve problems: A handbook for designing problem-solving learning environments*. New York: Routledge.
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
Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135–170.
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, USA, Taiwan, Australia*. UK: 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
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
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
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. https://doi.org/10.2307/2529310
Levy, S. T. (2012). Young children's learning of water physics by constructing working systems. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-012-9202-z
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.
Malik, X. (2014). *The future of Europe is science*. Luxembourg: Publications Office of the European Union. https://doi.org/10.2796/28973
Matan, A., & Carey, S. (2001). Developmental changes within the core of artefact concepts. *Cognition, 78*(1), 1–26. https://doi.org/10.1016/S0010-0277(00)00094-9
McFadden, A., & Williams, K. E. (2020). Teachers as evaluators: Results from a systematic literature review. *Studies in Educational Evaluation, 64*, 100830. https://doi.org/10.1016/j.stueduc.2019.100830
Meindertsma, H. B., Van Dijk, M. W., Steenbeek, H. W., & Van Geert, P. L. (2014). Assessment of preschooler's scientific reasoning in Adult-Child interactions: What is the optimal context? *Research in Science Education, 44*(2), 215–237.
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*. United States: 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
Nitko, A. J. (1996). *Educational assessment of students*. CA: ERIC.
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18.
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*. Paris, France: OECD Publishing. https://doi.org/10.1787/9789264190511-en
OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems*. Paris, France: OECD Publishing. https://doi.org/10.1787/9789264208070-en
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, 19*(3), 270–300.
Parziale, J. (2002). Observing the dynamics of construction: Children building bridges and new ideas. *Microdevelopment: Transition Processes in Development and Learning*, 157–180.
Pearson, G., & Young, A. T. (2002). *Technically speaking: Why all Americans need to know more about technology*. Washington DC: 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

Philpot, R., Ramalingam, D., Dossey, J. A., & McCrae, B. (2017). Factors that influence the difficulty of problem-solving items. In B. Csapó & J. Funke (Eds.), *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning* (pp. 141–158). OECD Publishing: Paris. https://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. Retrieved from www.platformbetatechniek.nl

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

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

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.

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/10.1080/13664530.2014.900516

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.

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

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

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. https://doi.org/10.1007/2Fs10763-013-9461-5

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.

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

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.

Schwartz, M. (2009). Cognitive development and learning: Analyzing the building of skills in classrooms. *Mind, Brain, and Education, 3*(4), 198–208. https://doi.org/10.1111/j.1751-228X.2009.01070.x

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

Siuty, M. B., Leko, M. M., & Knackstedt, K. M. (2018). Unravelling the role of curriculum in teacher decision making. *Teacher Education and Special Education, 41*(1), 39–57.

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

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. Retrieved from http://hdl.handle.net/11370/408b8e4e-2be4-4312-a48a-8898995dc273

Svensson, M., Zetterqvist, A., & Ingerman, Å. (2012). On young people's experience of systems in technology. *Design & Technology Education, 17*(1)

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.

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

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.

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 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

Verhavert, S. (2018). *Beyond a mere rank order: The method, the reliability and the efficiency of comparative judgment (unpublished doctoral thesis)* Available from repository.uantwerpen.be.

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*. 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*. https://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://doi.org/10.1037/0022-3514.49.2.436

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

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

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*. 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/10.1016/S0191-491X(99)00022-X