International Journal of Technology and Design Education (2023) 33:415–438
https://doi.org/10.1007/s10798-022-09734-5

Teacher judgement accuracy of technical abilities in primary education

Dannie Wammes[1] · Bert Slof[2] · Willemijn Schot[3] · Liesbeth Kester[4]

Accepted: 25 January 2022 / Published online: 14 February 2022

Abstract

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

Keywords Engineering · Primary education · Judgement accuracy · Formative assessment

This Study Was Supported By a Grant (023.007.027) from the Dutch Research Council (NWO) and Was Conducted in cooperation with HAN University of Applied Sciences.

✉ Dannie Wammes
dannie.wammes@han.nl

[1] HAN University of Applied Sciences, HAN PABO, Kapittelweg 35, 6525 EN Nijmegen, The Netherlands

[2] Netherlands Institute for Curriculum Development, Enschede, The Netherlands

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

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

Introduction

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

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

This study explores the merits of using diagnostic tasks in relation to teachers' judgement accuracy regarding their pupils' technical abilities. First, the challenges around teachers' judgement accuracy of their pupils' technical abilities will be discussed. Second, it is explained how teacher and pupil characteristics might influence teachers' judgements, pupils' performance and the associated judgement accuracy. Finally, teachers' views on the use of diagnostic tasks will be discussed.

Teacher judgement accuracy

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

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

Scientific research has proposed tasks that enable pupils to apply their knowledge, including its tacit aspects, in the domain of science and technology to diagnose technical ability (Hast, 2020; Swaak & de Jong, 1996). This study used two tasks about technical systems that allow pupils to interact with the materials in various ways (Wammes et al., (2021). Although these tasks can only capture a limited part of the technical skills spectrum, they provide objective data that gives us the first glimpse of teacher judgement accuracy in this domain. This is the first research question of the present study: How accurate are teacher judgements about pupils' prior knowledge of technical systems compared with the results of diagnostic tasks?

Teacher and learner characteristics

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

For engineering, it has been reported that teachers' ability to gain insight into their students' technical skills is related to their technical knowledge (Compton & Harwood, 2005; van Niekerk et al., 2010) and self-efficacy beliefs regarding teaching in this domain (Jones & Moreland, 2004). Teaching experience is likely of secondary importance (Nadelson et al., 2013; Hsu et al., 2011).

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

Use of diagnostic tasks

The tasks used in this study were developed for formative use in primary schools (Wammes et al., 2021). Whether primary school teachers will use these tasks depends partly on the effort and time required to use them in the classroom and partly on the value that teachers attribute to the data they produce (Kirton et al., 2007). That value will depend on the insight gained into pupils' technical skills, the possibilities a teacher sees for using that insight to adapt technology education, and the teacher's expectations about the impact of a more tailored approach on developing pupils' technical skills (Praetorius et al., 2017). The possibilities that teachers see for applying the acquired insight depend not only on personal didactic qualities but also on the freedom teachers have to design their lessons within the curriculum (Sach, 2015). The final research question of this study addresses the value of the diagnostic tasks from the teacher's perspective: How do teachers value the diagnostic tasks for their judgement and teaching practice?

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

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

Method

Participants

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

Measurements

Pupil performance

The question, 'How accurate are teachers' judgements about pupils' prior knowledge of technical systems as compared with the results of diagnostic tasks?' was explored by comparing teacher judgements with their pupils' performance on two diagnostic tasks: the Buzz-Wire (Fig. 1) and the Stairs Marble Track (Fig. 2). These tasks are non-verbal, allow pupils to change system variables independently, take only a few minutes to accomplish and allow pupils' knowledge to be assessed on a generic scale (i.e., Fischer, 1980).

The Buzz-Wire (BW) is an electric circuit (see Fig. 1). It has a switch made of a copper spiral and a ring with a handle. Touching the spiral with the ring closes the circuit and activates a lamp and buzzer. The Stairs Marble Track (SMT) is a mechanical device that transports marbles upwards to a descending track that brings the marbles back to the start (see Fig. 2). The mechanism is a camshaft with six eccentric wheels and six bars of increasing length with a slanted top.

Both tasks are introduced with a one-minute video that shows how the devices should function without revealing their mechanism or construction. When pupils start the

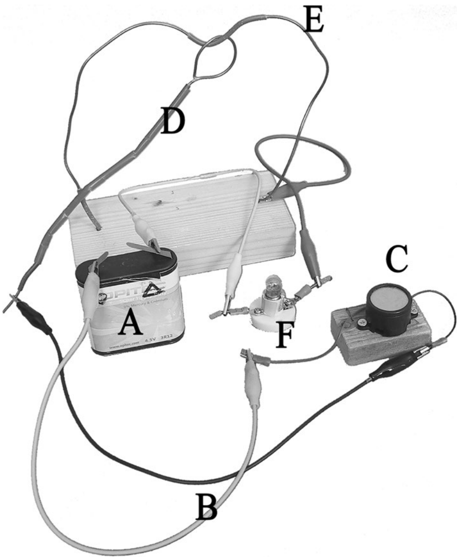

Fig. 1 Buzz-Wire. **a** Battery, **b** Wire with crocodile clips, **c** Buzzer, **d** Loop, **e** Copper spiral, **f** Lamp with isolation on outerside connectors (Wammes et al., 2021)

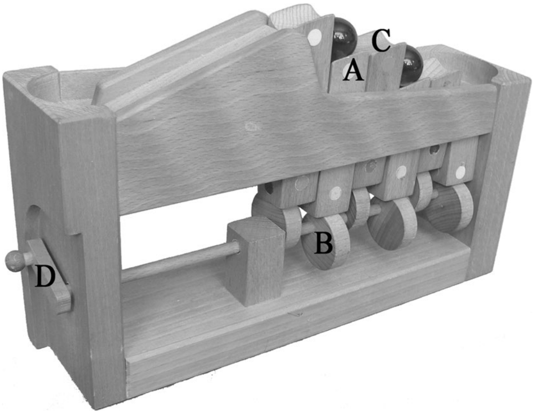

Fig. 2 Stairs Marble Track. **a** Marble (not available in task), **b** Camshaft with six eccentric wheels, **c** Bar with slanted top, **d** Handle, blocked (Wammes et al., 2021)

Buzz-Wire task, all parts are spread out on the table. The SMT task begins with the six bars scattered randomly on the table. In both cases, the pupils' task is to restore the devices.

The scale used to evaluate the pupils' ability levels was developed by Fischer (1980). The task-specific elaborations (see Appendix 1) were validated by Wammes et al. (2021). The Fischer scale describes how skills build up to more complex skills. A pupil's work product has performance level 1 when it results from a single action based on sensorimotor information. A work product that results from a combination (mapping) of two sensorimotor-based actions indicates a level 2 skill. When the work product results from the combined use of multiple sensorimotor actions, it points to a level 3 (system) skill.

The repeated use of sensorimotor actions results in the ability to remember causal or other relationships between task components. In terms of Fischer: it results in a (causal) relationship *representation*. The tasks restrict the possibility to find causal or other relationships between task components otherwise, for instance, by trial and error. When a work product reflects the use of a single representation, the pupil demonstrates a level 4 skill.

Level 5 skills are evident in work products that require the combination of two representations. At level 6, pupils' work products show that their actions are based on the combined use of multiple representations. At the next level (7), the work-products indicate that pupils can apply a generic feature of a phenomenon to solve an unfamiliar problem. The correct completion of the SMT task requires a level 6 skill, while the BW task allows pupils to demonstrate their understanding of electrical circuits up to level 7.

Teacher judgements

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

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

Teacher and learner characteristics

Our second research question was: How do teacher and learner characteristics relate to judgement accuracy of technical abilities? First, the measurement and scoring of the teacher characteristics, and then the measurement of the pupil characteristics will be described.

For all teachers, their gender, teaching experience and the time they had worked with the pupils were recorded. The teachers' technical knowledge was assessed using a multiple-choice test. This Technical Knowledge Test (TKT) comprised 43 items about technology from 2015, 2016 and 2017 editions of a Dutch national assessment on Science and Technology. This is an admission test for students who want to become primary school teachers. Cronbach's alpha of the TKT was 0.78.

Teachers' self-efficacy beliefs were administered using an adapted version of the Science Teaching Efficacy Belief Instrument (STEBI-b) (Bleicher, 2004; Riggs & Enochs, 1990). The references to science in the statements were changed to references to technology. For example, "I will continually find better ways to teach science" was changed to "I will continually find better ways to teach technology". The instrument combines the Personal Teaching Efficacy scale (PTE, 13 items) and the Teaching Outcome Expectancy scale (TOE, ten items). The PTE is about personal efficacy beliefs, which relate to teacher judgment accuracy (Nadelson et al., 2013). The TOE refers to ideas about the effectiveness of science, technology, and education in general. The STEBI-b was fully administered in this study, but only the PTE score was used as that scale indicates the teachers' self-efficacy. Cronbach's alpha of the PTE was 0.72.

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

Teachers' evaluation of the diagnostic tasks

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

Procedure

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

Scoring and analyses

Teacher judgement accuracy

The performance$_{\mathrm{BW}}$ and performance$_{\mathrm{SMT}}$ of pupils' work products were determined using the scoring rules of Appendix 1 (Wammes et al., 2021). A second rater independently assessed the work products of the pupils of one school. Inter-rater agreement was calculated in SPSS using the two-way mixed model of the Intraclass Correlation Coefficient (ICC) with the absolute agreement option. For 33 BW work products, the ICC was 0.85, and for 38 SMT work products, the ICC was 0.96, which is good.

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

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

Teacher and learner characteristics

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

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

There was a strong correlation between teachers' absolute judgements of their pupils' performance on the SMT and the BW. ($r = 0.811$). Therefore, we decided to use the average teachers' judgements over the two tasks as the dependent variable for analysis in SPSS with a set of two-level hierarchical models that nested learners (level 1) within teachers (level 2). Model 0 was unconditional and used to establish the amount of variance in judgements between teachers and within the teachers' classes. Model 1 added pupils' performance on both tasks as predictors. Model 2 included the other learner characteristics, and Model 3 introduced the teacher characteristics. SPSS hierarchical regression analysis was used for the relative judgements.

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

Teachers' evaluation of the diagnostic tasks

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

Results

Teacher judgement accuracy

The ICC values in Table 1 show considerable differences in judgement accuracy. The relative judgements were generally the most accurate, followed by the absolute judgements$_{\mathrm{SMT}}$. The absolute judgements$_{\mathrm{BW}}$ were the least accurate.

The frequencies of the deviations between judgement and performance in Table 2 reveal that underestimation of pupils' technical abilities occurred more frequently than overestimation. For three teachers, their underestimation was significant.

Teacher and learner characteristics

The second research question explored which teacher and learner characteristics might relate to the differences in judgement accuracy. Table 3 provides an overview of the characteristics of the teachers and their judgement accuracy.

The unconditional model of the multilevel analyses in Table 4 shows, with an ICC of 0.25, that 25% of the variance in judgements can be attributed to differences between the teachers and 75% to within-class judgement variance. Pupils' performance introduced in model 1 explained 20.7% of the variance, but only the SMT task explained a significant proportion, implicating that the teachers' judgements are especially biased for the BW

Table 1 Primary school teachers' judgement accuracy for technical ability

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

[a] Accuracy = ICC two-way mixed, absolute agreement

[b] Teacher 1 had the youngest pupils and selected only those with high general performance

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

Table 2 Differences between absolute judgements and task performance

Teacher		Average Fischer scale levels				Frequency of absolute judgement–performance deviations in levels					Paired Samples T-Test judgements – performance (BW and SMT)
Pupils		BW		SMT		over-estimate			under-estimate		
	N	abs-j$_{BW}$	perf$_{BW}$	abs-j$_{SMT}$	perf$_{SMT}$	≤ − 2	− 1	0	1	≥ 2	
1	9	$\bar{x}$=4.8, sd = 1.1	$\bar{x}$=4.2, sd = 1.2	$\bar{x}$=3.1, sd = 1.7	$\bar{x}$=4,1, sd = 0.6	3	4	3	2	6	t = .489 p = .631
2	23	$\bar{x}$=3.0, sd = 1.7	$\bar{x}$=3.7, sd = 1.6	$\bar{x}$=3.6, sd = 1.7	$\bar{x}$=3.4, sd = 1.4	7	10	10	7	12	t = -1.067 p = .292
3	29	$\bar{x}$=2.6, sd = 1.0	$\bar{x}$=3.5, sd = 1.4	$\bar{x}$=2.3, sd = 1.0	$\bar{x}$=3.3, sd = 1.0	1	5	14	21	17	t = 5.716** p < .001
4	25	$\bar{x}$=4.6, sd = 2.2	$\bar{x}$=4.7, sd = .9	$\bar{x}$=4.3, sd = 1.9	$\bar{x}$=4.5, sd = 1.0	11	8	11	6	14	t = .462 p = .646
5	21	$\bar{x}$=2.2, sd = .9	$\bar{x}$=4.0, sd = 1.5	$\bar{x}$=2.5, sd = .9	$\bar{x}$=4.0, sd = 1.1		4	8	7	23	t = 6.529** p < .001
6	17	$\bar{x}$=4.8, sd = 1.2	$\bar{x}$=4.5, sd = 1.7	$\bar{x}$=4.2, sd = 1.4	$\bar{x}$=3.9, sd = 1.0	9	8	7	5	5	t = -1.172 p = .249
7	25	$\bar{x}$=4.5, sd = 1.3	$\bar{x}$=3.8, sd = 1.4	$\bar{x}$=4.2, sd = 1.3	$\bar{x}$=4.1, sd = 1.1	12	13	13	5	7	t = -1.756 p = .085
8	29	$\bar{x}$=4.4, sd = 1.4	$\bar{x}$=5.1, sd = 1.5	$\bar{x}$=3.8, sd = 1.2	$\bar{x}$=4.7, sd = 1.2	4	8	16	11	19	t = 3.622** p = .001
All	178	$\bar{x}$=3.8, sd = 1.7	$\bar{x}$=4.2, sd = 1.5	$\bar{x}$=3.5, sd = 1.6	$\bar{x}$=4.0, sd = 1.2	47	60	82	64	103	t = 4.800** p < .001

abs-j. = absolute judgement; perf. = task-performance level

** Significant at the 0.001 level (2-tailed)

Table 3 Teacher characteristics ordered by relative judgement accuracy [a)]

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

a ICC two-way mixed, absolute agreement, T=teacher, G=teacher's gender, N=number of pupils, Exp.=Years of experience as a teacher, Pm.=months of exposure to pupils, rel-j acc.=relative judgement accuracy, abs-j acc.=absolute judgement accuracy,

b negative average covariance, violating reliability model assumptions,

* Correlation significant at the 0.05 level (2-tailed), ** Correlation significant at the 0.001 level (2-tailed)

task. The introduction of the learner characteristics in model 2 reduced the unexplained between-learner variance, increasing the proportion of unexplained between-teacher variance to 34%. Being controlled for performance, model 2 reveals that pupils' gender biases teacher-judgements. Teacher judgements for boys were higher than for girls. Boys did indeed outperform girls, but only on the SMT task where the teachers' mean difference was less than expected (0.61 level for performance, 0.84 for judgements).

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

The hierarchical regression analyses on the relative judgements showed that introducing pupils' performance rank in model 1 explained 15.9% of the variance. The introduction of the learner characteristics explained a further 16.5%. The R square change resulting from the introduction of teacher characteristics was 0.8% and non-significant. The final model showed that there was a similar gender bias (b=-0.17, p<0.001) and reading comprehension (b=0.09, p=0.002) as for the absolute judgements. In contrast with the absolute judgements, there was no significant bias for mathematics scores (b=0.04, p=0.194) but a significant bias for age (b=-,06, p=0.007), indicating that older pupils were less present in the higher ranks as judged by their teacher, whereas performance showed a small age-related increase.

For one class, the recording failed. Open coding of the thinking aloud recordings of the ranking of the remaining 145 pupils resulted in four categories: Learning behaviour (e.g., concentration, perseverance, posing questions), Cognitive ability (e.g., math scores; remarks like a 'clever' or 'average' pupil), Science and Technology (specific references to interest in science and technology), and Support at home (e.g., ' It's likely that there is no interest for technology at home'). Another feature of the explanations was their value;

Table 4 Results from the multilevel regression models of teacher judgement

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

Table 4 (continued)

Model 0		Model 1		Model 2Model 2		Model 3	
Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
χ^2 change		34.58***		65.06***		11.11*	
-2LL	628.684		594.101		529.043		517.938
AICc	634.822		604.450		548.128		543.852

$^{*}p < .05$,;$^{**}p < .01$, $^{***}p < .001$

they were positive, e.g., 'a real go-getter', neutral, e.g., 'average cognitive ability' or negative, e.g., 'not interested in technology'. For 75% of the pupils, explanations were given that addressed more than one category, e.g., 'Not a smart boy per se, but he is really good at analysing structures (categorised as Cognitive ability—neutral), but from him, I expect high performance in technical subjects' (categorised as Science and Technology—positive)'. Inter-rater agreement for the explanation categories was $\kappa = 0.726$, and for the value ratings (positive, neutral or negative), it was 0.857.

Table 5 shows that, in line with the bias found on reading comprehension and math scores, teachers tend to overestimate the importance of cognitive ability as an indicator of technical ability. In 60% of the explanations about high-ranked pupils by judgment, the teacher mentioned their high cognitive ability. Among those high-ranked by their performance, there were fewer (39%) pupils for whom a high cognitive ability had been mentioned. For the low-ranked judgements, the teachers mentioned a low cognitive ability for 39% and a high cognitive ability for 8% of the pupils. Contrary to these expectations, a low performance lacked any relationship to cognitive ability, as mentioned by the teachers (see Table 5: Low-rank; av-perf).

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

Teachers' evaluation of the diagnostic tasks

In the interviews, the teachers were especially surprised by pupils with non-expected high task performance. Illustrative is the reaction of teacher 5: "For K30, I thought it would be nothing. This was pure because of the image I have of her. She scored significantly higher than I expected". Another example comes from Teacher 7, "I immediately noticed K24. I had given her the lowest score for the BW task, and then she scores six". Teachers 3 and 8 found it striking that only a few students had lower scores than expected, which corresponds with their significant systematic underestimation shown in Table 2. Some teachers expressed that they were glad that the results confirmed their suspicions. "I knew that K36 was very interested in technology, but I found it difficult to estimate whether he could do it. From the results, I see that he is technically skilled." (teacher 7). "I am delighted to receive confirmation that my view of the students is broadly correct." (teacher 4). Teachers also related pupils' task performance level with their reading comprehension and mathematics abilities. Teacher 6 commented: "If you look at K17. Her test scores do not indicate that she is a brilliant student, but she is knowledgeable. She scored highly on both tasks. That certainly says something about her." Teacher 7 wondered why three pupils did the task so well while they had such low mathematics scores.

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

Table 5 Teachers' explanations by pupils' [predicted] and performance rank

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

rel-j = rank by relative judgement, av-perf = rank by average task performance level

[a] Number of pupils with explanations of this category

[b] within top 25th percentile range,

[c] below 75th percentile range

[d] Percentage of pupils of this rank for whom their teachers expressed this type of explanation

[e] Statements about 75% of the pupils included explanations from more than one category, causing the sum of column percentages to exceed 100%

experience to use the results of the diagnostic tasks. Four teachers would communicate the results to pupils and their parents.

Discussion

Teacher judgement accuracy

The participating teachers differed in their judgement accuracy. With one exception, these were below the mean of r = 0.64 reported by Südkamp et al. (2012) for language and mathematics. For three of the eight teachers, there was a significant systematic underestimation of the task performance level of their students.

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

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

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

Teacher and learner characteristics

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

The higher expectations of boys' performance compared to girls were partly confirmed by the better performance of boys on the SMT task. However, they did not correspond to the equal performance of boys and girls on the BW task. The higher estimation of boys' technical ability has been described before (Buccheri et al., 2011; Seiter, 2009). This might be domain-specific since primary school teachers usually estimate girls' skill level to be higher (Timmermans et al., 2015). The better performance of boys on the SMT task may be related to the Spatio-temporal skill that this task requires. Tasks that require Spatio-temporal skills are known to be performed better by boys than by girls (Reilly et al., 2017; Wang, 2017).

The correlation of teacher judgements with pupils' proficiency scores in reading comprehension and mathematics, and the absence of such a correlation of these proficiency scores with pupils' task performance, resulted in a bias in judgement accuracy. The lack of such a correlation is in line with the estimates made by Wammes et al. (2021), who reported that scores on reading comprehension and mathematics explained no more than 10% of the variance in pupils' scores on the BW and SMT task.

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

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

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

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

Teachers' evaluation of the diagnostic tasks

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

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

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

Implications

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

Limitations and future research

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

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

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

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

Conclusion

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

Appendix I

Scoring rules for the Buzz-Wire and Stairs Marble Track tasks.

Start at the highest level and work downwards.

Level	Buzz-Wire	Stairs Marble Track
Single abstractions (Rp4/Sa1) An abstraction is used to arrive at 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). {[mapping iii]iiiiv} 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		

Level	Buzz-Wire	Stairs Marble Track
6	The work product demands the combination of multiple representations 1. An electric circuit in which both lamp and buzzer function by default *and* all connections on metal OR 2. Loop and spiral connected as a switch that turns a lamp or buzzer on/off, *disregarding* whether all connections are on metal	The correct configuration. It demands the combination of all representations i. All bars ordered according to their length ii. The orientation of the slanted bar tops potentially allows a marble to roll onto the slanted top of an adjacent bar iii. The slope of all slanted tops will cause the marble to roll down in the direction of the high roll-off point iv. A correct estimate of the effect of a turning camshaft on the movement and height of adjacent bars
	If not: go to Rp2	
Representational mappings (Rp2) Causal relationships with an intermediate step, linking single causal relations		
5	The outcome contains a mapping 1. A connected lamp or buzzer will function in an electric circuit, disregarding whether all connections are on metal OR 2. 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	The outcome contains a mapping: All bars are in the correct order, *and* the direction of all the bar-top slides allows a marble to roll onto the slide of an adjacent bar (any correct combination of correct or 180°- rotated slide positions) OR The slope of all slanted 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)
	If not: go to Rp1	
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 i. There is a connection from one pole of the battery to the other pole through at least one other component (lamp, buzzer, loop, spiral) ii. Both poles of the lamp or buzzer are connected to the battery iii. All connections should be on metal, conducting electricity. Both poles of the lamp should be connected in this way iv. The ring and spiral are linked. Not directly, but via at least one other component	Use of a single representation i. All bars in the frame are ordered by their length at their correct position in the frame OR ii. The direction of all slides potentially allows a marble to roll onto the slide of an adjacent bar. (any combination of correct or 180° rotated slide positions)
	If not: go to Sm3	

Level	Buzz-Wire	Stairs Marble Track
Sensory-motor 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 are positioned to fill the gap in the frame between the roll-on and roll-off point i. There are bars vertically positioned in the frame with the slanted tops upwards (but bars missing or at least one slanted top rotated by 90° or 270°) OR ii. 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	
Sensory-motor 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) i. At least one combination [mapping] of a single property (e.g. length or top-shape) of two or more bars OR ii. 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 sensory-motor actions (Sm1) Use of single feature of an object or task. Observable		
1	The work product does not include a connection between two components by a wire	The work product does not include a combination of bar or bar and frame features

Generic rule (Van der Steen, 2014)

References

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). *Educational psychology: A cognitive view* (Vol. 6). Austin: Holt, Rinehart and Winston.

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

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

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

Buccheri, G., Gürber, N. A., & Brühwiler, C. (2011). The impact of gender on interest in science topics and the choice of scientific and technical vocations. *International Journal of Science Education, 33*(1), 159–178. https://doi.org/10.1080/09500693.2010.518643

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

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. https://doi.org/10.1007/s10798-004-5401-6

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

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

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

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

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

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

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

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. https://doi.org/10.1007/s10798-014-9285-9

Hast, M. (2020). "It is there but you need to dig a little deeper for it to become evident to them": Tacit knowledge assessment in the primary science classroom. *Diversifying Learner Experience*. https://doi.org/10.1007/978-981-15-9861-6_2

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

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

International Technology Education Association. (2007). *Standards for technological literacy: Content for the study of technology* (E3 ed). Reston: International Technology Education Association.

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

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

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

Kärkkäinen, K., & Vincent-Lancrin, S. (2013). *"Sparking Innovation in STEM Education with Technology and Collaboration: A Case Study of the HP Catalyst Initiative"*. OECD Education Working Papers, No. 91, OECD Publishing. https://doi.org/10.1787/5k480sj9k442-en

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

Klug, J., Bruder, S., & Schmitz, B. (2016). Which variables predict teachers diagnostic competence when diagnosing students' learning behavior at different stages of a teacher's career? *Teachers and Teaching, 22*(4), 461–484.

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

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

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

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

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

National Assessment Governing Board. (2013). *Technology and engineering framework for the 2014 NAEP*. Retrieved from http://www.nagb.gov

Nitko, A. J. (1996). *Educational assessment of students*. Prentice-Hall Order Processing Center.

OECD. (2014a). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problem*s (volume V). OECD Publishing. https://doi.org/10.1787/9789264208070-en

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

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

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

Praetorius, A. K., Berner, V. D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *The Journal of Educational Research, 106*(1), 64–76. https://doi.org/10.1080/00220671.2012.667010

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

Rasinen, A. (2003). An analysis of the technology education curriculum of six countries. *Journal of Technology Education, 15*(1), 31–47.

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

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

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

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]. Expertisecentrum Nederlands, Nijmegen.

Schrader, F., & Helmke, A. (2001). Alltägliche leistungsbeurteilung durch lehrer [Daily performance judgements by teachers]. *Leistungsmessungen in Schulen, 2*, 45–58.

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

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

Stang, J. (2016). *Zur urteilsgenauigkeit von mathematiklehrkräften: Genauigkeitsbeeinflussende faktoren, stabilität und auswirkungen* [Judgement accuracy of teachers of mathematics: Factors influencing accuracy, consistency and impact] (Doctoral dissertation, Universität Passau).

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

Swaak, J., & de Jong, T. (1996). Measuring intuitive knowledge in science: The development of the what-if test. *Studies in Educational Evaluation, 22*(4), 341–362. https://www.learntechlib.org/p/81580/.

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

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

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

Urhahne, D., & Wijnia, L. (2020). A review on the accuracy of teacher judgments. *Educational Research Review.* 100374.

van Eijk, R., Evers, J., Haan, F., Klapwijk, T., Kooistra, B., Klink, I., de Kraker, L., Majoor, D., van Ommen, J., Snoek, M., Versfelt, J., Wassink, G-J., & Wouda, S. (Eds.) (2015). *Development prospects and careers of teachers: advice from the Critical Friends of the Teachers Agenda to the Minister and State Secretary.* The Critical Friends of the Teachers’ Agenda. Amsterdam University of Applied Sciences.

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

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal, 47*(2), 497–527. https://doi.org/10.3102/0002831209353594

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, University of Groningen]. http://hdl.handle.net/11370/408b8e4e-2be4-4312-a48a-8898995dc273.

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

Wammes, D., Slof, B., Schot, W., & Kester, L. (2021). Pupils’ prior knowledge about technological systems: Design and validation of a diagnostic tool for primary school teachers. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-021-09697-z

Wang, L. (2017). *Various spatial skills, gender differences, and transferability of spatial skills.* Visual-spatial ability in STEM education (pp. 85–105) Springer. https://doi.org/10.1007/978-3-319-44385-0_5

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

Zhu, M., Urhahne, D., & Rubie-Davies, C. M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students’ academic outcomes. *Educational Psychology, 38*(5), 648–668.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.