Metacognition and Learning
https://doi.org/10.1007/s11409-024-09383-0

Effects of availability of diagnostic and non-diagnostic cues on the accuracy of teachers' judgments of students' text comprehension

Janneke van de Pol[1] · Eleanor Rowan[1] · Eva Janssen[1] · Tamara van Gog[1]

Received: 18 January 2024 / Accepted: 8 March 2024
© The Author(s) 2024

Abstract

Accurately judging students' comprehension is a key professional competence for teachers. It is crucial for adapting instruction to students' needs and thereby promoting student learning. According to the cue-utilization framework, the accuracy of teachers' judgments depends on how predictive (or diagnostic) the information (or cues) that teachers use to make judgments is of student performance. It is, however, unclear from prior studies if merely providing access to diagnostic cues aids accuracy, or whether this only helps if non-diagnostic cues are unavailable or ignored. Therefore, we investigated, using a within-subjects experimental design, the accuracy of secondary school teachers' ($N=33$) judgments of anonymous students' text comprehension under four cue availability conditions: 1) non-diagnostic cues only; 2) diagnostic cues only; 3) a mix of diagnostic and non-diagnostic cues; and, 4) after an intervention informing them of the diagnosticity of cues, again a mix of diagnostic and non-diagnostic cues. Access to diagnostic cues enhanced teachers' judgment accuracy, while access to non-diagnostic cues hindered it. While teachers' judgment accuracy was not enhanced by the intervention (presumably because it was already relatively high), their diagnostic cue utilization increased, and non-diagnostic cue utilization decreased. In addition, teachers' calibration increased after the intervention: They knew better when their judgments were (in)accurate. Furthermore, teachers were quite aware that diagnostic cues are diagnostic, but their awareness that non-diagnostic cues (especially students' interest) are not, could be improved. These results could be useful in designing effective interventions to further foster teachers' judgment accuracy.

Keywords Teacher judgment accuracy · Cue-utilization framework · Diagnostic cue utilization · Calibration · Confidence · Judgment accuracy interventions

Introduction

Teachers are continually monitoring their students, making judgments about their learning progress (Shavelson, 1983; Shulman, 1998; Urhahne & Wijnia, 2021). These teacher judgments affect students' learning, as teachers base instructional decisions on their judgments

✉ Janneke van de Pol
j.e.vandepol@uu.nl; jvandepol@kohnstamm.uva.nl

[1] Department of Education, Utrecht University, PO Box 80.140, 3508 TC Utrecht, The Netherlands

Published online: 05 April 2024

of students' current level of understanding (Box et al., 2015; Ready & Wright, 2011; Ruiz-Primo & Furtak, 2007). If a teacher's judgment is not accurate, this can negatively affect students' academic self-concept and hinder their learning (Pielmeier et al., 2018; Urhahne, 2015). That is, if teachers underestimate students' understanding, they may provide instruction or tasks that are too easy, not furthering students' learning. In contrast, if teachers overestimate students' understanding, they may provide instruction or tasks that are too difficult, which may lead to comprehension breakdowns (Wittwer & Renkl, 2008). For teachers' instructional decisions to assist students in their learning progress, teacher judgments of student learning must be as accurate as possible (Schrader & Helmke, 2001; Thiede et al., 2019; Van de Pol et al., 2019). However, research indicates that the degree to which teachers' judgments relate to actual student learning, or the judgment accuracy, varies widely between teachers and is moderate overall (Hoge & Coladarci, 1989; Südkamp et al., 2012; Urhahne & Wijnia, 2021).

One cause of teachers' moderate judgment accuracy seems to be the kind of cues (i.e., pieces of information) that they use to inform their judgments. The cue-utilization framework of metacognitive monitoring (Brunswik, 1956; Koriat, 1997) provides a model for understanding how the accuracy of judgments depends on the *cues used* to make a judgment, and more specifically, on the *diagnosticity* of those cues. First, the model posits that judgments are inferential. That is, people use certain information, or cues, when making judgments. The cue-utilization model is specified for individuals making judgments of their own learning, in which three types of cues are distinguished: intrinsic cues (i.e., characteristics of the study material such as text length), extrinsic cues (i.e., conditions of learning such as the number of times material has been studied or encoding operations such as the level of processing, performed by the learner during learning), or mnemonic (i.e., indicators that signal to the learner to what extent material has been learned such as the ease of processing). When teachers have to make judgments of students' learning, however, the available cues differ, because mnemonic cues are internal to the learner and typically not available to teachers (unless students would report on them). Therefore, in the literature on teacher judgments of student learning, the following categories of cues are often used: task cues (comparable to intrinsic cues; characteristics of the task), student cues (student characteristics or behaviour (e.g., gender, IQ, extraversion, effort), or performance cues (e.g., a completed worksheet) (Oudman et al., 2018; Thiede et al., 2019; Urhahne & Wijnia, 2021; Van de Pol et al., 2021a).

Second, the accuracy of judgments depends on the degree to which these cues are predictive, or *diagnostic*, of the judged outcome (Dunlosky & Thiede, 2013). For example, how well a student did on a practice task that is related to the judged outcome is more diagnostic than a student's level of extraversion in class (Van de Pol et al., 2021b). The cue-utilization framework posits that using diagnostic cues to make judgments leads to more accurate judgments, while using non-diagnostic cues leads to lower judgment accuracy (Van de Pol et al., 2021a, b; Brunswik, 1956; Koriat, 1997).

Cue diagnosticity can be determined by operationalizing and measuring potential cues, such as students' overall academic ability or extraversion, resulting in a *cue value* (Van de Pol et al., 2020, 2021b; Thiede et al., 2019). By calculating the correlation between this cue value and a student's performance on a task with a measurable grade, a cue's diagnosticity for a given task can be determined (Dunlosky & Thiede, 2013; Thiede et al., 2019; Van de Pol et al., 2021b). Performance cues, such as student answers or scores on previous tests on the same and related tasks, generally correlate the highest with student performance (Fiorella & Mayer, 2015; Kostons & de Koning, 2017; Van de Pol et al., 2019, 2021b). Student cues, such as students' in-class behaviour, general academic ability, or background, have lower correlations with student performance on a specific task (Kaiser et al., 2013; Paleczek et al., 2017; Ready & Wright, 2011; Schnitzler et al., 2020; Van de Pol et al., 2019, 2021b). Task cues, such as the length or difficulty of a text, also have relatively low correlations with students' performance on a specific task (Van de Pol et al., 2021a).

There are quite some experimental studies that investigated to what extent the *height* of teachers' judgments is affected by non-diagnostic student cues, such as ADHD (Klapproth & Brink, 2024), special educational needs and immigrant background (Pit-ten Cate & Glock, 2018). Moreover, several studies have investigated to what extent teachers' judgment *accuracy* is affected by non-diagnostic student cues such as student minority status (Kaiser et al., 2017), student ethnicity (Glock et al., 2015; Pit-ten Cate et al., 2016) or student engagement (Kaiser et al., 2013).

Yet, only a few studies investigated the effects of the availability of both performance cues and students cues on teachers' judgment accuracy. The available evidence suggests that, in line with the central tenet of Koriat's cue-utilization framework, teachers' judgment accuracy is affected by the diagnosticity of the cues used. For instance, in a study by Van de Pol et al., (2021b), secondary education teachers were asked to report the cues they had used when making judgments of their own students' learning. As they knew the students, they had access to student cues, and they were provided with performance cues from a previously completed –related– task. When teachers reported using non-diagnostic student cues, their judgment accuracy was lower.

Oudman et al. (2018) experimentally manipulated primary education teachers' access to cues. Teachers made judgments of their own students' mathematics performance under different conditions, in which teachers had access to: 1) only student cues, as they were given the students' names and could therefore use their knowledge of the student; 2) only performance cues, as they were provided with students' anonymized answers on earlier practice tasks from which performance cues could be deduced; and 3) both student and performance cues, as they were provided with students' names and their answers on the practice tasks. Teachers made more accurate judgments when they only had access to performance cues than when they only had access to student cues or student and performance cues. Moreover, teachers' judgment accuracy did not differ in the latter two conditions.

These findings by Oudman et al. (2018) suggest that giving teachers access to diagnostic cues is not enough; they may have to simultaneously refrain from using non-diagnostic student cues. A later study by Oudman et al. (2023) suggests that teachers may still have relied on student cues when available, because it was hard for them to rapidly infer performance cues from students' answers. When teachers were provided either with students' names only or with students' names and their *scores* on practice tasks (rather than answers, which made it easier to infer performance cues), their judgments were more accurate when having access to both student and performance cues (Oudman et al., 2023).

In contrast to the findings by Oudman et al. (2018), Van de Pol et al. (2021a) found that secondary education teachers judging their students' reading comprehension, were most accurate when having access to student and performance cues, compared to only performance cues. Yet, data from another study using similar materials (Van de Pol et al., 2021b) suggested that again, teachers may have had difficulties with accurately inferring performance cues (i.e., teachers' judgment of the number of correct relations in a diagram of the causal relations in the text that students completed, deviated substantially from the actual number of correct relations in that diagram), and when teachers' inferences about performance cues were inaccurate, their judgments of students' performance were also less accurate. Furthermore, only when they judged the cue-values of diagnostic cues correctly, was the use of these cues related to more accurate judgments of students' learning.

One experimental study in which the values of the cues were provided using vignettes of anonymous students (i.e., teachers did not have to judge the cue values themselves) is the study by Kaiser et al. (2015). Teachers' judgments of students' mathematics grades were more accurate when provided with (the values of) diagnostic performance cues (prior oral and written mathematics achievement), than when additionally provided with non-diagnostic student cues (i.e., gender, background, IQ, family background, and academic self-concept). However, the number of cues provided in the condition with both diagnostic and non-diagnostic cues was higher (seven) than in the condition with only diagnostic cues (two). Differences in judgment accuracy may therefore not only have been due to differences in the diagnosticity of the information available, but also to the fact that in the condition with diagnostic and non-diagnostic cues, they had to process more information.

In sum, prior research suggests that to improve teachers' judgment accuracy, they need access to information from which they can easily and accurately infer performance cues, and may need to refrain from using student cues when these are also available in addition to performance cues. However, evidence currently supporting this claim is correlational (e.g., Van de Pol et al., 2021b), comes from studies that have not measured the actual diagnosticity of the cues (e.g., Oudman et al., 2018; Thiede et al., 2015; Van de Pol et al., 2021a), required teachers to interpret the cues, which is a difficult task in itself and might therefore distort judgment accuracy of students' learning (e.g., Oudman et al., 2018; Thiede et al., 2018; Van de Pol et al., 2021b), or did not provide teachers with the same amount of information between experimental conditions (Kaiser et al., 2015).

Therefore, the first aim of the present study was to experimentally test the assumed link between cue utilization and teacher judgment accuracy by manipulating teachers' access to cues in different conditions. The second aim was to investigate whether teachers are aware of the diagnosticity of the cues, and whether teachers' cue utilization, judgment accuracy, and calibration (explained in more detail below) can be promoted by a brief intervention in which they are informed about the diagnosticity of the cues and instructed to use diagnostic cues and ignore non-diagnostic cues.

As for awareness, there is not much research on teachers' cue utilization in general, and hardly any on whether teachers are aware of cue diagnosticity. The study by Zhu (2019), is an exception. Two-hundred-and-sixty primary school teachers were asked to rank several cues with regard to their diagnosticity within certain categories of cues (e.g., Abilities and attitudes [containing cues such as general intelligence, interest]; Behaviour during class [containing cues such as concentration, hand raising]; Tests [containing cues such as last test performance, grade for other subjects]; Student demographics [containing cues such as age, gender]) and to rank the categories in order of diagnosticity. Within and between categories, the teachers' ranking matched the ranking based on the actual diagnosticity (as determined by information from meta-analyses) to a great extent. However, Zhu (2019) focused on cue diagnosticity for general judgments of students' achievement, as opposed to the task specific judgments we focus on in the present study. The studies reviewed earlier suggest that teachers may not be aware of the low diagnosticity of student cues for many task-specific judgments, as teachers still report using student cues even when more diagnostic performance cues are available (Oudman et al., 2018, Van de Pol et al., 2021a).

The intervention we tested was focused on making teachers aware of cue diagnosticity, and encouraging them to use diagnostic cues and ignore non-diagnostic cues when making judgments. This approach is similar to that of Thiede et al. (2015), who found that teachers who had been trained to generate and focus on diagnostic performance cues in student-centred instruction made more accurate judgments of their students' performance than teachers who had not (but they did not measure cue diagnosticity or cue utilization).

We will not only measure effects of the intervention on judgment accuracy, but also on calibration, which is the relation between teachers' judgment accuracy and their confidence in their judgment. This has been suggested to be important for student learning (Gabriele et al., 2016) because being accurate might not be enough for teachers to actually act on those judgments; only if teachers feel confident that their judgments are accurate, they will translate their judgments into appropriate instructional decisions (Gabriele et al., 2016; Praetorius et al., 2013). Moreover, if they correctly feel that their judgments might not be accurate (thus unconfident about inaccurate judgments), they are more likely to first gather more information about the students' understanding before deciding how to proceed (e.g., by providing extra instruction). Gabriele et al. (2016) found that it was not so much teachers' judgment accuracy as it was their calibration that predicted students' mathematics achievement. Being made more aware of the diagnosticity of the cues they use through the intervention, can be expected to foster teachers' calibration as well: if teachers have more information about the diagnosticity of the cues, they might know better what cues to use and what cues to ignore and therefore whether their judgment is (in)accurate.

The present study

The present study experimentally tested the effects of cue diagnosticity on teachers' judgment accuracy, as well as the effects of an informative intervention on teachers' cue utilization, judgment accuracy, and calibration. In a within-subjects design, teachers made judgments about anonymous students' reading comprehension test performance under four different conditions, having access to: 1) non-diagnostic cues only; 2) diagnostic cues only; 3) a mix of diagnostic and non-diagnostic cues; and, 4) after an intervention informing them of the diagnosticity of cues, again a mix of diagnostic and non-diagnostic cues. We used cues generated by real students and measured against these students' test performance so we could rely on the actual diagnosticity of the cues. Furthermore, we provided teachers with actual cue values to make sure findings are not confounded by teachers' ability to accurately infer the cues. Finally, in all conditions, teachers were presented with the same number of cues so that the amount of information that had to be processed was similar.

In each condition, teachers made judgments about three students' understanding of three texts, reported the confidence in each judgment, and reported to what extent they had based each judgment on each available cue. After condition 3, teachers were asked to rank all cues from most diagnostic to least diagnostic to measure their awareness of the diagnosticity of the cues and received a "cue diagnosticity" intervention, in which they were explicitly informed about the concept of cue diagnosticity and the actual diagnosticity of each of the cues, encouraged to base their judgments on diagnostic cues, and ignore non-diagnostic cues. After that, in the fourth condition, they again saw a mix of diagnostic and non-diagnostic cues (same cues as in condition 3). During the experiment, they thought out loud. The first research question of the current study focused on the first three conditions (without intervention) and was: To what extent does the diagnosticity of available cues affect teachers' judgment accuracy? We hypothesized that teachers' judgment accuracy would be higher when diagnostic cues were available than when no diagnostic cues were available (H1.1: Condition 2 & 3 > condition 1). We also expected teachers' judgment accuracy to be higher when only diagnostic cues were available than when a mix of diagnostic and non-diagnostic cues were available (H1.2: condition 2 > condition 3).

Our second research question focused on the effects of the intervention and was: To what extent can teachers' (a) judgment accuracy, (b) calibration, and (c) use of only diagnostic cues be promoted by an intervention that informs them about cue diagnosticity and stimulates them to apply this information? We expected teachers' judgment accuracy (H2.1), calibration (H2.2) and diagnostic cue-utilization (H2.3) to all be higher in condition 4, after the intervention, compared to condition 3, and their non-diagnostic cue-utilization to be lower (H2.4). Finally, the third research question was: To what extent do teachers express an awareness of cue diagnosticity? We explored this by analysing to what extent teachers' ranking of the cues in terms of their diagnosticity was similar to the ranking of the cues based on the actual diagnosticity, and by analysing the think-aloud data they generated while making this ranking. Furthermore, we coded to what extent the teachers' idea of the diagnosticity of a cue was aligned with the actual diagnosticity as presented in the intervention.

Investigating whether and how teachers' judgment accuracy can be affected by available cues and knowledge of cue diagnosticity, not only has theoretical but also practical relevance. That is, it will not only show whether the central assumptions of Koriat's cue-utilization framework also apply to teacher judgments, but the findings from this study can also be useful for designing future interventions: If teachers' cue utilization can indeed be steered by making diagnostic cues available, making non-diagnostic cues unavailable, or informing teachers on cue-diagnosticity, these elements can be used in designing and testing interventions for teachers in practice.

Method

Participants

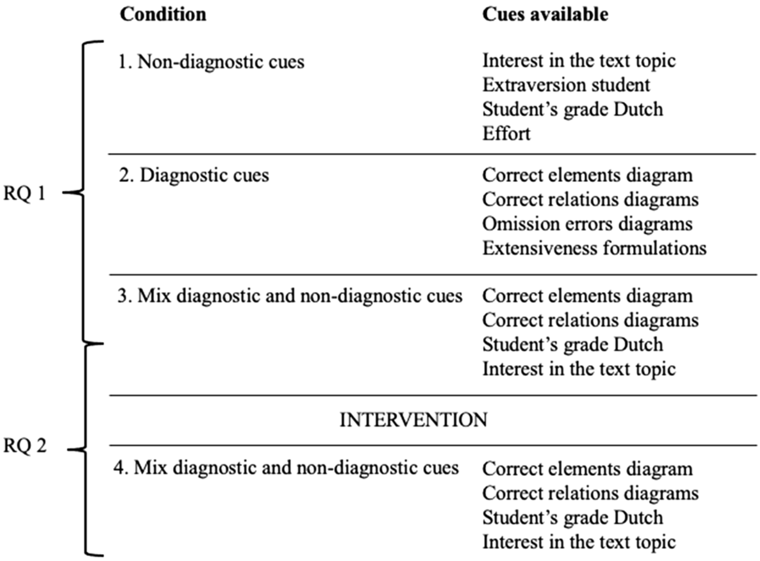
Participants were 33 teachers in Dutch secondary schools (11 male, 22 female; 90.9% born in the Netherlands). The age of the teachers ranged from 22 to 61 ($M_{\text{age}} = 39.9$, $SD = 11.1$), and their years of teaching experience ranged from one to 46 ($M_{\text{exp}} = 12.7$, $SD = 11.2$).[1] All teachers taught subjects in which text comprehension is an important skill, including Dutch (60.6%), other languages (e.g., French, English; 12.1%), History (6.1%), Geography (12.1%), Biology (3%), and Economics (3%).[2] The sample size was based on a multilevel a-priori power analysis in SPA-ML (Moerbeek & Teerenstra, 2015) using a power of 0.80 and effect size of -0.381 (Van de Pol et al., 2021b). Teachers were recruited from the network of the researchers and via social media. They could participate when they taught (1) a subject in which reading comprehension plays a role, and (2) when they taught students in grade 8–11 of pre-university education; grade 8–10 of senior general secondary education or grade 9–10 of preparatory secondary vocational education.[3]

Participants provided active informed consent and received a €20 gift certificate for their participation. They also received a report with general findings from the entire sample, as

[1] Multilevel analyses showed that there was no effect of age on teachers' judgment accuracy in any of the conditions, and no effect of years of experience on teachers' judgment accuracy in conditions 1–3. In condition 4, teachers' judgment accuracy was higher when teachers had more years of experience.

[2] One teacher did not report the subject they taught.

[3] VMBO-g or VMBO-t in Dutch.



Condition	Cues available
1. Non-diagnostic cues	Interest in the text topic Extraversion student Student's grade Dutch Effort
2. Diagnostic cues	Correct elements diagram Correct relations diagrams Omission errors diagrams Extensiveness formulations
3. Mix diagnostic and non-diagnostic cues	Correct elements diagram Correct relations diagrams Student's grade Dutch Interest in the text topic
INTERVENTION	
4. Mix diagnostic and non-diagnostic cues	Correct elements diagram Correct relations diagrams Student's grade Dutch Interest in the text topic

Fig. 1 Overview of Conditions

well as a confidential report of their own individual results. This study was approved by the Ethical Review Board of the [information blinded for review].

Design

The study had a within-subjects design,[4] with teachers making judgments about student achievement on a reading comprehension test under four conditions: having access to: 1) non-diagnostic cues only (student cues); 2) diagnostic cues only (performance cues); 3) a mix of diagnostic and non-diagnostic cues (student cues and performance cues); and, 4) after an intervention informing them of the diagnosticity of cues, again a mix of diagnostic and non-diagnostic cues (student cues and performance cues; Fig. 1). To rule out any order effects, the order of the first two conditions was randomised.

For each judgment, teachers were given a vignette about a student (Fig. 3), containing four cues in the form of information about this student (student cues) and their performance on a previous –related– learning task (performance cues). Though teachers did not know the students whose performance they had to judge, this study used data from real students. The reading comprehension test was administered to students as part of Van de Pol et al. (2021b), and student data from this study was anonymised and used for both student test scores (to calculate teachers' judgment accuracy) and the cue values presented to the teachers.

[4] We used a within-subjects design because this is a very powerful design in which participant characteristics are kept constant across conditions, because each participant takes part in each condition.

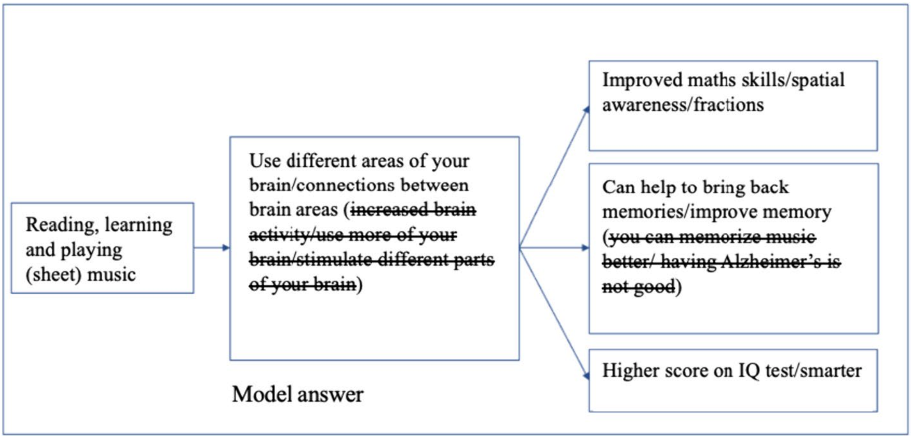

Reading, learning and playing (sheet) music
Use different areas of your brain/connections between brain areas (~~increased brain activity/use more of your brain/stimulate different parts of your brain~~)
Improved maths skills/spatial awareness/fractions
Can help to bring back memories/improve memory (~~you can memorize music better/ having Alzheimer's is not good~~)
Higher score on IQ test/smarter
Model answer

Fig. 2 Example of diagramming task shown to teachers. Note. The exemplars shown to teachers in this study included possible correct answers, and crossed-out possible incorrect answers

Materials

The experiment was programmed in Gorilla Experiment Builder and the study was conducted online with an experimenter present via Zoom.

To-be-judged text comprehension tasks

Teachers were asked to judge student performance on a reading comprehension test, which real students had completed as part of Van de Pol et al. (2021b). Teachers were informed about the test the students had taken, in which students read three short texts (stemming from Van Loon et al., 2014), each containing five clauses describing causal relations, and then completed pre-structured diagrams representing the relations in each text (Fig. 2). Students then had to write a response to three test questions (one about each text) asking for an explanation of the causal relations described in the text. For example, for the text "Music Makes You Smarter", the test question was, "There are several positive effects of learning, reading and playing music. Describe these four effects as completely as possible. In your answer, indicate the order of the four causes, using linking words like first, second, because or therefore. Use the following sentence in your answer: *learning, reading, and playing music…*". Students could receive a total of eight points for each question: four points for correctly identifying the four causal relations, and four points for having these in the correct order. For more information about the student tasks, see Appendix 1.

Performance and confidence judgments

Teachers were asked to judge students' achievement on the abovementioned reading comprehension test by answering the following question: "How many points out of eight do you think this student received on the test question for the text [text title]?". After making each performance judgment, teachers were also asked to report their confidence in each judgment they had made, answering the question, "How sure are you of your estimate of

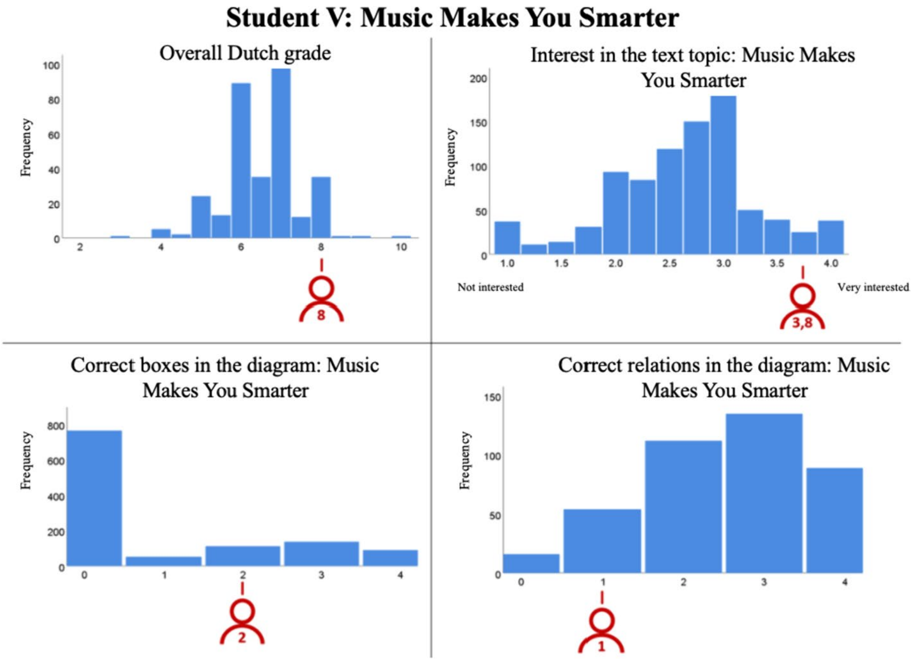


Fig. 3 Vignette for student V for text 'Music Makes You Smarter' (Condition 3, Mix Diagnostic and Non-Diagnostic Cues)

the amount of points this student scored on the test question?" on a Likert scale from *very unsure* (0), *unsure* (1), *slightly unsure* (2), *slightly sure* (3), *sure* (4), to *very sure* (5). This was needed to calculate their calibration.

Vignettes

For each judgment they made, teachers were provided with cues in the form of information about each student. The students were given pseudonyms (e.g., Student V), and the cues were presented in the form of four histograms (Fig. 3). These histograms indicated the cue value for the student (e.g., that Student V scored an 8/10 overall in the school subject Dutch) as well as the distribution of grades in the entire sample (e.g., a score of eight is relatively high in this sample, as the majority of students scored lower than an eight in Dutch). The order in which the histograms were displayed differed per judgment, to stimulate teachers to study the histograms carefully and prevent them from noticing that students were the same across conditions. The used values for the diagnostic cues and non-diagnostic cues can be found in Appendix 2, Tables 5 and 6.

Though teachers were told that they would be making judgments about the achievement of 12 different students, data from the same three students was used in all conditions, in randomised order and under different pseudonyms. This ensured that the students' actual test scores, cue values, and cue diagnosticity were stable across all conditions. Three students were chosen to ensure variation in test scores both within and between students (range = 0 – 8, $M_{\text{score}} = 3.11$, $SD = 2.85$). As cues were chosen based on diagnosticity across

Table 1 Cue values and diagnosticity

Category	Cue	Range	Reliability	Diagnosticity	Utilization
Diagnostic cues	Number of correct boxes in diagram	0 – 4	0.99	0.63	0.58
	Number of correct relations in diagram	0 – 4	0.91	0.59	0.67
	Number of omission errors in diagram	0 – 4	0.96	0.45	0.55
	Extensiveness of formulations	0 – 10	NA	0.38	0.32
Non-diagnostic cues	Dutch grade	0 – 10	NA	0.03	0.28
	Extraversion	0 – 7	0.89	-0.04	0.06
	Effort in class	0 – 4	0.76	0.08	0.31
	Interest in text topic	1 – 4	1.00	0.19	0.19

Cue diagnosticity is a Pearson correlation between cue value and test score. Cue utilization is the proportion of judgments this cue was reported as being used in. Reliability for diagnostic cues is the interrater reliability (Krippendorff's alpha), reliability for non-diagnostic cues is the internal consistency (Ω). For more information, see Van de Pol et al. (2021b)

the entire sample, cue diagnosticity was checked for each individual to ensure that diagnosticity was consistent.

Cue utilization

After making each judgment, teachers were asked to report the extent to which they had utilised each available cue, answering the question, "How much did you base your judgment on the information given?" for each cue in turn, *not at all* (0), *a little* (1), *somewhat* (2), and *a lot* (3). The presence of a cue list has been shown to not affect teachers' judgment accuracy, judgment height, or cue use (Van de Pol et al., 2021b), and so we used this self-report measure for cue utilization. There was also a text box available for teachers to fill in any other information they had used to make their judgment.[5]

Diagnostic and non-diagnostic cues

The information representing diagnostic and non-diagnostic cues was selected based on the actual diagnosticity values as determined in Van de Pol et al. (2021b), using intra-individual Pearson correlations between actual cue values and student test scores. Of the different performance cues measured in Van de Pol et al. (2021b), the four most diagnostic cues were chosen to be used in the current study, with moderate to high diagnosticity (correlation > 0.30), high reliability (Ω > 0.75) and if possible, moderate to high cue utilization (> 0.30) (Table 1). The diagnostic cues used were: The number of correct boxes in the diagram; the number of correct relations in the diagram; the number of omission errors in the diagram (e.g., empty boxes or question marks); and the extensiveness of formulations (average number of words per diagram box). Of these, the two most diagnostic cues were

[5] Only three teachers filled this out (they reported that they looked at the previous student or other students' scores on the same text and the level of the text).

used in the mixed diagnosticity condition (condition 3) and the mixed diagnosticity + intervention condition (condition 4).

As non-diagnostic cues, four cues that were of low diagnosticity (correlations between actual cue values and students' test scores < 0.30), high reliability ($\Omega > 0.75$), and if possible, moderate to high cue utilization (> 0.30[6]), were chosen to be used in the present study (cf. Van de Pol et al., 2021b). The non-diagnostic cues used were: The students' overall grade in Dutch, as provided by the student; extraversion, measured with the Big Five extraversion scale (Goldberg, 1992); effort in class, measured with the Ongoing Engagement Subdomain scale (IRRE, 1998); and students' interest in the text topic, measured using a 5-item situational interest scale (Linnenbrink-Garcia et al., 2010). More information about these instruments can be found in Van de Pol et al. (2021b). For the mixed diagnosticity condition (condition 3) and the mixed diagnosticity + intervention condition (condition 4), interest in the text topic was used, as this varied per text, and Dutch grade, as this had relatively low diagnosticity and high utilization.

Ranking task

After the mixed diagnosticity condition (condition 3), teachers were asked to rank all cues they had encountered from most predictive to least predictive. After having read a short introduction about this ranking task (Appendix 3), they received the following instruction: "How predictive are the following information sources in general for a student's test score?". The cues were listed and teachers could click and drag the cues into any order they wanted.

Intervention

In the intervention, teachers were informed about cue diagnosticity. They were informed (in written text) that according to previous research, the number of correct boxes and relations in students' diagrams were predictive of a students' test score, and that a students' Dutch grade, extraversion, effort in class, and interest in the text topic were not predictive. They were also informed that according to previous research, using predictive information and ignoring information that is not predictive, helps to make more accurate judgments. Finally, teachers were requested to only use the predictive information provided and ignore the non-predictive information, when making judgments in the subsequent block of judgments (i.e., the mixed diagnosticity + intervention condition [condition 4]).

Measures

Judgment accuracy

We analysed teachers' absolute accuracy as which is the absolute difference between the judgment value and the student's comprehension test score (Schraw, 2009; Thiede et al., 2015; Van de Pol et al., 2021b). It can range, in our study, from 0 to 8, with a score closer to zero indicating a more accurate judgment. We also report descriptive

[6] In the study where this information and the vignette data stems from (Van de Pol et al., 2021b), cue utilization is expressed as the proportion of judgments for which the particular cue is used by the teachers in that study.

statistics of bias, which is calculated by subtracting a student's test score from the value of a teacher's judgment (Schraw, 2009). Scores in our study range from –8 to + 8, with scores closer to zero indicating more accurate judgments. Negative scores indicate underestimation while positive scores indicate overestimation. For our analyses, we used absolute accuracy as the outcome measure, as bias scores are more suitable for describing general tendencies because bias scores can cancel each other out in analyses (Südkamp et al., 2008).

Calibration

Calibration captures the extent to which teachers' judgment accuracy is aligned with their confidence in this judgment (Gabriele et al., 2016). To calculate calibration, we reverse-coded absolute accuracy scores and transformed confidence scores into a 9-point scale,[7] so that a common scale was used for accuracy and confidence. To calculate calibration, we subtracted each judgment accuracy score from the corresponding judgment confidence score, and took the absolute value of this number. Scores range from 0 – 8, with a score of 0 indicating that confidence and accuracy are aligned (which can be low confidence in a less accurate judgment or high confidence in an accurate judgment) while a higher calibration score indicates that a teacher's confidence in their judgment and the accuracy of this judgment are not aligned (which can be high confidence in a less accurate judgment or low confidence in a more accurate judgment) (Gabriele et al., 2016).

Procedure

The experiment was conducted online (in Gorilla experiment builder). Teachers completed the study individually while on a Zoom call with a researcher. After providing general information, teachers were informed about the reading comprehension task that they would be estimating student performance on –making task cues available– and given information about the cues they would be using. First, teachers were shown the texts students had read and filled-in diagram exemplars, and then teachers saw the diagram task and reading comprehension test questions students had also seen. Teachers were then shown examples of the histograms they were going to see with information about each student, with an explanation of how the histogram conveyed the student score relative to the sample score. For the judgment task, teachers made judgments about student test scores in four conditions. In each condition, they made judgments about three students, and three texts per student. Think-aloud data was collected throughout.[8] Teachers could take a short break after the second condition (diagnostic cue condition) and the sessions took between 60 and 75 min.

[7] We used the following formula: (confidence score/6)*9.

[8] The recording of one teacher was absent due to technical difficulties. One other teacher had great difficulty with thinking aloud and hardly thought out loud so this transcript was not useful. One teacher did not think out loud while reading the cue diagnosticity intervention. Therefore, we used transcripts of 31 teachers for RQ2, and transcripts of 30 teachers for the part of RQ2 that focuses on teachers' responses to the intervention.

Analyses

To test hypotheses 1 and 2, a multilevel regression was performed in MPlus version 8.7 (Muthén & Muthén, 2017). Multilevel analysis is recommended in the analysis of judgment accuracy research, as teachers vary in their judgment accuracy at the individual level, as well as the level of the judgment task, giving the data a nested structure (Ready & Wright, 2011; Südkamp et al., 2012). We defined three levels in our data: teacher was the highest level (level 3), with conditions nested within teachers (level 2), as all teachers made judgments in all conditions, and then student/text as the lowest level (level 1). Student and text was considered the same level, as teachers made judgments about the same three student-text combinations. To account for the nested structure of the data, the 'complex two-level' function in Mplus was used, with the maximum likelihood estimation with robust standard errors (MLR). The teacher level was modeled using the "Complex" function, because we were not interested in the (fixed or random) effects on this level; we only wanted to account for the non-independence of observations within teachers.

Before conducting our analysis, assumptions were checked and no violations were found. We examined the data for univariate and multivariate outliers, using the median absolute deviation for univariate outliers and Mahalanobis distance for multivariate outliers (Leys et al., 2019). We ran the analyses without outliers to check for differences in results, but as the scales of our outcome variables were small, all outliers were meaningful and so we report on results with the outliers left in (Leys et al., 2019). R^2 is reported as an indication of effect size. An effect of 0.02 is considered small, 0.13 medium, and ≥0.26 large (Cohen, 2013). As a manipulation check, we tested whether teachers' cue utilization was actually affected by the manipulation in the first three conditions. Teachers' use of the non-diagnostic cues (Dutch grade and interest in the text topic) decreased significantly when having both diagnostic and non-diagnostic cues (condition 3) compared to only having non-diagnostic cues (condition 1): Dutch grade (B = -0.67, SE = 0.09, $p < 0.001$), and interest in the text topic (B = -0.55, SE = 0.09, $p < 0.001$). Teachers' use of the number of correct boxes (i.e., a diagnostic cue) decreased significantly (B = -0.19, SE = 0.09, p = 0.028) when having a mixture of diagnostic and non-diagnostic cues (condition 3) compared to having only diagnostic cues (condition 2). For the other diagnostic cue, the number of correct relations, there was no difference in cue use between these two conditions. Thus, overall, the manipulation of the availability of diagnostic and non-diagnostic cues affected teachers cue use (except for one diagnostic cue).

Teachers' awareness of cue diagnosticity

We explored teachers' awareness of the diagnosticity of cues in several ways. For the ranking task data, we calculated the correlation between the rank based on the actual diagnosticities of the cues and each teacher's rank, using the Spearman rank correlation coefficient. To get an overall idea of the degree to which the teachers' ranking matched the actual ranking, we calculate the mean (and SD) of all correlations.

Furthermore, we analysed the think-aloud data while performing the ranking task. We coded which cues were mentioned in connection with an idea of their diagnosticity. This means connecting the cue to being predictive of the students' score, using synonyms such as important, relevant, useful. Examples are: "The extensiveness of the answer is also not very predictive", "Well, extraversion is not

Table 2 Teachers' judgment accuracy, calibration, and cue utilization per condition

	Condition 1 (only non-diagnostic cues)		Condition 2 (only diagnostic cues)		Condition 3 (diagnostic and non-diagnostic cues)		Condition 4 (diagnostic and non-diagnostic cues after intervention)	
	M	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Judgment accuracy	3.18	(2.53)	1.12	(0.95)	1.35	(1.33)	1.20	(0.96)
Bias	2.71	(3.02)	−0.19	(1.45)	0.21	(1.88)	−0.26	(1.51)
Calibration	2.74	(1.85)	1.80	(1.30)	2.07	(1.38)	1.51	(1.39)
Non-diagnostic cue utilization								
Effort	2.05	(0.91)	–		–		–	
Extraversion	0.65	(0.77)	–		–		–	
Dutch	2.42	(0.74)	–		1.08	(1.01)	0.08	(0.27)
Interest	2.35	(0.88)	–		1.26	(1.18)	0.09	(0.35)
Diagnostic cue utilization								
Boxes	–		2.74	(0.67)	2.55	(0.77)	2.86	(0.48)
Relations	–		2.79	(0.60)	2.66	(0.70)	2.90	(0.40)
Extensiveness	–		1.42	(1.07)	–		–	
Omissions	–		2.06	(1.09)	–		–	

Accuracy ranges from 0–8, 0 = perfect accuracy. Calibration ranges from 0–8, 0 = perfect calibration. Cue utilization ranges from 0–3 with 0 = not at all used

so predictive, I'm going to put that one at the bottom [rank]", " most predictive I think", and "The number of correct boxes is also important". For each cue, we coded a 0 for diagnosticity not mentioned and a 1 for diagnosticity mentioned. We also indicated whether the teacher's idea of the cue diagnosticity aligned with the diagnosticity as determined based on Van de Pol et al. (2021b). For example, the utterance "Well, extraversion is not so predictive, I'm going to put that one at the bottom [rank]" would be coded as aligned (1) as extraversion is indeed not diagnostic. The utterance "Interest is the most predictive I think" would be coded as not aligned (0) as interest is not diagnostic. Twenty percent of the data was coded by two researchers. The interrater reliability was 'substantial' for whether or not the diagnosticity of the cue was mentioned (Cohen's Kappa: 0.75) and 'almost perfect' for the alignment (Cohen's Kappa: 0.91; Landis & Koch, 1977).

In addition, we analysed the think-aloud data of the part where the teachers read the information about the actual diagnosticity of the cues during the intervention. We determined, for each cue, whether teachers mentioned (dis)agreement with the information about the diagnosticity of the cues. Then, we determined whether their idea of the diagnosticity of a cue was aligned (1) or not aligned (0) with the actual diagnosticity. Twenty percent of the data was coded by two researchers. The interrater reliability was 'almost perfect' for whether the teachers mentioned (dis)agreement with the information about the cue diagnosticity (Cohen's Kappa: 0.89) and 'almost perfect' for whether the teachers' own idea of the diagnosticity of a cue was aligned with the actual diagnosticity (Cohen's Kappa: 0.91; Landis & Koch, 1977).

Table 3 Teachers' ranking per cue

		Rank							
		1	2	3	4	5	6	7	8
Diagnostic cues	Correct boxes in diagram	11	11	6	1	3	1	0	0
	Correct relations in diagram	11	12	5	3	2	0	0	0
	Omission errors in diagram	2	5	13	5	3	5	0	0
	Extensiveness of formulations diagram	0	0	0	4	4	3	16	6
Non-diagnostic cues	Interest in the text topic	6	1	2	7	7	5	3	2
	Effort	1	2	5	6	7	8	4	0
	Extraversion	0	2	0	1	1	2	3	24
	Grade Dutch	2	0	2	6	6	9	7	1

Teachers ranked the cues from most diagnostic (1st rank) to least diagnostic (8th rank)

Results

Of the variation in judgment accuracy, 18.1% occurred on the teacher level and 27.9% of the variation in calibration occurred on the teacher level. Descriptives including the judgment accuracy and bias scores can be found in Table 2. The order of conditions did not affect teachers' judgment accuracy, as a t-test was not significant ($p = 0.40$).

Relation between cue diagnosticity and teachers' judgment accuracy (RQ1)

We found confirmation for all hypotheses regarding the effect of condition on judgment accuracy, with large effect sizes. As hypothesized (H1.1), teachers' judgments were significantly more accurate in condition 2 (diagnostic cues only; $B = -2.06$, $p < 0.001$, $SE = 0.12$, $R^2 = 0.997$) and 3 (mixed cues; $B = -0.92$, $SE = 0.07$, $p < 0.001$; $R^2 = 0.995$) than in condition 1 (non-diagnostic cues only). As hypothesised (H1.2), judgment accuracy was also significantly higher in condition 2 (diagnostic cues only) than condition 3 (mixed cues) ($B = 0.23$, $SE = 0.08$, $p = 0.003$; $R^2 = 0.606$).

Effects of the cue-diagnosticity intervention (RQ2)

Opposite to what we expected (H1.2), teachers' judgment accuracy did not significantly differ before (condition 3) and after (condition 4) the cue diagnosticity intervention ($B = -0.14$, $p = 0.09$, $SE = 0.08$). However, as expected (H2.2), teachers' calibration was higher after the intervention than before ($B = -0.53$, $SE = 0.16$, $p = 0.001$; $R^2 = 0.097$ [small effect]).

As hypothesised (H2.3), teachers' self-reported use of diagnostic cues increased following the intervention. Teachers reported using students' correct boxes significantly more after the intervention than before ($B = 0.25$, $SE = 0.08$, $p = 0.003$; $R^2 = 0.057$ [small effect]). Teachers also reported using correct relations significantly more after the intervention than before ($B = 0.19$, $SE = 0.07$, $p = 0.004$; $R^2 = 0.040$ [small effect]). As hypothesised (H2.4), teachers' self-reported use of non-diagnostic cues decreased following the intervention.

Table 4 Mentioning of cues and alignment with actual diagnosticity during the ranking task and intervention

		During ranking task		During intervention	
	Cues	Mentioned	Aligned	Mentioned	Aligned
Diagnostic cues	Correct boxes	63%	100%	39%	92%
	Correct relations	69%	100%	35%	91%
	Omissions	72%	91%	0%	0%
	Extensiveness formulations	63%	25%	0%	0%
Non-diagnostic cues	Dutch grade	56%	50%	32%	20%
	Extraversion	69%	86%	26%	38%
	Effort	63%	5%	26%	13%
	Interest	66%	19%	52%	6%

Teachers reported using students' Dutch grade as a cue significantly less after the intervention than before (B=-1.00, SE=0.14, p<0.001; R^2=0.412 [medium effect]). Teachers also reported using students' interest in the text topic significantly less after the intervention than before (B=-1.16, SE=0.17, p<0.001; R^2=0.398 [medium effect]).

Teachers' awareness of the diagnosticity of cues (RQ3)

We explored the extent to which teachers were already aware of the actual diagnosticity of the cues before the intervention. The mean Spearman rank correlation between the teachers' ranking and the actual ranking was 0.55 (SD=0.36; range: -0.48 to 0.93).

The majority of the teachers ranked the number of correct boxes (n=28), the number of correct relations (n=28), and the number of omission errors (n=20) in the top three, correctly seeing those as diagnostic cues (Table 3). The extensiveness of formulations was –although it was a diagnostic cue– most often ranked (n=25) in one of the last three places (rank 6–8) and thus seen as non-diagnostic.

Students' extraversion (n=29) and their grade for Dutch (n=17) were relatively often ranked in one of the three last places and thus often –rightly– seen as non-diagnostic. Teachers were less unanimous about the diagnosticity of students' effort and interest in the text topic. Some teachers saw these as non-diagnostic, as 12 teachers ranked effort and 10 teachers ranked interest as one of the three least diagnostic cues. Yet, effort and interest were to some extent also incorrectly seen as diagnostic, as these were by respectively eight and nine teachers ranked in the top three. Interestingly, a student's interest in the text topic was ranked as the most diagnostic cue (i.e., rank 1) by six teachers.

The think-aloud data while completing the ranking task support the findings of the analysis of the ranking task. When thinking aloud, the idea of diagnosticity in relation to one of the cues was mentioned by 50–70% of the teachers (Table 4). For the cues correct boxes, correct relations, omissions, and extraversion, the teachers' idea of the diagnosticity of the cues aligned with the actual diagnosticity of the cues (86%-100%). Examples of utterances in which the diagnosticity was aligned with the actual diagnosticity are: "the number of correct relations is in my opinion the most important, because then one has really grasped the task" or "I did not attach much value to Dutch grade". For effort, interest, and extensiveness of formulations in the diagram, the teachers' idea of the diagnosticity was not so often aligned with the actual diagnosticity (5%-25%). Teachers for example stated: “I found interest very important” or “general effort plays a role”.

Furthermore, we coded teachers’ think-aloud data on the cues presented during the intervention, in terms of the degree to which teachers’ ideas of the diagnosticity of the cues as presented during the intervention was aligned with the actual diagnosticity. All cues except omission errors and extensiveness of the formulations in the diagram were mentioned when thinking aloud (Table 4). In most cases, teachers were aware of the diagnosticity of the correct boxes and relations, as their idea of the diagnosticity was aligned in respectively 92% and 91% of the cases with the actual diagnosticity (Table 4). For the non-diagnostic cues, the alignment was much lower, with an especially low degree of alignment for students’ interest in the text topic (Table 4). That is, teachers were often surprised by the fact that interest was not diagnostic. One teacher for example stated: “[reading aloud from the intervention text] ‘the information sources that were found to be non-predictive are Dutch grade, extraversion, effort, and interest’. O, that is funny. I was completely wrong.” And another teacher stated: “Ok, interest is also not predictive. I thought about that very differently.”

Discussion

In this study, we investigated the effects of the diagnosticity of available cues on teacher judgment accuracy, whether an intervention improved the use of diagnostic cues, calibration, and judgment accuracy, and teachers’ awareness of cue diagnosticity.

Judgment accuracy and cue utilization without intervention

As expected, teachers’ judgments were more accurate when having diagnostic cues available (only or in combination with non-diagnostic cues) than when having only non-diagnostic cues available. This finding aligns with Kaiser et al. (2015). When they only had diagnostic cues available, teachers were most accurate; more so than when a mix of diagnostic and non-diagnostic cues were presented. The manipulation check suggests that teachers’ use of non-diagnostic cues decreased significantly when they also had diagnostic cues available, which may explain the higher judgment accuracy in the mixed vs. non-diagnostic cues condition. However, they still used non-diagnostic cues to a limited extent in the mixed condition before the intervention (Table 2), which may explain why accuracy was highest when no non-diagnostic cues were available. Another explanation for that finding could be that they had more diagnostic information available in the condition with only diagnostic cues (four diagnostic cues) than in the mixed condition (two diagnostic cues).

Thus, having (more) diagnostic cues available seems to lead to more accurate judgments, but (additionally) having non-diagnostic cues available decreases teachers’ judgment accuracy. This is an important finding, as previous studies have only found correlational evidence for this (Van de Pol et al., 2021b), have not measured the diagnosticity of the cues (Oudman et al., 2018, Kaiser et al., 2015; Van de Pol et al., 2021a), and/or did not provide teachers with the same amount of cues in different conditions (Kaiser et al., 2015).

Effects of the intervention on teachers' judgment accuracy, cue utilization and awareness of cue diagnosticity

Opposite to what we expected, teachers' judgment accuracy did not improve after the intervention, even though their use of the diagnostic cues increased significantly, and the use of non-diagnostic cues decreased significantly. A reason for the absence of an effect on teachers' judgment accuracy may be that teachers already made reasonably accurate judgments before receiving the intervention (only a 15% deviation). Previous research with the same materials has shown lower teacher judgment accuracy: a deviation of 23% when having student cues and performance cues available (Van de Pol et al., 2021a). Yet, teachers in Van de Pol et al. (2021a) were not provided with actual cue values which was the case in the current study. In the present study, teachers did not have to interpret the cues (e.g., students' scores on practice tasks) and could use these objective scores directly for their judgments. In one previous study in which teachers also had the cue values available of a mix of diagnostic and non-diagnostic cues, teachers were also relatively accurate (deviation of 16%) in terms of absolute accuracy (Kaiser et al., 2015). Having cue values might thus have resulted in relatively accurate judgments in our study, also before the intervention.

Furthermore, the results of the analyses of teachers' awareness of the diagnosticity of the cues may explain why teachers were already relatively accurate before the intervention, given that they were quite aware. That is, they knew that the number of correct boxes and relations and the number of omissions in the students' diagrams was diagnostic of their test performance and during the intervention, they expressed agreement with the fact that these were diagnostic. Yet, the extensiveness of students' formulations in their diagrams was often incorrectly seen as non-diagnostic. However, of the diagnostic cues, this was the least diagnostic cue, so teachers were able to pick out the top three of the most diagnostic cues.

In addition, teachers knew that students' extraversion and grade for Dutch were non-diagnostic as these were often classified in the bottom three of least diagnostic cues. Yet, interest was often seen as diagnostic and even sometimes seen as *the* most diagnostic cue, whereas it was actually not diagnostic. Also, during the intervention, teachers expressed their disagreement or surprise about the fact that it was a non-diagnostic cue and there were teachers who still used this non-diagnostic cue after the intervention. It might be worthwhile to explore why teachers still use students' interest as a cue, even when they are informed that it is not diagnostic, and find ways to lower the use of this cue. The fact that our relatively short intervention affected teachers' cue utilization implies that cue utilization can be steered, both increasing utilization of diagnostic cues *and* decreasing utilization of non-diagnostic cues. Of course, teachers may have simply been following the instructions (Burger, 2009), and transfer to their professional behaviour in their own classrooms, where cues are seldomly presented so clearly to teachers, but have to be sought out, is not guaranteed. However, taken together, these findings offer initial support for the value of informing teachers about cue diagnosticity. As noted, future research could expand the intervention and examine effects on teachers' classroom practice. These findings also support the inclusion of measures such as cue diagnosticity and utilization in research on teacher judgment accuracy.

Effects of the intervention on teachers' calibration

An encouraging finding in this study was that even though teachers were not significantly more accurate in their judgments after the intervention, their confidence and accuracy were more aligned, as indicated by calibration scores. That is, after the intervention, teachers were more confident about their accurate judgments and/or less confident about their inaccurate judgments than before the intervention.

Previous findings indicate that calibration, not judgment accuracy, predicts variance in student achievement (Gabriele et al., 2016). Thus, the finding that our intervention increased calibration is important, as this suggests that explicit instruction about cue diagnosticity may have the potential to motivate teachers to enact instructional decisions based on their accurate judgments, and seek more information when they feel a judgment is less accurate. However, the effect size was small, and these conclusions require testing in studies where teachers' behaviour in their own classrooms and understanding of their own professional knowledge is examined (Brodie et al., 2018; Pit-ten Cate et al., 2016).

Limitations, implications, and future research

A strength of this study is that it provides experimental (i.e., causal) evidence for the effects of the availability and use of (non-)diagnostic cues on teachers' judgment accuracy and calibration. As the judgment process is not well understood, controlled studies in which causal effects can be isolated offer important contributions to the knowledge base (Fiedler et al., 2002). However, a limitation of this experimental approach is its lower ecological validity. In our online experiment, teachers made judgments about students they did not know personally, based on cue values that were provided rather than had to be inferred, in a process that is very different to how most teachers will make judgments in practice. So, while the finding that teachers' cue utilization can be steered is encouraging, it is important for future research to test an intervention of a similar nature in a more ecologically valid setting.

Another potential limitation of the present study is that we measured teachers' calibration by linking their confidence to their judgment accuracy at one time point, following Gabriele et al. (2016). This means we have to be cautious in interpreting these findings and cannot infer whether teachers are actually *aware* of the accuracy of their judgments, as this would require data showing that higher confidence is accompanied by higher accuracy and lower confidence by lower accuracy within the same person across two time points (Oudman et al., accepted). Yet we did find that teachers' calibration improved after the intervention, which is important as their confidence in their judgment accuracy may determine how they proceed to act, which is important for student learning.

To test our findings on the effect of the intervention on teachers' cue utilization in a more ecologically valid context, future research could consider teachers' prior knowledge of cue diagnosticity and develop interventions to address increasing the use of diagnostic cues and decreasing the use of non-diagnostic cues separately. As teachers do seem to have some awareness that performance cues are diagnostic, and indicate consciously using these to make judgments –when cue-values are given– interventions could train teachers in formative assessment approaches (Fiorella & Mayer, 2015; Fisher & Frey, 2014; Furtak et al., 2018; Thiede et al., 2018). These would help teachers to *generate* more diagnostic cues, providing them with information about students' understanding and performance on similar learning tasks. As part of an intervention, teachers could be explicitly instructed to use these diagnostic performance cues when making judgments of student learning. Second, as teachers do not always seem to be aware that student cues are less diagnostic for certain assessment tasks, interventions could inform teachers about bias, and how student cues such as interest and effort can affect the accuracy of teachers' judgments. It could be important in such interventions to also help teachers in interpreting the cues (cf. Oudman et al., 2023; Van de Pol et al., 2021b).

In the present study, we included measures that future research could also use to clarify the judgment process and the effects of any interventions. Studies on judgment accuracy and cue utilization generally take one of two approaches: interventions aiming to increase judgment accuracy through diagnostic cue utilization (Thiede et al., 2015, 2018), or more experimental studies analysing cue diagnosticity and utilization (Oudman et al., 2018; Van de Pol et al., 2021a, b). In the former, cue utilization diagnosticity is generally assumed, not measured, and in the latter, no action is taken to increase either cue utilization or judgment accuracy. The current study combines these approaches, implementing an intervention designed to increase judgment accuracy, but also measuring the utilization of cues based on real student data, the diagnosticity of which was calculated. These measures enable us to better understand the processes underlying accurate and less accurate judgments, and therefore, we hope that future studies will follow suit.

Conclusion

The cue utilization framework posits that judgment accuracy is affected by the diagnosticity of the cues people use to make a judgment of student learning. This study confirmed that relationship. This is an important finding for teacher judgment accuracy research, and further emphasises the value of the cue utilization framework for studies on teacher judgment accuracy. It was unclear, however, whether availability and use of diagnostic cues is sufficient to improve judgment accuracy, or whether use of non-diagnostic cues simultaneously needs to be suppressed. We found evidence that having diagnostic cues and non-diagnostic cues available leads to higher accuracy as having non-diagnostic cues available, but not as high as when only having diagnostic cues available. This suggests that making diagnostic cues available is insufficient if teachers do not learn to suppress the use of non-diagnostic cues. Our intervention, which was designed to teach them that, showed promising results suggesting that teachers' cue utilization can be steered, and that learning about cue diagnosticity can improve teachers' alignment between their accuracy and confidence.

Thus, cue diagnosticity seems a meaningful concept that teachers can (learn to) apply in evaluating students' comprehension. Future interventions to improve teachers' judgment accuracy should thus not only focus on making diagnostic cues available, but also help teachers to supress the use of non-diagnostic cues.

Appendix 1

Student tasks

Reading comprehension task

Students read three short expository texts: "Music makes you smarter" (167 words), "Sinking metro cars" (158 words) and "Renovating concrete buildings" (166 words). Each text contained five clauses describing causal relations (e.g., learning to play an instrument can have many benefits, because reading music and playing different notes on an instrument involves using different areas of the brain).

Diagramming task

After reading all texts, students completed pre-structured diagrams representing the relations in each text. Students were given a diagram for each text in which blank boxes were connected with arrows, indicating causal relations, and were asked to complete the boxes. Students' diagrams were assessed for the number of correct elements in the diagram (i.e., a box was filled in correctly), omissions (i.e., blank boxes or question marks), correct relations (i.e., two boxes correctly filled in, in the correct order), and the average amount of words per box (all Krippendorff's alpha > 0.91 when double coding 60 diagrams).

Test

After completing all diagrams, students answered a test question on each text, writing a short explanation of the causal relations described in the text. For example, for the text "Music Makes You Smarter, the test question was, "There are several positive effects of learning, reading and playing music. Describe these four effects. Give an answer that is as complete as possible. In your answer, indicate the order of the four causes, using linking words like first, second, because or therefore. Use the following sentence in your answer: *learning, reading and playing music…"*. See for more information Van de Pol et al. (2021b).

Appendix 2

Table 5 Values for non-diagnostic student cues in the vignettes

	Extraversion (0–7)	Interest (1–4)			Dutch (0–10)	Effort (0–7)
Student		Text 1	Text 2	Text 3		
1	4.17	2.75	1	2.5	6	3.2
2	3.67	3.5	2.75	1.75	7	3
3	5	3.75	3.75	3.75	8	3.6

Text 1: Music Makes You Smarter, Text 2: Metro Cars, Text 3: Renovation of Concrete Buildings

Table 6 Values for diagnostic student cues in the vignettes

	Omissions (0–4)			Correct relations (0–4)			Correct boxes (0–4)			Extensiveness (0–10)		
Student	Text 1	Text 2	Text 3	Text 1	Text 2	Text 3	Text 1	Text 2	Text 3	Text 1	Text 2	Text 3
1	1	4	0	1	0	1	1	0	1	3.60	1	4.20
2	0	0	1	3	4	0	3	4	2	2.80	6	5.83
3	1	4	4	1	0	0	2	0	0	5.40	1	1.40

Text 1: Music Makes You Smarter, Text 2: Metro Cars, Text 3: Renovation of Concrete Buildings

Appendix 3

Explanation ranking task teachers

When making judgments, you have used several information sources. The information sources differ in the extent to which they are predictive of the student's test score. On the next page, you will find a list with the information sources that you have encountered. We will ask you to sort these information sources from most predictive to least predictive for a student's test score. It's about the test about the three texts that you have seen earlier, in which students are asked to describe the relations in the text in the right order.

Acknowledgements We would like to thank Joëlle Canisius, Suzanne Gerritsen, and Odet Yadava for their help in the data gathering.

Funding During the realization of this research the first author was funded by a Veni grant from the Dutch Research Council awarded to the first author (grant number: 451–16-012).

Data availability Data have been stored anonymously on a secure server at Utrecht University and can be made available upon request.

Declarations

Ethics approval We complied with the APA ethical standards for treatment of human participants, informed consent, and data management. The study was approved by the Ethics committee of the Faculty of Behavioural Sciences of Utrecht University (file number 20–596).

Competing interests The authors declare that they have no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

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

Brodie, K., Marchant, J., Molefe, N., & Chimhande, T. (2018). Developing diagnostic competence through professional learning communities. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic competence of mathematics teachers: Unpacking a complex construct in teacher education and teacher practice* (pp. 151–171). Springer International Publishing. https://doi.org/10.1007/978-3-319-66327-2_8

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univ of California Press. https://doi.org/10.1525/9780520350519-017

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist, 64*(1), 1. https://doi.org/10.1037/a0010932

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

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

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom—A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88*(1), 527–561. https://doi.org/10.1006/obhd.2001.2981

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity*. Cambridge University Press. https://doi.org/10.1017/cbo9781107707085.003

Fisher, D., & Frey, N. (2014). *Checking for understanding: Formative assessment techniques for your classroom*. ASCD. https://doi.org/10.4135/9781483365633.n2

Furtak, E. M., Circi, R., & Heredia, S. C. (2018). Exploring alignment among learning progressions, teacher-designed formative assessment tasks, and student growth: Results of a four-year study. *Applied Measurement in Education, 31*(2), 143–156. https://doi.org/10.1080/08957347.2017.1408624

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49–60. https://doi.org/10.1016/j.learninstruc.2016.06.008

Glock, S., Krolak-Schwerdt, S., & Pit-ten Cate, I. M. (2015). Are school placement recommendations accurate? The effect of students' ethnicity on teachers' judgments and recognition memory. *European Journal of Psychology of Education, 30*, 169–188. https://doi.org/10.1007/s10212-014-0237-2

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26. https://doi.org/10.1037/1040-3590.4.1.26

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

IRRE. (1998). *Research Assessment Package for Schools (RAPS) manual for elementary and middle school assessments*. Institute for Research and Reform in Education. http://www.irre.org/publications/research-assessment-package-schools-raps-manual

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

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift für Erziehungswissenschaft, 18*(2), 279–302. https://doi.org/10.1007/s11618-015-0619-5

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology, 109*(6), 871–888. https://doi.org/10.1037/edu0000156

Klapproth, F., & Brink, C. (2024). Does students' ADHD diagnosis affect teachers' school-track decisions? An experimental study. *European Journal of Psychology of Education,* 1–23. https://doi.org/10.1007/s10212-024-00795-9

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349. https://doi.org/10.1037/0096-3445.126.4.349

Kostons, D., & de Koning, B. B. (2017). Does visualization affect monitoring accuracy, restudy choice, and comprehension scores of students in primary education? *Contemporary Educational Psychology, 51*, 1–10. https://doi.org/10.1016/j.cedpsych.2017.05.001

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data (pp. 159e174). *Biometrics*. https://doi.org/10.2307/2529310

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1). https://doi.org/10.5334/irsp.289

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70*(4), 647–671. https://doi.org/10.1177/0013164409355699

Moerbeek, M., & Teerenstra, S. (2015). *Power analysis of trials with multilevel data*. CRC Press. https://doi.org/10.1201/b18676-10

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

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

Oudman, S., van de Pol, J., & van Gog, T. (2023). Effects of cue availability on primary school teachers' accuracy and confidence in their judgments of students' mathematics performance. *Teaching and Teacher Education, 122*, 103982. https://doi.org/10.1016/j.tate.2022.103982

Oudman, V. S., Van de Pol, J., Janssen, E. M., & Van Gog, T. (accepted). Primary school students' awareness of their monitoring and regulation juddment accuracy. Learning and Instruction.

Paleczek, L., Seifert, S., & Gasteiger-Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students: A multilevel analysis. *Psychology in the Schools, 54*(3), 228–245. https://doi.org/10.1002/pits.21993

Pielmeier, M., Huber, S., & Seidel, T. (2018). Is teacher judgment accuracy of students' characteristics beneficial for verbal teacher-student interactions in classroom? *Teaching and Teacher Education, 76*, 255–266. https://doi.org/10.1016/j.tate.2018.01.002

Pit-ten Cate, I. M., & Glock, S. (2018). Teacher expectations concerning students with immigrant backgrounds or special educational needs. *Educational Research and Evaluation, 24*(3–5), 277–294. https://doi.org/10.1080/13803611.2018.1550839

Pit-ten Cate, I. M., Krolak-Schwerdt, S., & Glock, S. (2016). Accuracy of teachers' tracking decisions: Short-and long-term effects of accountability. *European Journal of Psychology of Education, 31*(2), 225–243. https://doi.org/10.1007/s10212-015-0259-4

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/10.1016/j.learninstruc.2016.06.008

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

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84. https://doi.org/10.1002/tea.20163

Schnitzler, K., Holzberger, D., & Seidel, T. (2020). Connecting judgment process and accuracy of student teachers: Differences in observation and student engagement cues to assess student characteristics. *Frontiers in Education, 259*. https://doi.org/10.3389/feduc.2020.602470

Schrader, F.-W., & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. *Leistungsmessungen in Schulen, 2*, 45–58. https://doi.org/10.1007/978-3-322-97583-6_4

Schraw, G. (2009). Measuring metacognitive judgments. In *Handbook of Metacognition in Education*. https://doi.org/10.1007/978-1-4419-1428-6_4928

Shavelson, R. J. (1983). Review of research on teachers' pedagogical judgments, plans, and decisions. *The Elementary School Journal, 83*(4), 392–413. https://doi.org/10.1086/461323

Shulman, L. S. (1998). Theory, practice, and the education of professionals. *The Elementary School Journal, 98*(5), 511–526. https://doi.org/10.1086/461912

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Psychologische Psychologie, 22*(34), 261–276. https://doi.org/10.1024/1010-0652.22.34.261

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

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., Oswalt, S., Snow, J. L., Sutton, J., & 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

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

Thiede, K. W., Oswalt, S., Brendefur, J. L., Carney, M. B., & Osguthorpe, R. D. (2019). Teachers' judgments of student learning of mathematics. In *The Cambridge Handbook of Cognition and Education* (pp. 678–695). https://doi.org/10.1017/9781108235631.027

Urhahne, D. (2015). Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion. *Teaching and Teacher Education, 45*, 73–82. https://doi.org/10.1016/j.tate.2014.09.006

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

Van de Pol, J., De Bruin, A. B., van Loon, M. H., & Van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology, 56*, 236–249. https://doi.org/10.1016/j.cedpsych.2019.02.001

Van de Pol, J., van Loon, M., van Gog, T., Braumann, S., & de Bruin, A. (2020). Mapping and drawing to improve students' and teachers' monitoring and regulation of students' learning from text: Current findings and future directions. *Educational Psychology Review, 32*(4), 951–977. https://doi.org/10.1007/s10648-020-09560-y

Van de Pol, J., Muilenburg, S. N., & van Gog, T. (2021a). Exploring the relations between teachers' cue-utilization, monitoring and regulation of students' text learning. *Metacognition and Learning, 16*(3), 769–799. https://doi.org/10.1007/s11409-021-09268-6

Van de Pol, J., van Gog, T., & Thiede, K. (2021b). The relationship between teachers' cue-utilization and their monitoring accuracy of students' text comprehension. *Teaching and Teacher Education, 107,* 103482. https://doi.org/10.1016/j.tate.2021.103386

Van Loon, M. H., de Bruin, A. B., van Gog, T., van Merriënboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist, 43*(1), 49–64. https://doi.org/10.1080/00461520701756420

Zhu, C. (2019). Understanding the formation and improving the accuracy of teacher Judgment. [Doctoral dissertation, Universität Passau]. https://opus4.kobv.de/opus4uni-passau/frontdoor/index/index/docId/738

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