International Journal of Science and Mathematics Education (2022) 20:459–480
https://doi.org/10.1007/s10763-021-10171-4

Dealing with Student Errors in Whole-Class Discussions of Biology Lessons at German Secondary Schools

Lena von Kotzebue[1] · Christian Förtsch[2] · Sonja Förtsch[2] · Birgit J. Neuhaus[2]

Received: 25 October 2019 / Accepted: 17 February 2021 / Published online: 8 April 2021

Abstract

Dealing with student errors is a central feature of instructional quality. Teachers' reactions to a student's error and classmates' errors can be crucial to the success of a lesson. A teacher should respond appropriately in terms of motivational and learning-related issues so that the error can become a learning opportunity for students. Currently, error situations have rarely been directly recorded and explored in empirical zstudies. This gap is the central focus of the current study in which we investigated errors in biology instruction within a cross-sectional design where biology lessons in German secondary schools were videotaped, teachers' dealings with errors analyzed, and student achievement documented with pretests and posttests. The study found that constructively dealing with student errors had a significant positive effect on student achievement at the class level. Results confirmed the relevance of teachers' appropriate dealing with student errors on learning in biology instruction.

Keywords Student errors · Dealing with errors · Error management sequences · Biology instruction · Multilevel modeling

Introduction

Theoretically, instructional quality is described as a main influence on student outcomes such as interest or achievement. The supply-usage models describe instruction

Lena von Kotzebue and Christian Förtsch shared first authorship

✉ Lena von Kotzebue
lena.vonkotzebue@sbg.ac.at

✉ Christian Förtsch
christian.foertsch@bio.lmu.de

[1] School of Education, University of Salzburg, Hellbrunnerstraße 34, 5020 Salzburg, Austria

[2] Biology Education, Faculty of Biology, LMU Munich, Winzererstr. 45/II, 80797 Munich, Germany

as a teacher's offering, which has to be used by their students resulting in learning (Kunter et al., 2013). Meta-analyses indicated positive effects of different high-quality instructional features on student outcomes (Kyriakides, Christoforou, & Charalambous, 2013; Seidel & Shavelson, 2007; e.g., constructivist-based instructional approaches, teacher questioning, feedback). Additionally, Steuer and Dresel (2015) noted that dealing with student errors in biology instruction was a key feature of instructional quality.

The term *error* has a range of meanings in the literature (Senders & Moray, 1991); one possible interpretation is that there is a discrepancy between a student's current understanding and the scientific knowledge previously presented in instruction (Kobi, 1994). Recent views on errors consider them as natural element in classroom settings and emphasize the key role for cognitive and affective outcomes. For example, the error culture theory (Oser & Spychiger, 2005; Steuer & Dresel, 2015) assumed that errors are an important way to gain knowledge and, within the constructivism view of learning, are considered natural elements of learning and reflection processes (e.g., Käfer, Kuger, Klieme, & Kunter, 2019). Therefore, it is important to identify strategies on how to constructively deal with student errors during instruction (e.g., Steuer & Dresel, 2015). However, there is little empirical research on such dealings with student errors (e.g., Tulis, 2013). Error situations were rarely recorded directly (e.g., Santagata, 2005); rather, studies indirectly explored errors with questionnaire surveys (e.g., Käfer et al., 2019; Kreutzmann, Zander, & Hannover, 2014). This is why this study systematically analyzed teachers' reactions to student errors in videotaped German biology instruction using the error management sequences (EMS) that included a student's error, the teacher's dealing with this error, and the teacher's question; finally, the effect of teachers' dealing with student errors on student achievement was explored.

Theoretical Background

From an constructivist point of view, learning is described as an active process of knowledge construction, which is based on students' misconceptions and prior knowledge (Loyens & Gijbels, 2008; Mayer, 2009). Errors can give teachers and their students information about such underlying misconceptions, students' faulty learning processes, and teachers' ineffective instruction (Hesketh, 1997; Vosniadou & Brewer, 1987). Consequently, the right to be wrong is of central importance for interactive constructivist approaches to science learning and teaching (e.g., Dole & Sinatra, 1998) as errors can be the beginning of a learning opportunity for students. However, teachers' dealing with errors is also important. Teachers should respond appropriately and constructively to student errors as it offers the possibility to create correct mental models on which students can reflect on their errors and thus avoid future errors (Chi, 2013; Kapur, 2012).

Positive Error Culture

Heinze's (2004) definition of an error in mathematics was modified to target biology as "a statement that contravenes the generally accepted statements and definitions of [biology] and generally accepted [biological] methodologies" (p. 223); it served as the foundation for this study. The error culture theory (Oser & Spychiger, 2005) suggests

that students learn what or why something is not working by making and engaging with errors, thereby gaining negative knowledge (i.e., to know what is not correct). Negative knowledge is important for the acquisition, consolidation, and preservation of positive knowledge or the sense of knowing what is correct (Gartmeier, Bauer, Gruber, & Heid, 2008). A positive error culture is characterized by the fact that errors are learning opportunities. Teachers should be willing to talk, take time, give adaptive emotional responses, empathize with students' thinking processes, and show trust by awakening their faith in being capable of learning (Heinze, 2004; Keith & Frese, 2005). This requires that teachers identify the errors before they deal with them in class and take advantage of the learning opportunity. However, Mindnich, Wuttke, and Seifried (2008) found that errors are repeatedly ignored in class.

Teacher Reactions to Student Errors

Teacher reactions to student errors can be divided into two interconnected aspects: motivational and learning. There are theoretical frameworks (e.g., behavioralism, constructivism) that describe reactions to student errors in class; however, there is not a clear practical understanding of how teachers should motivate and help students learn from their errors (Tulis, 2013).

An important motivational reaction to a student's error is that these students are not laughed at or scolded, as this tends to build a negative attitude toward errors that can lead to error-avoidance behaviors (Tulis & Riemenschneider, 2008). Realization of an error as a learning opportunity requires that students not be embarrassed by overt consideration of their error (Oser, Hascher, & Spychiger, 1999). Therefore, adequate motivational reactions are important as students should not be inhibited from taking an active part in a teaching-learning environment. Even if they are not sure about the right answer to a question, they should not be afraid to engage the opportunity and make errors. This context of encouragement can have a major impact on student performance (Hattie & Timperley, 2007).

The learning-related reactions to student errors are especially important for enabling learning opportunities, which enhance their learning processes leading to improved achievement (Steuer & Dresel, 2015). Five possible reactions have been described in the literature concerning teachers' reactions to student errors:

- A teacher redirects the same question to the student who made the error, the whole class, or other students. The strategy of redirecting the question to the same student focuses on a specific student's thinking and leads to a personalized learning opportunity (Hiebert et al., 2003). A teacher redirecting the question to another student or to the whole class broadens the engagement; however, this does not take advantage of the personalized learning opportunity for the student who made the error. Therefore, this reaction is considered problematical in view of student learning (Oser et al., 1999; Oser & Spychiger, 2005; Tulis, 2013).
- A teacher clarifies the question and then provides more information to the student or whole class, which can be used to answer the question (Mindnich, 2012). This strategy focuses on the critical aspects of the question context and supports students' thinking.

- A teacher tries to elaborate the cause of the error and the student's underlying learning process. This can be done by asking questions about how the student constructed knowledge and by probing for more detailed explanations of the answer (Chin, 2006a; Türling, 2014).
- A teacher provides assistance for the student to recognize the error. Explaining the question or defining relevant terms (Oser et al., 1999) can eliminate misconceptions through targeted instruction.
- A teacher gives content-related feedback to the student answer, which can be divided into simple and elaborate forms of feedback (Mory, 2004). Simple feedback provides information about the correctness or incorrectness of the statement; it can include the correct answer to the question. Elaborate feedback may also include the correct answer, but it helps students to directly correct the error or avoid it in the future (Hattie & Timperley, 2007).

Teacher Questions in Classroom and Wait Time

Student errors might occur due to the type of question or the way questions are formulated, and not all student errors are assumed to be negative. Chin (2006b) compared the types and purposes of questions in traditional and constructivist teaching. Teachers usually ask closed questions as an accountability function to find out what students know in traditional lessons that mostly require short answers, where facts and lower level knowledge are recalled. Students are usually not encouraged to articulate their thoughts or revise their understandings in these lessons. Chaining questions and redirecting unfinished questions to other students are practices that can be part of traditional lessons (Glas & Schlagbauer, 2008). However, different types of questions are used in constructivist teaching in which a teacher asks, encourages, and supports students to engage in real dialog (Lemke, 1990). The teacher tries to use constructivist-based instructional approaches to induce students to change their conceptual understanding (Smith, Blakeslee, & Anderson, 1993). The aim of questioning here is not only to diagnose their current state of knowledge but also to challenge and scaffold their knowledge and use more open-ended questions to promote higher order thinking (Baird & Northfield, 1992).

Studies have repeatedly shown that primarily procedural and factual questions are asked in the classroom (e.g., Bartek, 2002; Myrick & Yonge, 2002). An analysis of questions asked in biology classes revealed that students have to mainly reproduce facts and occasionally explain relationships (Förtsch, Werner, von Kotzebue, & Neuhaus, 2018b). Questions that require an overview of the biological context or an overarching concept are virtually nonexistent. Heavy reliance on teacher questioning can have problematic consequences since it does not promote the students' autonomy (Choi, Land, & Turgeon, 2008; Ismail & Alexander, 2005) and does not encourage them to think, reflect, or ask their own questions (Cooper, 2010; Sardareh, Saad, Othman, & Me, 2014). Sullivan and Liburn (2004) proposed different criteria for good tasks/questions: They should not require reproduction of facts nor limit desired responses to one correct answer; rather, in answering the question, students should learn something, and the teachers should learn something about the students' way of thinking.

The time teachers allow between asking a question and calling on a specific student to provide a response is called wait time. Enabling students to think more deeply and formulate a longer answer than a single word or short sentence requires an extended wait time. Several studies in the 1970s and 1980s investigated the role of wait time on the quality of students' answers. Rowe (1974) and Nunan (1990) pointed out that an increasing wait time leads to changes in classroom discourse—longer and more comprehensive student answers or less student failures to answer the question. Tobin (1987) reviewed studies on wait time, which ranged from kindergarten through grade 12 with numerous subjects and content areas, finding that an average wait time of between 3 and 5 s leads to a higher achievement, a decrease in student confusion, more student discourse, and less failure to respond. He also suggested that, depending on the objective of the question, a shorter wait time can be beneficial for learning, for example, drill and practice activities or recall of facts. Therefore, lengthened wait time between a teacher's question and the student's expected answer might occur more frequently in cognitively activating learning settings; it can also be associated with increased learning success.

More recent studies on wait time based on a process-oriented approach (e.g., Ingram & Elliott, 2014; Kirton, Hallam, Peffers, Robertson, & Gordon, 2007; Smith & King, 2017; Sun, 2012; Tincani & Crozier, 2008) have reported mixed results for the efficacy on student outcomes. Ingram and Elliott (2016) argued that the initiation-response-feedback/follow-up framework dominates the interactions of extended wait time in classroom interactions. Extending wait time can lead to a variety of changes in those interactions that may have both desirable and undesirable effects. They advocate a more differentiated approach to both the understanding of wait time and the desired behavior of students, as well as the interaction of these two factors.

Domain-Specific Error Management

Oser and Spychiger (2005) and Tulis (2013) emphasized domain-specific differences in both the frequency of student errors and teachers dealing with such errors. Tulis (2013) analyzed teachers' reactions to student errors in mathematics, German, and economic lessons in a video study and reported domain-specific differences. The most common reaction of mathematics teachers to student errors was to redirect the question to another student or to the whole class, whereas German teachers mostly corrected the error themselves, and economic teachers discussed the error with the whole class.

Current studies of error situations in classrooms rarely recorded the events directly but rather studied them indirectly by questionnaire surveys. Most of these studies took place in mathematics instruction (e.g., Heinze & Reiss, 2007; Kreutzmann et al., 2014; Oser & Spychiger, 2005; Santagata, 2005; Steuer & Dresel, 2015; Tulis, 2013). There are isolated studies in the context of other subjects, for example, history (Oser & Spychiger, 2005), economics (Mindnich et al., 2008; Tulis, 2013), German (Kreutzmann et al., 2014; Tulis, 2013), and English as a foreign language (e.g., Käfer et al., 2019). However, the only study on student errors in science education dealt with physics teaching (Seidel, Prenzel, & Kobarg, 2005), thereby leaving a lack of studies on student errors in biology instruction involving biology teachers.

Aim and Research Questions

The aim of this study is to analyze how biology teachers deal with student errors in whole-class discussions and if they use instructional practices of effective dealing with students' errors as described in theory. Additionally, we were interested in whether the use of these practices would lead to better student learning. The research questions of this study are:

1. How do teachers deal with student errors in terms of error management sequences in biology instruction?
2. How do effective error management sequences in biology instruction affect student achievement?

Methods

This analysis was part of the cross-sectional project ProwiN that aimed to analyze the effects of teachers' professional knowledge and different features of instructional quality on student achievement and interest. ProwiN was conducted in grades 8 and 9 of German secondary schools within the disciplines biology, chemistry, and physics (Tepner et al., 2012). This study was done in the biological part of ProwiN with an overview of the whole study published earlier (von Kotzebue et al., 2015).

Sample

The sample for this study consisted of biology teachers and their grade 9 students in secondary schools (gymnasium) in Bavaria, a federal state of Germany. All teachers and students participated voluntarily and signed consent forms. Two different samples of the project data were used to address the research questions (RQ). The sample for RQ1 was used to describe teachers' dealing with student errors in instruction; the complete video data set of the project was used, which included videotaped lessons of 43 biology teachers (60% female) and their 43 classes. All teachers were videotaped twice, except for one teacher who could be videotaped only once ($N = 85$ videos). On average, teachers were 35 years old ($SD = 8$; min = 25, max = 52). All teachers completed their studies at university to become a biology teacher for German secondary schools. Following this, all teachers completed a 2-year practical training program while teaching in secondary schools (traineeship; for a detailed description of teacher education in Germany, see Cortina & Thames, 2013). They had about 6 years of teaching experience after their traineeship ($SD = 5.5$). The lessons were on average 42.46 min long ($SD = 7.34$; min = 29.35, max = 84.34).

A subsample of the biology teachers was used for addressing RQ2 on the effects of teachers' dealing with student errors in instruction on student achievement. Student achievement data were needed for analysis. Unfortunately, we were not able to collect the pre-achievement test (pretest) and post-achievement test (posttest) data for four classes. Therefore, these classes and their teachers were not included in the subsample for RQ2. The complete pretest and posttest data for these teachers and their classes were used in this part of the study. The resulting subsample consisted of 39 biology teachers (53% female; 78 videos) and their classes. After their traineeship, these teachers had an average teaching experience of 6.1 years ($SD = 5.7$) and were 35.6 years

old ($SD = 8.3$) on average. The student subsample of these 39 classes consisted of 827 students ($M = 21.2$ students per class; 49.7% female; age in years: $M = 14.3$; $SD = .60$).

Curriculum Context

A neurobiology unit and related topics in 18 lessons explicitly determined by the Bavarian biology curriculum (Bavarian State Ministry for Education and Culture [StMUK], 2004) served as the context for this study. Comparable data from the first lesson on reflex arc was the same for all teachers; however, the second lesson was based on various teacher-selected neurobiology topics from the remaining lessons. Teachers had no further guidelines concerning how to teach these topics and did not know the focus of the study. The videotaping was guided by established standards (Seidel et al., 2005).

Data Collection

Data collection followed three steps. First, students completed a pretest on neurobiology topics. Then, biology teachers were videotaped for two Grade 9 neurobiology lessons. After the whole neurobiology topic was taught by the teachers, students completed a posttest on neurobiology and a questionnaire on motivational aspects (e.g., willingness to make an effort in their learning), which were used as control variables.

Instruments

Student Knowledge Tests. The achievement pretest and posttest focused on neurobiology and included all the topics specified in the curriculum (StMUK, 2004). The achievement tests were matched to the videotaped lessons taught by the teachers and were considered to have high curricular validity. Furthermore, the construct validity was ensured by having the tasks designed to include factual knowledge, conceptual knowledge, and scientific reasoning (Förtsch, Förtsch, von Kotzebue, & Neuhaus, 2018a), as indicated by the curriculum and the German National Education Standards (Conference of the Ministers of Education, 2005).

The pretest included 12 open-partial credit tasks and 6 multiple-choice items ($N_{pre} = 18$), whereas the posttest included all the tasks and multiple-choice items from the pretest and 4 additional open-partial credit tasks ($N_{post} = 22$). The additional tasks were only used in the posttest as they were too hard for the students to solve before they were taught the associated topic. The electronic supplementary materials (ESM) show examples of tasks that were used in the pretest and posttest. The open-partial credit items required an objective coding of the student responses; the pretest and posttest items were scored by two independent markers using a coding scheme with sample solutions to ensure objective coding. Interrater agreements or intraclass correlations (ICC) showed a high agreement and, therefore, objective and reliable coding: for pretest, $ICC_{(unjust)} = .99$, $F\ (1277,1277) = 77.92$, $p < .001$, $N = 1278$; for posttest, $ICC_{(unjust)} = .98$, $F\ (2477,2477) = 56.50$, $p < .001$, $N = 2478$ (Förtsch, Werner, von Kotzebue, & Neuhaus, 2016). Furthermore, both tests were analyzed using the Rasch partial credit model (PCM; Bond & Fox, 2007) that can be used to make both tests comparable as

long as both tests overlap at least in a few items where item difficulty results could be established for both tests on the same scale (Boone, Staver, & Yale, 2014). This analysis revealed good fit values: item reliability of 1.0 for both pretest and posttest; person reliability of .63 (pretest) and .78 (posttest); and all Infit-MNSQ/Outfit-MNSQ <1.3, which pointed to good reliability.

Student Questionnaire. A post-instruction questionnaire was used to control for possible motivational effects on student achievement (Wild, Gerber, Exeler, & Remy, 2001), which had 4-point Likert-type items with a response scale (1 = strongly disagree … 4 = strongly agree). This questionnaire contained a *willingness to make an effort* subscale ($n = 3$; $\alpha = .72$) that was assumed to affect student achievement and was used as a control variable in further analyses. Willingness to make an effort describes a student's readiness to independently deal with the lesson topic and the effort made to be successful. An example item of this scale is "When I will be examined in biology lessons (e.g., oral examination), I make a big effort to be successful."

Description of the Category System for Analyzing EMS in Biology Instruction. The recorded biology lessons were analyzed using a theoretically devised, event-based coding system that was adapted, modified, and developed from previous systems (e.g., Oser et al., 1999; Santagata, 2005; Tulis, 2013). An EMS event was defined to include three components: the teacher's question that leads to the student's error, the student's error, and the teacher's response. Therefore, an EMS always includes at least one student error and one teacher response. The teacher's question is included only if it leads to the student's error. If a student makes an incorrect statement without being prompted by a question previously asked by the teacher, only the student's error and the teacher's response are coded. One EMS continues until the error is either cleared or the teacher goes on to a new topic or ignores the error. Therefore, an EMS consists of a maximum of one teacher question, one student error, and one teacher response. This unit of analysis was similar to that in Santagata's (2005) study, but we did not take just the first response of a teacher into account; we also coded the classroom activities when the EMS occurred. The coding process was conducted separately in two steps using Videograph (Rimmele, 2012), a software program that allows users to play a video and at the same time code its contents (e.g., classroom discourse in instruction). All categories and codes were determined prior to analyzing the classroom videos based on the theoretical foundations established for this study; no emerging categories were added during or after coding. Coding followed an event-based approach. The first step used variables defining the three EMS components to identify all student errors, the teacher's questions, and responses (events); the second step used several categories for in-depth analyses of the identified EMS events (Fig. 1; see ESM for the detailed category system).

Teacher Questions. The teachers' questions that led to a student's error were analyzed based on two main aspects: unfavorable question type and question type. Unfavorable question types are to be avoided due to didactic, educational, or linguistic reasons. For example, a leading question can result in students being influenced by the way the question is posed so that they give a specific answer that the teacher expects. However, the students' answers do not reflect their real understanding of the content necessary to

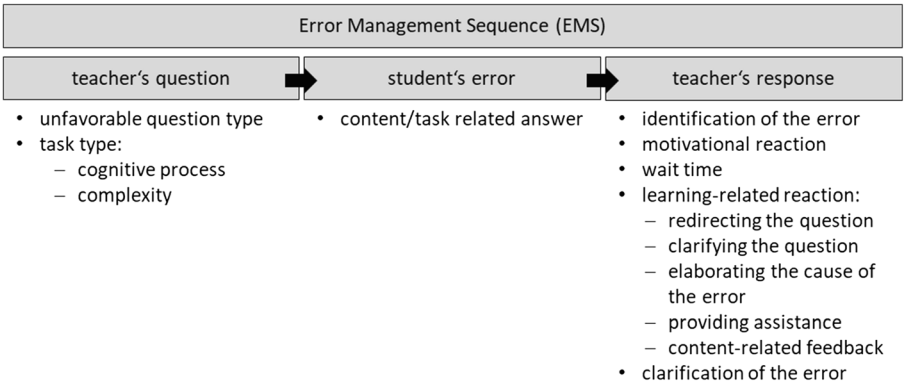


Fig. 1 Components of an EMS and analyzed categories

answer the question. One of nine possible unfavorable question types could be chosen in this category, for example, closed-ended question or yes-or-no (decision) questions (Glas & Schlagbauer, 2008). The question type is subdivided into two categories—cognitive processes and complexity—that describe the demand and difficulty of the question (Förtsch et al., 2018b; Kremer et al., 2012). The cognitive processes category focuses on students' thinking needed to complete the tasks, which were divided into four facets: reproduce, select, organize, and integrate. The complexity category was divided into three facets depending on the task's complexity; it was coded as fact, relationship, or overarching concept.

Student Errors. A student's error is generally understood to be a deviation from the expected correct answer according to the class norm (Heinze, 2004). Such errors were coded if there were false statements in their answers to teacher questions and if they were not able to answer or gave no answer.

Teacher Responses. Teacher responses were coded into nine different categories. The first category is *identification of the error*—It is important not to ignore errors, but it is difficult for students to study if the teacher addresses them too overtly in front of the class (Oser & Spychiger, 2005). The second category is *motivational reaction* of the teacher (Spychiger, Kuster, & Oser, 2006), and the third category is *wait time* until the teacher intervenes after the student has made an error (Tobin, 1987); these two responses were coded as in other studies (e.g., Tulis, 2013). The learning-related responses were coded using the remaining five categories: *redirecting of the same question* (Hiebert et al., 2003), *clarifying the question* (Mindnich, 2012), elaborating the *cause of the error* (Chin, 2006a; Türling, 2014), *providing assistance* (Oser et al., 1999), *content-related feedback* (Mory, 2004), and determined whether and by whom the error of the student was *clarified*.

Quality Criteria and Psychometrics for the Category System and Coding Categories

Typical examples for all categories in the category system were provided to ensure objective coding, which helped coders find the right coding for each teacher's question,

student's error, or teacher's response. The reliability of the coding was explored using 11% of the videos ($N = 13$ videos) that were coded by two trained independent coders; the percentage of their agreement and Cohen's kappa coefficients were calculated for each category. The percentage of agreement ranged between 90 and 99% ($\kappa = .86$ to .98) for teacher questions, between 96 and 99% ($\kappa = .93$ to .98) for student errors, and between 92 and 100% ($\kappa = .66$ to 1.00) for teacher responses. These values indicated satisfactory interrater agreement and, therefore, pointed to an objective and reliable measure (Landis & Koch, 1977).

The data from the category system for analyzing teacher responses to the students' errors and teachers' effective dealing with student errors were analyzed using Rasch measurement techniques based on the item response theory, which is used to consider both participants' abilities and test item difficulties; each item is assumed to be of a different difficulty (Boone et al., 2014). The Rasch measurement techniques are necessary for analyzing partial credit data such as coding from category systems (Wright & Masters, 1982). Partial credit data and the rating scale data were considered to be ordinal data that must be converted to a linear scale before statistical tests such as multilevel analyses were conducted. Rasch analysis results in linear person measures and item measures (referred to as the categories in this study), which are then expressed on the same linear scale. Consequently, the Wright Maps, which can be used to explain data patterns in a meaningful manner, can be constructed (Bond & Fox, 2007; Boone et al., 2014).

The Rasch model computer program Winsteps (Linacre, 2012) was used in this study. The data were analyzed using the PCM that extends the original dichotomous Rasch model (Rasch, 1960) and can be used for polytomous data. The general idea of PCM is that the problems in the items can be solved completely or partially, and the responses with partial solution can be expressed by partial credits. Additionally, responses can be ordered hierarchically so that a higher partial credit means higher quality responses. This was judged to be appropriate for all categories of the teachers' responses, which had different partial credit rating manuals (Boone et al., 2014).

Additionally, even if different categories had the same coding steps, it cannot be assumed that a higher coding in each of two categories represents similarities in difficulty (Boone et al., 2014; Linacre, 2012). An example would be that, for the category redirecting questions, the meaning of a change from a rating of 0 to 1 is not assumed to be the same as a change in the category asking a more specific question from 0 to 1. Therefore, we were able to use the single categories to refer to a teacher's ability to deal with student errors in biology instruction and to compute an overall teacher measure for further quantitative multilevel analyses. Furthermore, we could verify the reliability of the category system in this way. The psychometric analysis used the categories on the level of each single event of EMS for coding the teachers' responses in the analyzed videos. If we coded one variable in a single EMS using more than one category, the highest observed measure in the coding of the variable was used in the aggregation of teacher measures. This was done to describe teachers' ability to deal effectively with student errors. Final person measures for multilevel analyses were obtained by averaging the Rasch measures for all single EMS events for each teacher. Therefore, one value per teacher describing the ability to effectively deal with

student errors was obtained. This was possible as there was no significant difference in teachers' ability in dealing with errors in the two lessons, $t(39) = -1.53$, $p = .133$.

The evaluation of the item fit utilized established practices (Bond & Fox, 2007; Boone et al., 2014; Linacre, 2012). The results indicated that item reliability was .99 and all Item Infit-MNSQ/Outfit-MNSQ values were less than 1.3, which pointed to a productive measurement. These fit values also provided evidence that the used categories can be combined to compute a suitable person measure for teachers' ability to effectively deal with student errors (Bond & Fox, 2007; Boone et al., 2014). Person reliability was .32, which is not unexpected as a relatively small number (9) of items with a small number of categories per item measured the teachers' ability (Boone et al., 2014; Linacre, 2012).

Quantitative Multilevel Analyses

The collected data in this study were hierarchically structured and can be ordered on two different levels. Student achievement in the pretest and posttest and willingness to make an effort were measured separately for each student and, therefore, are on a student level (level 1). In contrast, teachers' ability to effectively deal with student errors was measured on the class level (level 2); this variable affects all students, which are nested in the same class. We assumed that students within one class tend to be more similar to each other than students from other classes. Consequently, student measures from one class are not completely independent from each other; multilevel analysis can take this issue into account (Hox, 2010).

We used multilevel analyses to address our research aim, which was to estimate effects of teachers' dealing with student errors in biology instruction on student achievement on the class level (level 2). We calculated a path model, which includes posttest student achievement as the outcome variable, simultaneously on the class level (level 2) and student level (level 1). Level 1 included student achievement in the pretest and willingness to make an effort as two control variables, which affect student achievement in the posttest. Therefore, we can interpret differences in student achievement in the posttest as effects of teachers' ability to effectively deal with student errors (level 2 predictor variable). All multilevel analyses were conducted using Mplus 7.3 (Muthén & Muthén, 2012) and were shown as standardized values, meaning that one unit change represents a standard deviation change in the original measure (z-standardized). Model fit was evaluated by comparative fit index (CFI) > .90; root mean square error of approximation (RMSEA) < .05; and standardized root mean square residual (SRMR) < .08, separately for the between-class and within-class covariance matrices ($SRMR_{between}$, $SRMR_{within}$; Hu & Bentler, 1998).

Results

Descriptive Findings of How Teachers Deal with Student Errors (EMS) in German Biology Instruction

Before the analysis of the descriptive data was implemented to address RQ1, two preliminary steps were taken. First, all careless mistakes, such as slips where something

that was actually known was overlooked, were excluded from the analysis. Second, only the errors made in whole-class work were included in the analysis. Most of the student errors flowed from teacher-generated questions in the whole-group phase of the biology lessons; the frequency of errors and type of teacher responses varied across the questions asked and errors made. Preliminary analysis indicated that whole-class work was by far the most frequent classroom activity where 81% of the errors occurred; thus, 752 EMS were identified. The other classroom activities were less frequent, and fewer errors were identified: 14% ($n = 137$ EMS) group work and 5% ($n = 52$ EMS) individual work.

Therefore, the remainder of the analyses focused on the 752 EMS within whole-class work where all students can perceive the errors and benefit from them. An average of 8.85 errors per lesson ($SD = 3.01$; min = 3.5, max = 18.5) was identified in whole-class work in the participating German biology classes. Coding of the student errors and teacher responses identified 585 teacher questions for the final analyses. The lower number of teacher questions compared to student errors and teacher responses is due to the coding procedure where a teacher's question was only coded if it triggered a student error. Furthermore, there were cases where student errors occurred without a teacher question having been posed. Analyses of teacher questions that led to one or more student errors indicated the following: almost half of these questions were categorized as unfavorable questions, a large majority of teacher responses were indirect or neutral and occurred quickly after the students' answers, most feedback on the errors came from the teacher but some from other students, and teachers used a limited variety of strategies to remediate student errors.

Teacher Questions. Classification of the 585 teachers' questions identified that (a) almost half were categorized as an unfavorable question and (b) all teacher questions that led to a student error were of the unfavorable question type. The most common unfavorable question type was closed-ended questions (25%) requiring one-word answers. Furthermore, a yes-or-no (decision) question (7%) and unfinished questions/supplementary questions (7%) were identified. All other unfavorable question types occurred in less than 5% of the teacher questions that led to a student error.

The teachers' questions or tasks were further classified into the categories of cognitive processes and complexity. The cognitive processes errors occurred mainly in tasks that require reproduction (85%). Errors were rarely identified in more demanding tasks (select, organize, integrate). The same pattern was found for the complexity subcategory where errors occurred in factual tasks (77%), and none were found for overarching concepts.

Teacher Responses. Teacher responses were coded for all 752 student errors (Table 1). Few teachers did not identify or ignore the error; however, most responded often indirectly. Their motivational reaction to the error was neutral in almost all responses. Their wait time was most often 1 to 3 s after the error. Teachers' clarification of the error was more evenly distributed amongst the teacher, another student, or the student making the error.

Six strategic categories of teachers' learning-related responses to student errors were identified; their responses in those categories ranged from 56 to 200 (Table 2). Most teachers redirected the same question to either another student in the class or the whole class; they rarely asked the same student who made the error. Teachers who clarified the question after the student error occurred mostly addressed the whole class. They

elaborated the cause of the error and followed up almost all of the cases by asking questions such as "Are you sure?". Teachers provided assistance mainly by giving content-related hints. Teachers more often gave simple feedback, such as just telling the student that the answer was wrong, rather than giving elaborated feedback, such as explaining the correct answer.

Teachers' Ability to Deal with Student Errors and Its Effects on Student Achievement

RQ2 was addressed with two steps of analysis. First, teachers' abilities to deal with student errors were psychometrically analyzed using Rasch analysis. Second, teachers' abilities resulting from Rasch analysis were used to analyze their effects on student achievement using two-level path modeling.

Teachers' Ability to Deal with Student Errors—Psychometric Results. The descriptive results of teacher responses involved computing a suitable person measure for teachers' ability to effectively deal with student errors. The fit values indicate that the used categories are suitable to be combined into one variable. The results of the Rasch analyses are plotted in a Wright Map (Fig. 2) that represents the teachers' ability to effectively deal with student errors in a specific EMS (person measures) on the left side of the vertical axis and the categories describing effective responses of dealing with student errors (item measures) on the right side axis (details are provided in ESM). A higher person (teacher) measure results from a teacher's response receiving higher coding on the nine categories of the coding system. Such higher measure can be interpreted as the teacher showed more often the specific behaviors for effectively dealing with student errors in a specific EMS. A teacher's response in EMS 1 was given higher scores in the nine categories compared to a teacher's response in EMS 2

Table 1 Frequency of teacher responses to student errors ($N = 752$) by category and subcategory

Category	Subcategory	%
Identification of the error	No	5.7
	Yes, without clear focus on the error	67.0
	Yes, with clear focus on the error	27.2
Motivational reaction	Devaluation/doubt	4.6
	Neutral reaction	94.8
	Praise/encouragement	0.6
Wait time	Under 1 s	17.9
	1 to 3 s	80.1
	Over 3 s	1.9
Clarification of the errors	Not clarified	7.3
	By the student who made the error	14.4
	By another student	43.5
	By the teacher	34.8

Table 2 Frequency of teacher's learning-related responses to student errors ($N = 752$) by category and subcategory

Category	Responses % / (frequency)	Subcategory (if category was coded)	% within the category
Redirecting of the same question	17.4 (131)	To the same student	9.6
		To another student	64.0
		To the whole class	26.5
Clarifying the question	16.5 (124)	To the same student	30.8
		To another student	3.3
		To the whole class	65.8
Elaborating the cause of the error	19.2 (145)	Follow up	95.1
		Ask questions on learning process	4.9
Providing assistance	12.8 (96)	Content-related	80.4
		Process-related	19.6
Simple feedback	26.6 (200)	Negative feedback	63.9
		Number of errors/ degree of correctness	18.8
		Correct answer is provided directly by the teacher	17.3
Elaborated feedback	7.5 (56)	Correct answer by teacher with explanation	81.4
		Correct answer by teacher with explanation and help	18.6

because EMS 1 is located at a higher position in the Wright Map. This indicates that the teacher's response in EMS 1 provided instruction for a higher student learning outcome. The mean score for teacher responses in all EMS was plotted with an "M" on the left side of the vertical axis. Therefore, the teacher's response in EMS 1 was coded with a higher score than that of the average teacher's response in the study sample.

A higher item measure (top of the Wright Map) means that these items (categories of the coding system) were rarely observed in the EMS of all lessons. The three categories on the bottom of the Wright Map (clarification of the error, redirecting the question, identification of the error) indicate that these items have a low item measure, which can be interpreted in the following way: Most of the student errors were *identified* and were *redirected* to another student; at the end of most EMS, the error was *clarified*. These categories were observed in nearly every lesson. Other categories of teacher responses to student errors, which are at the top of the Wright Map (i.e., elaborating the cause of the error, providing assistance, or clarifying the question), were rarely observed in an EMS in the video data.

Effects of Effective Dealing with Student Errors on Student Achievement—Results of Multilevel Analyses. The null model indicated that 19% of the variance in student achievement in the posttest can be explained by the differences between classes (ICC = .192). The two-level (teacher and student) path model revealed that (a) the teachers' ability to effectively deal with student errors predicted student achievement in the posttest on the class level and (b) the control variables, student achievement in the pretest, and willingness to make an effort predicted student achievement in the posttest

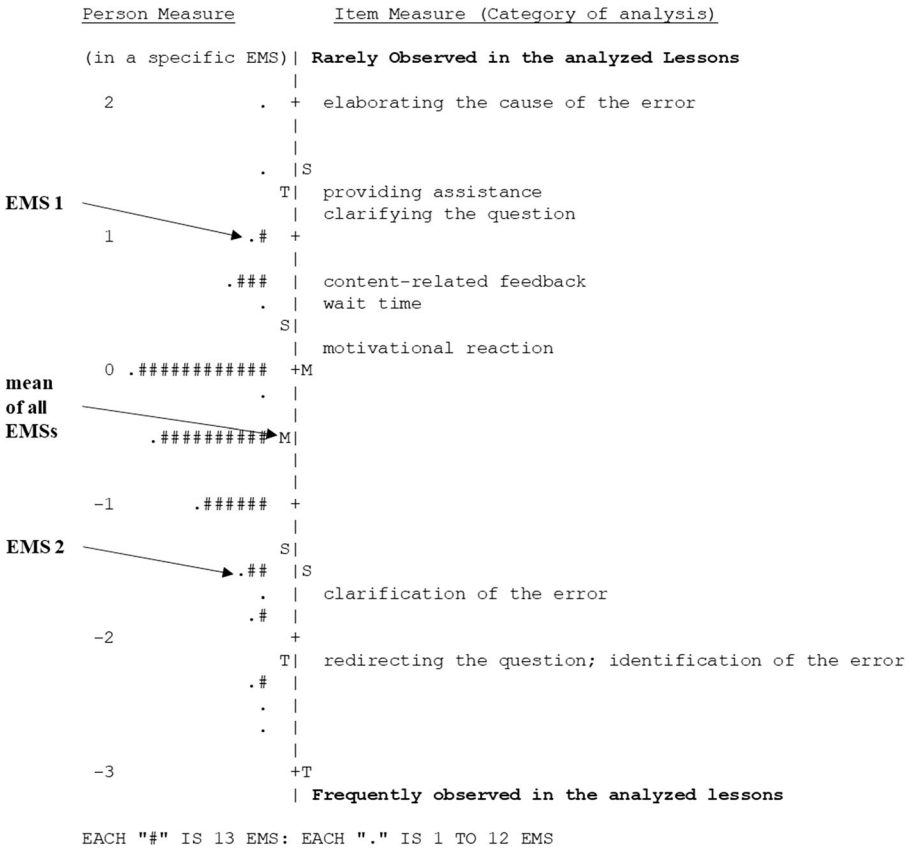


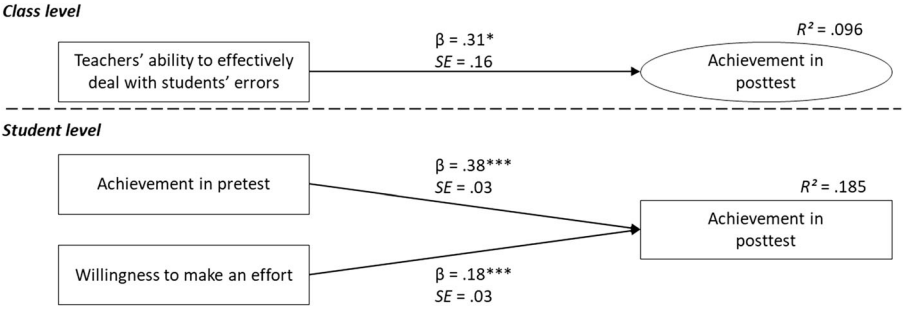
Fig. 2 Wright Map of teacher responses in dealing with student errors in a specific EMS. Person measures are plotted with the number sign "#" or a dot against item measures using the names of the categories (teacher responses)

on the student level (Fig. 3). The results indicated a significant small effect of teachers' ability to effectively deal with student errors on student achievement in the posttest ($\beta = .31$, $SE = .16$, $p = .026$) with 9.6% of the variance of student posttest achievement on the class level being explained by this predictor ($R^2 = .096$). Cohen (1988) suggested that 2% of the variance being explained ($R^2 = .02$) was a small effect and 13% ($R^2 = .13$) was a moderate effect. Therefore, this finding suggests that teacher responses, which were analyzed as effectively dealing with student errors, led to higher student understanding of the content and represented a small moderate effect in their improved performance in the achievement test. Furthermore, both predictors (control variables) showed significant moderate effects on student posttest achievement and explained 18.5% ($R^2 = .185$) of its variance (student achievement in the pretest: $\beta = .38$, $SE = .03$, $p < .001$; willingness to make an effort: $\beta = .18$, $SE = .03$, $p < .001$). Fit values of the model also showed good values: χ^2 (3) = 135.61, $p < .001$; CFI = 1.000, RMSEA < .001, $SRMR_{within} < .001$, $SRMR_{between} < .001$.

Discussion

The first research question of this study was to describe how biology teachers react in an error situation. A theory-based category system was used to examine the error management sequences, and a Wright Map was used to analyze the teachers' ability to deal with student errors. The results showed that student errors were often preceded by teacher questions that were of an unfavorable question type, which included closed-ended questions that required a one-word answer. Most student errors were identified following low-level cognitive and low-complexity questions. However, a comparison of all 2660 tasks identified in the 85 videos (Förtsch et al., 2018b) revealed no abnormalities between cognitive level and complexity. The respective proportions were of comparable size; thus, it was assumed there was no impact of the cognitive level and complexity of the tasks on the occurrence of student errors in this study (see ESM). Furthermore, different conceptual contents in the lessons did not influence the number of student errors. Teachers' learning-related reactions to student errors varied greatly, with simple feedback being the most common, followed by the questions about the cause of the error. However, the analysis of the Wright Map (Fig. 2) revealed that the most difficult reaction for the teacher was the category of elaborating the cause of the error. The reason was that in 95% of the cases, the teachers just followed up their reaction and in only 5% of the cases did they ask questions about a student's learning process, which was assumed to be a better reaction. Santagata (2005) found that mathematics teachers asked students how they came up with their answers in only about 4% of the teachers' reactions. Therefore, this category might be of interest for teacher education.

Teachers redirected the same question in more than one-sixth of the cases in this study. However, this was usually redirected to another student or to the whole class (15.7%). Accordingly, the Wright Map indicates that redirecting the question was done very often by the teachers. But such reactions occurred less often than in other studies that mostly analyzed mathematics teaching (e.g., Oser & Spychiger, 2005; Santagata, 2005; Tulis, 2013). Santagata's study of mathematics instruction in the USA indicated that redirecting the same question to other students in the class could be identified in over 30% of teachers' reactions. Individual students were not given the opportunity to



$N_{students}$ = 788; $N_{teachers}$ = 39. *p < .05. **p < .01. ***p < .001.
ICC = .192
$\chi^2(3)$ = 135.61, p < .001; *CFI* = 1.000, *RMSEA* < .001, *SRMR(within)* < .001, *SRMR(between)* < .001

Fig. 3 Two-level path model estimating the effect of teachers' ability to effectively deal with student errors on student achievement in the posttest on the class level

correct their answer as the teacher redirected the question to another student or the whole class. This maladaptive reaction was also conspicuously frequent in German economics and in Swiss history lessons (Oser & Spychiger, 2005; Tulis, 2013). Teachers in our study rarely provided assistance to the students, especially process-related help. In only a few cases was the same student asked a more specific question with more information (rephrasing), affording another chance to give the right answer. Elaborated feedback was rarely given by the teachers in our study (7.5%), especially feedback that provides additional helpful comments/wordings beyond the correct answer (1.4%). The Wright Map indicates that elaborated feedback was quite difficult for biology teachers to provide.

Teachers usually gave the students wait time of less than 3 s (88%) after their error. However, numerous studies showed that wait time between 3 and 5 s lead to a higher achievement and to more student discourse (Tobin, 1987). Almost no errors were ignored by the teachers in our study (5.7%). This can also be seen in the Wright Map since identifying the error is considered as very simple. Tulis (2013) found similar results in economics teaching (5.4%), but Mindnich et al. (2008) found that German economics teachers ignored about 40% of the errors. Santagata (2005) investigated mathematics teachers from Italy and the USA and found that they often asked the student who made the error to correct it (Italy: 41%; USA: 29%), asked another student to correct it (Italy: 12%; USA: 32%), or clarified it themselves (Italy: 32%, USA: 25%). A study of mathematics teaching in Germany showed that more than 25% of teachers' reactions were to correct the error themselves (Hiebert et al., 2003). In our study, the error was mostly clarified by a student who did not make the error (43.5%) or by the teacher (34.8%). Compared to the other studies, there are conspicuously fewer errors corrected by the student who made them (14.4%); this may mean that student errors are not used to represent real learning opportunities in these biology classes.

The second research question in this study was how teachers' dealing with student errors in biology instruction affects student achievement. We found a positive significant effect of effective dealing with student errors on student achievement. These results are in line with results for mathematics lessons (Heinze & Reiss, 2007; Steuer & Dresel, 2015). Therefore, positive dealing with student errors is not only important for acquiring mathematical routines but also for building a conceptual knowledge in biology based on the analysis of whole-class discussions. Furthermore, our finding can imply that the effects on learning for the individual that gets the feedback could be higher. This study and other studies investigated the class-level effects of other instructional quality features on student performance using multilevel modeling. The effect sizes in our study were slightly higher but within a comparable range for single instructional quality features in biology lessons: use of models, $R^2 = .17$; cognitive activation, $R^2 = .23$ or .15; tasks, $R^2 = .16$; and conceptual orientation, $R^2 = .14$ (Förtsch et al., 2016, 2018a, 2018b, 2020; Förtsch, Werner, Dorfner, von Kotzebue, & Neuhaus, 2017) or for all the three basic dimensions of instructional quality together in mathematics, $R^2 = .37$ (Kunter et al., 2013). Furthermore, there are many other instructional quality features reported in the literature that were not controlled in this study; for example, this cross-sectional study did not prescribe guidelines for the teachers. Investigating "isolated" effects would require an experimental design controlling for other instructional quality features.

Limitations

The results of this study cannot be generalized to classroom teaching in German schools due to different subjects and topics being taught, but the results showed some exemplary situations of teachers dealing with student errors in grade 9 biology classes while teaching neurobiology. Dealing with student errors is a highly complex situation that cannot be fully understood by analyzing just a few lessons. This study was not an experimental study, but the videotaped lessons were conducted by different teachers in their intact classes. The content of the lessons considered was roughly the same (1st lesson, reflex arc; 2nd lesson, teacher-selected neurobiology topics within those specified by the curriculum); however, the conceptual contents of the second lesson could vary between teachers. Teachers were free to design the content as they saw fit, and this resulted in different scientific processes and practices across the lessons. However, this instructional freedom provides important insights into typical biology lessons, which are particularly valuable for descriptive results. Future research into the effects of dealing with student errors on performance should be an experimental study where the design enables controlling these confounding factors. Furthermore, we did not specify the classroom environment that predominates in the individual classes. This could have an impact on whether students respond at all and give (incorrect) answers. However, we have demonstrated that effectively dealing with student errors has a positive impact on student performance.

Conclusions

Considering the results in this study, we can describe which aspects of dealing with student errors were difficult for the biology teachers to implement and which were implemented well. Teachers seemed to have difficulty in elaborating the cause of student error, providing assistance, giving elaborated feedback, or clarifying the question, whereas identifying and clarifying the error, redirecting the question, and giving simple feedback were readily implemented effectively in the lessons. These results illustrate that there are primarily two areas where teachers can exert positive influence on student learning: their questions and their reactions to a student's error. The analysis in this study showed that the type and complexity of the teachers' questions are probably not decisive reasons for whether or not a student makes an error because there were no unusual frequencies of student errors across the entire set of different teacher questions asked in class. But there could be a connection between a student's error and the questioning technique of the teacher. This inference is based on the result that 50% of the questions leading to student errors were of an unfavorable question type. The identification of student errors and teachers' motivational reactions were evaluated and found to be appropriate because almost all errors were identified and hardly any negative motivational reactions were observed. Deficits in teachers' learning-related reactions and their clarifications were also apparent. Teachers should make more effort to ensure that the student who made the error gets the chance to correct it. Teachers should critically examine their learning-related reactions to student errors and should profoundly inquire about a student's learning process. These aspects of teaching must be included in teacher education and professional development initiatives. However, further research must be

conducted to make empirically verified and reliable statements regarding the appropriateness of teachers' reactions to student errors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10763-021-10171-4.

Acknowledgements We are grateful to the German Federal Ministry of Education and Research (grant number 01 JH 0904) for supporting our study. We would also like to thank Larry and Shari Yore for their many constructive support.

Funding Open access funding provided by Paris Lodron University of Salzburg.

References

Baird, J. R., & Northfield, J. R. (Eds.). (1992). *Learning from the PEEL experience*. Monash University Printing.

Bartek, M. M. (2002). Paving the road to critical thinking. *Understanding Our Gifted, 14*(4), 10–12.

Bavarian State Ministry for Education and Culture [StMUK] (Ed.). (2004). *Lehrplan für das Gymnasium in Bayern* [Curriculum for the Bavarian gymnasium]. Kastner.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Chi, M. T. H. (2013). Two kinds and four sub-types of misconceived knowledge, ways to change it, and the learning outcomes. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 49–70). Routledge Press.

Chin, C. (2006a). Classroom interaction in science: Teacher questioning and feedback to students' responses. *Journal of Science Education, 28*(11), 1315–1346.

Chin, C. (2006b). Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching, 44*(6), 815–843.

Choi, I., Land, S. M., & Turgeon, A. (2008). Instructor modeling and online guidance for peer questioning during online discussion. *Journal of Educational Technology Systems, 36*(3), 255–275.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Conference of the Ministers of Education. (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Jahrgangsstufe 10)* [Education standards for the subject biology (grade 10)]. Luchterhand.

Cooper, R. (2010). *Those who can, teach*. Wadsworth Cengage Learning.

Cortina, K. S., & Thames, M. H. (2013). Teacher education in Germany. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers. Results from the COACTIV project* (pp. 49–62). Springer.

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist, 33*(2–3), 109–128.

Förtsch, C., Werner, S., von Kotzebue, L., & Neuhaus, B. (2016). Effects of biology teachers' professional knowledge and cognitive activation on students' achievement. *International Journal of Science Education, 38*(17), 2642–2666. https://doi.org/10.1080/09500693.2016.1257170

Förtsch, C., Werner, S., Dorfner, T., von Kotzebue, L., & Neuhaus, B. J. (2017). Effects of cognitive activation in biology lessons on students' situational interest and achievement. *Research in Science Education, 47*(3), 559–578. https://doi.org/10.1007/s11165-016-9517-y

Förtsch, S., Förtsch, C., von Kotzebue, L., & Neuhaus, B. J. (2018a). Effects of teachers' professional knowledge and their use of three-dimensional physical models in biology lessons on students' achievement. *Education Sciences, 8*(3), 118. https://doi.org/10.3390/educsci8030118

Förtsch, C., Werner, S., von Kotzebue, L., & Neuhaus, B. J. (2018b). Effects of high-complexity and high-cognitive-level instructional tasks in biology lessons on students' factual and conceptual knowledge. *Research in Science & Technological Education, 36*(3), 353–374. https://doi.org/10.1080/02635143.2017.1394286

Förtsch, C., Dorfner, T., Baumgartner, J., Werner, S., Kotzebue, L. von, & Neuhaus, B. J. (2020). Fostering students' conceptual knowledge in biology in the context of German National Education Standards. *Research in Science Education, 50*, 739–771. https://doi.org/10.1007/s11165-018-9709-8

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning, 1*, 87–103.

Glas, R., & Schlagbauer, J. (2008). *Pädagogik am Gymnasium – Praxiswissen für den Berufseinstieg* [Pedagogy at the gymnasium - practical knowledge for career entry]. Brigg, Friedberg, Germany: Brigg.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Heinze, A. (2004). Zum Umgang mit Fehlern im Unterrichtsgespräch der Sekundarstufe I [Dealing with errors in the classroom discourse on the lower secondary level]. *Journal für Mathematik-Didaktik, 25*, 221–245.

Heinze, A., & Reiss, K. (2007). Mistake-handling activities in the mathematics classroom: Effects of an inservice teacher training on students' performance in geometry. In H. L. Chick & J. L. Vincent (Eds.), *Proceedings of the 31st conference of the international group for the psychology of mathematics education* (pp. 9–16). Korean Society of Educational Studies in Mathematics.

Hesketh, B. (1997). Dilemmas in training for transfer and retention. *Applied Psychology, 46*(4), 317–339. https://doi.org/10.1111/j.1464-0597.1997.tb01234.x.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., . . . Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 Video Study (NCES 2003-013)*. U.S. Department of Education. Washington, DC: NCES.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453.

Ingram, J., & Elliott, V. (2014). Turn taking and 'wait time' in classroom interactions. *Journal of Pragmatics, 62*, 1–12.

Ingram, J., & Elliott, V. (2016). A critical analysis of the role of wait time in classroom interactions and the effects on student and teacher interactional behaviours. *Cambridge Journal of Education, 46*(1), 37–53.

Ismail, H., & Alexander, J. (2005). Learning within scripted and non-scripted peer-tutoring session: The Malaysian context. *Journal of Educational Research, 99*, 67–77.

Käfer, J., Kuger, S., Klieme, E., & Kunter, M. (2019). The significance of dealing with mistakes for student achievement and motivation. Results of doubly latent multilevel analyses. *European Journal of Psychology of Education., 34*(4), 731–753.

Kapur, M. (2012). Productive failure in learning math. *Cognitive Science, 38*(5), 1008–1022.

Keith, N., & Frese, M. (2005). Self-regulation in error management training: Emotion control and metacognition as mediators of performance effects. *Journal of Applied Psychology, 90*, 677–691.

Kirton, A., Hallam, S., Peffers, J., Robertson, P., & Gordon, S. (2007). Revolution, evolution or a Trojan horse? Piloting assessment for learning in some Scottish primary schools. *British Educational Research Journal, 33*(4), 605–627.

Kobi, E. (1994). Fehler [Error]. *Die neue Schulpraxis [The new school practice], 64*(2), 5–10.

Kremer, K., Fischer, H. E., Kauertz, A., Mayer, J., Sumfleth, E., & Walpuski, M. (2012). Assessment of standards-based learning outcomes in science education: Perspectives from the German project ESNaS. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible: Learning outcomes in science education* (pp. 201–218). Waxmann.

Kreutzmann, M., Zander, L., & Hannover, B. (2014). Der Umgang mit Fehlern auf Klassen- und Individualebene [Managing errors on the class and individual level]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46*(2), 101–113.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. Springer.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*, 143–152.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310.

Lemke, J. L. (1990). *Talking science: Language, learning and values*. Ablex.

Linacre, J. M. (2012). *A user's guide to Winsteps/Ministep: Rasch-model computer programs.* Retrieved from http://www.winsteps.com/a/winsteps.pdf. [12 April 2020].

Loyens, S. M. M., & Gijbels, D. (2008). Understanding the effects of constructivist learning environments: Introducing a multi-directional approach. *Instructional Science, 36*(5–6), 351–357. https://doi.org/10.1007/s11251-008-9059-4.

Mayer, R. E. (2009). Constructivism as a theory of learning versus constructivism as a prescription for instruction. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure?* (pp. 184–200). Routledge.

Mindnich, A. (2012). *Lehrerurteile in unterrichtlichen Fehlersituationen. Theoretische Rekonstruktion eines schulischen Alltagsphänomens* [Teacher judgments in teaching error situations]. *bwp@ Berufs- und Wirtschaftspädagogik – online, 22*, 1–19.

Mindnich, A., Wuttke, E., & Seifried, J. (2008). Aus Fehlern wird man klug? Eine Pilotstudie zur Typisierung von Fehlern und Fehlersituationen [Learning from errors? A pilot study on the typification of mistakes and mistake situations]. In E.-M. Lankes (Ed.), *Pädagogische Professionalität als Gegenstand empirischer Forschung [Pedagogical professionality in empirical research]* (pp. 153–163). Münster: Waxmann.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Lawrence Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Myrick, F., & Yonge, O. (2002). Preceptor questioning and student critical thinking. *Journal of Professional Learning, 18*(3), 176–181.

Nunan, D. (1990). The questions teachers ask. *JALT Journal, 12*(2), 187–202.

Oser, F., Hascher, T., & Spychiger, M. B. (1999). Lernen aus Fehlern: Zur Psychologie des negativen Wissens learning from errors: On the psychology of "negative" knowledge. In W. Althof (Ed.), *Fehlerwelten: Vom Fehlermachen und Lernen aus Fehlern [Making errors and learning from errors]* (pp. 11–41). Springer.

Oser, F., & Spychiger, M. B. (2005). *Lernen ist schmerzhaft: Zur Theorie des negativen Wissens und zur Praxis der Fehlerkultur* [Learning is painful: On the theory of negative knowledge and the practice of mistake culture]. Beltz.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Padagogiske Institut.

Rimmele, R. (2012). *Videograph* (version 4.2.1.22.X3) [computer software]. http://www.dervideograph.de/enhtmStart.html

Rowe, M. B. (1974). Relation of wait-time and rewards to the development of language, logic, and fate control. *Journal of Research in Science Teaching, 11*(4), 291–308.

Santagata, R. (2005). Practices and beliefs in mistake-handling activities: A video study of Italian and US mathematics lessons. *Teaching and Teacher Education, 21*(5), 491–508.

Sardareh, S. A., Saad, M. R. M., Othman, A. J., & Me, R. C. (2014). ESL teachers' questioning technique in an assessment for learning context: Promising or problematic? *International Education Studies, 7*(9), 161–174. https://doi.org/10.5539/ies.v7n9p161.

Seidel, T., Prenzel, M., & Kobarg, M. (Eds.). (2005). *How to run a video study: Technical report of the IPN video study*. Waxmann.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. https://doi.org/10.3102/0034654307310317.

Senders, J. W., & Moray, N. P. (1991). *Human error: Cause, prediction and reduction*. Lawrence Erlbaum Associates.

Smith, E. L., Blakeslee, T. D., & Anderson, C. W. (1993). Teaching strategies associated with conceptual change learning in science. *Journal of Research in Science Teaching, 20*, 111–126.

Smith, L., & King, J. (2017). A dynamic systems approach to wait time in the second language classroom. *System, 68*, 1–14.

Spychiger, M. B., Kuster, R., & Oser, F. (2006). Dimensionen von Fehlerkultur in der Schule und deren Messung [Dimensions of error culture in school and their measurement]. *Schweizerische Zeitschrift für Bildungswissenschaften, 28*(1), 87–110.

Steuer, G., & Dresel, M. (2015). A constructive error climate as an element of effective learning environments. *Psychological Test and Assessment Modeling, 57*, 262–275.

Sullivan, P., & Liburn, P. (2004). *Open-ended math activities: Using "good" questions to enhance learning in mathematics*. Oxford University Press.

Sun, Z. (2012). An empirical study on new teacher-student relationship and questioning strategies in ESL classroom. *English Language Teaching, 5*(7), 175–183.

Tepner, O., Borowski, A., Dollny, S., Fischer, H. E., Jüttner, M., Kirschner, S., . . . Wirth, J. (2012). Modell zur Entwicklung von Testitems zur Erfassung des Professionswissens von Lehrkräften in den Naturwissenschaften [Item development model for assessing professional knowledge of science teachers]. *Zeitschrift für Didaktik der Naturwissenschaften, 18*, 7–28.

Tincani, M., & Crozier, S. (2008). Comparing brief and extended wait-time during small group instruction for children with challenging behavior. *Journal of Behavioral Education, 17*(1), 79–92.

Tobin, K. (1987). The role of wait time in higher cognitive level learning. *Review of Educational Research, 57*(1), 69–95.

Tulis, M. (2013). Error management behavior in classrooms: Teachers' responses to student mistakes. *Teaching and Teacher Education, 33*, 56–68.

Tulis, M., & Riemenschneider, I. (2008). Self-concept, subject value and coping with failure in the math classroom—Influences on students' emotions. *International Journal of Psychology, 43*(3–4), 163 Abstracts of the XXIX International Congress of Psychology.

Türling, J. M. (2014). *Die professionelle Fehlerkompetenz von (angehenden) Lehrkräften* [Professional error competence of (prospective) teachers]. Springer.

von Kotzebue, L., Förtsch, C., Reinold, P., Werner, S., Sczudlek, M., & Neuhaus, B. J. (2015). Quantitative Videostudien zum gymnasialen Biologieunterricht in Deutschland – Aktuelle Tendenzen und Entwicklungen [Quantitative Video Studies in Biology Instruction in Secondary Schools in Germany: Current Trends and Developments]. *Zeitschrift für Didaktik der Naturwissenschaften, 21*(1), 231–237. https://doi.org/10.1007/s40573-015-0033-9

Vosniadou, S., & Brewer, W. F. (1987). Theories of knowledge restructuring in development. *Review of Educational Research, 57*(1), 51–67.

Wild, E., Gerber, J., Exeler, J., & Remy, K. (2001). *Dokumentation der Skalen- und Item-Auswahl für den Kinderfragebogen zur Lernmotivation und zum emotionalen Erleben* [Documentation of the scales and items of the questionnaire on motivation and emotional experience]. Universität Bielefeld.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis. Rasch measurement*. MESA Press.