Tutors' Assessments of a Tutee's Understanding in One-on-One Tutoring

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Grundprogramm Biologie
der Georg-August University School of Science (GAUSS)

vorgelegt von

Stephanie Herppich

aus Bayreuth

Göttingen, 2013

Betreuungsausschuss

Prof. Dr. Susanne Bögeholz, Abteilung Didaktik der Biologie, Albrecht-von-Haller-Institut für Pflanzenwissenschaften, Georg-August-Universität Göttingen

Prof. Dr. Jörg Wittwer, Arbeitsbereich Empirische Bildungsforschung mit dem Schwerpunkt Lehr-/ Lernforschung, Institut für Erziehungswissenschaft, Georg-August-Universität Göttingen (Erstmitgliedschaft: Sozialwissenschaftliche Fakultät, Zweitmitgliedschaft: Fakultät für Biologie und Psychologie)

Mitglieder der Prüfungskommission

Referentin: **Prof. Dr. Susanne Bögeholz**, Abteilung Didaktik der Biologie, Albrecht-von-Haller-Institut für Pflanzenwissenschaften, Georg-August-Universität Göttingen

Korreferent: **Prof. Dr. Jörg Wittwer**, Arbeitsbereich Empirische Bildungsforschung mit dem Schwerpunkt Lehr-/ Lernforschung, Institut für Erziehungswissenschaft, Georg-August-Universität Göttingen (Erstmitgliedschaft: Sozialwissenschaftliche Fakultät, Zweitmitgliedschaft: Fakultät für Biologie und Psychologie)

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Roland H. Grabner, Abteilung Pädagogische Psychologie, Georg-Elias-Müller Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Matthias Nückles, Abteilung Empirische Unterrichts- und Schulforschung, Institut für Erziehungswissenschaft, Albert-Ludwigs-Universität Freiburg

Prof. Dr. Hannes Rakoczy, Abteilung Biologische Entwicklungspsychologie, Georg-Elias-Müller Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Michael R. Waldmann, Abteilung Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller Institut für Psychologie, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 22.08.2013

Table of Contents

General Introduction

Good education is a general social concern in many countries (cf. Drechsel, Prenzel, & Seidel, 2009). This concern is emphasized by increasing political interest in educational outcomes, first and foremost, in school students' learning outcomes (i.e., achievement; e.g., Brookhart, 2011; Drechsel et al., 2009; Organisation for Economic Co-operation and Development (OECD), 2013). A pivotal means to good education and students' learning outcomes is instruction that is effective in terms of students' learning (Gage & Needels, 1989). Instruction is of major interest because it can be modified to improve education. To do so, it is necessary to know as to what forms of instruction are effective. Moreover, it is necessary to know as to which mechanisms make these forms of instruction effective (e.g., Gage & Needels, 1989; Lipowsky, 2009; Smith & Ragan, 2005). A recent synthesis of meta-analyses (Hattie, 2009) has documented that there, indeed, is much scientific interest in forms of instruction that optimally foster learning. Hattie (2009) synthesized meta-analyses that examine influences on the learning outcomes of school-aged students. In Hattie's synthesis (2009), two chapters that summarize 365 meta-analyses are exclusively devoted to influences from different "teaching approaches" (e.g., p. 161), that is, from forms of instruction.

This doctoral thesis contributes to the research on effective instruction and the mechanisms at work within these forms of instruction. It examines one-on-one human tutoring, which has been found to be a very effective form of instruction (for an overview, cf. Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Graesser, D'Mello, & Cade, 2011). More precisely, in the doctoral thesis the assessments and the assessment accuracy of tutors in one-on-one tutoring are studied. Assessments and assessment accuracy of instructors are deemed central mechanisms with regard to the effectiveness of instruction (e.g., Black & Wiliam, 1998; Furtak et al., 2008; Klug, Bruder, Kelava, Spiel, & Schmitz, 2013; Van de Pol, Volman, & Beishuizen, 2010; Vogt & Rogalla, 2009; Wiliam, Lee, Harrison, & Black, 2004).

Tutoring – An Effective Form of Instruction

In this doctoral thesis, *tutoring* is defined as a form of instruction where a human instructor (i.e., a *tutor*) teaches *one* learner (i.e., a *tutee*) on some subject matter. Moreover, the tutor is defined to be *more knowledgeable* about the subject matter than the tutee. The focus of interest is on the tutee's learning. This form of tutoring has to be distinguished from

computer tutoring where the tutee is taught by an artificial tutoring system (e.g., VanLehn, 2011). It also has to be distinguished from (small) group tutoring where one tutor teaches several learners (e.g., Schmidt, & Moust, 1995). In this doctoral thesis, tutoring is, furthermore, distinguished from peer tutoring. In peer tutoring, a learner teaches another learner and the focus of interest often is on the learning of both peers. However, there are no sharp boundaries, neither between the concepts of tutoring and peer tutoring (cf. the typology by Topping, 1996, 2005) nor between the lines of research that examine these forms of instruction (cf. Chi et al., 2001; P.A. Cohen, Kulik, & Kulik, 1982).

In Germany, tutoring is mostly conceptualized and implemented as private tutoring (Haag, 2010). Thus, it is conceived of as instruction that is provided outside of regular education at school. Nevertheless, private tutoring is not the only implementation of tutoring (cf. e.g., P. A. Cohen et al., 1982, for tutoring as a substitute to classroom instruction; Lonigan & Whitehurst, 1998, for parents tutoring their pre-school children).

Evidence for the Effectiveness of Tutoring

Research has found tutoring to be a very effective form of instruction (Bloom, 1984; Elbaum, Vaughn, Hughes, & Moody, 2000; Ritter, Barnett, Denny, & Albin, 2009; VanLehn, 2011). To test for its effectiveness, tutoring has usually been compared with classroom instruction or with other forms of instruction (P. A. Cohen et al., 1982; VanLehn, 2011). Effect sizes reported for the effectiveness of tutoring in terms of a tutee's learning vary between approximately 0.4 and 2.0 standard deviations (Chi et al., 2001; Graesser et al., 2011). J. Cohen (1988) interpreted effect sizes of 0.2 standard deviations as small, effect sizes of 0.5 standard deviations as medium, and effect sizes of 0.8 standard deviations as large. According to this interpretation, tutoring yields medium to large effects on a tutee's learning.

Tutoring by tutors with particular training in teaching or teaching experience (i.e., *experienced tutors*) is sometimes reported with effect sizes between 0.8 and 2.0 standard deviations to be more effective than tutoring by tutors without particular training in teaching or without teaching experience (i.e., *inexperienced tutors*). For inexperienced tutors average effect sizes of 0.4 standard deviations have been reported (for an overview, see Graesser et al., 2011). Inexperienced tutors can be, for example, parents, older peers, volunteers from the community, or university students (e.g., Chi et al., 2001; Graesser et al., 2011; Ritter et al., 2009; VanLehn, 2011). Experienced tutors can be, for example, classroom teachers, graduate students, university teachers, or professional tutors (e.g., Chi,

Roy, & Hausmann, 2008; Lehman, D'Mello, Cade, & Person, 2012; Putnam, 1987). However, there are only few studies that employ experienced tutors. Moreover, definitions of what constitutes an experienced tutor vary between studies. Thus, evidence on the effectiveness of different tutors is still inconclusive and deserves further study (Graesser et al., 2011; Lehman et al., 2012; VanLehn, 2011).

Approaches to the Study of Tutoring

Knowing that tutoring is effective leaves the question unanswered as to why it is effective. Graesser et al. (2011) have identified three approaches of research to answering this question. The first approach relates general characteristics of the subject matter, the tutee, the tutor and the structure of the tutoring session to the learning of a tutee (e.g., P. A. Cohen et al., 1982; Elbaum et al., 2000; Ritter et al., 2009; Wasik, & Slavin, 1993). Studies that adopt this approach, for example, examine the effectiveness of tutoring by tutors with varying levels of instructional training or instructional experience (e.g., Elbaum et al., 2000; Wasik, & Slavin, 1993). The second approach conducts in-depth analyses of the structures and processes of tutoring sessions (e.g., Cade, Copeland, Person, & D'Mello, 2008; Chi et al. 2008; Chi et al., 2001; Graesser, Person, & Magliano, 1995; McArthur, Stasz, & Zmuidzinas, 1990; Putnam, 1987; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). The third approach manipulates activities of tutors and tests the effect of the manipulation on a tutee's learning (e.g., Blom-Hoffman, O'Neil-Pirozzi, Volpe, Cutting, & Bissinger, 2006; Chi et al., 2001; Whitehurst et al., 1988).

Results of the second approach are particularly informative because one learns about the mechanisms that work within tutoring and that assumedly make tutoring effective. Knowledge about these mechanisms is essential when we want to even improve tutoring. This knowledge is also essential when we study tutoring to draw conclusions for other forms of instruction with the aim of making them more effective (e.g., classroom instruction; cf., Haag, 2010).

Important Structures and Processes in a Tutoring Session

The Tutoring Dialogue Frame. Studies in line with the second approach have deemed the structure of the tutoring dialogues within one tutoring session important for tutoring effectiveness. More precisely, these studies have observed that certain communication processes are embedded into this structure. These communication processes, in turn, are

seen as an important key to the effectiveness of tutoring (Cade et al., 2008; Chi et al., 2001; Graesser et al., 1995; VanLehn et al., 2007).

Assumedly due to the one-on-one situation (Snow & Swanson, 1992), the dialogues between a tutor and a tutee are more expanded than are dialogues in classroom situations. In class, the teacher usually, first, asks an *initiating* question about a given topic or problem, second, the student *responds*, and third, the tutor *evaluates* the correctness of the response giving short feedback (IRE; Mehan, 1979). In tutoring, there is a fourth and a fifth step (*5-step dialogue frame*; Graesser et al., 1995, p. 504). Fourth, tutor and tutee exchange several contributions to improve the response the tutee gave in the second step. The exchanges can be very interactive in nature. That is, the tutor gives feedback on a tutee's contribution or scaffolds the tutee to elicit new constructive responses from a tutee (Chi, 2009; Hmelo-Silver & Barrows, 2006). Scaffolding refers to a tutor's contributions such as questions or hints that are aimed at helping the tutee to proceed in a line of reasoning or in a task that the tutee would not be able to accomplish alone (Chi et al., 2001; Van de Pol et al., 2010). The exchanges can also be less interactive in nature. This is the case when the tutor predominantly provides instructional explanations (Chi, 2009). Fifth, the tutor assesses whether the tutee has understood the response. Usually the tutor takes the responsibility for the progression through the dialogue (Chi et al., 2001; Graesser et al., 1995).

Opportunities for the Tutor to Assess a Tutee's Understanding. The extended dialogue about a single topic or problem provides tutors with several opportunities to assess a tutee's understanding. These opportunities can occur during the third step, the fifth step and, particularly, during the fourth step of the tutoring dialogue frame. The more interactive a tutor organizes the exchanges during the fourth step the more opportunities to assess a tutee's understanding arise in the course of the exchanges. This is because a tutor can learn what a tutee does and does not know from the tutee's responses to a tutor's interactive contributions, for example, from the answer to a tutor's question (Chi, 2009; Hmelo-Silver & Barrows, 2006).

Assessing a tutee's understanding should enable the tutor to adapt instruction to the tutee's current understanding on a moment-to-moment basis (Snow & Swanson, 1992; Van de Pol et al., 2010). The more thoroughly a tutor assesses a tutee's understanding the better this tutor should be able to adapt instruction. The effectiveness of tutoring is partly ascribed to its adaptiveness on a moment-to-moment basis (Snow & Swanson, 1992;

Graesser et al., 2011; Lehmann et al., 2012; see Chi & Roy, 2010 for a deviating conceptualization and view of adaptation).

Moreover, due to the expanded dialogue, during tutoring the tutor gets the opportunity to gather a multitude of information about the tutee's understanding. After tutoring the tutor could aggregate this information to comprehensively assess the tutee's understanding (e.g., Black, 1993; Black & Wiliam, 2009; Perie, Marion, & Gong, 2009; cf. the section *Forms of Assessment*). This assessment might, in turn, serve the tutor to select materials for a subsequent tutoring session that are also adapted to the tutee's understanding (Chi, Siler, & Jeong, 2004; Kalyuga, 2007; Shepard, 2001; cf. also Perie et al., 2009). A tutor's assessments are pivotal to these considerations of the mechanisms that make tutoring effective. Nevertheless, research has not yet intensively studied these assessments. The next section reviews what is known about assessments of tutors and of other instructors.

Tutors' Assessments

An *assessment* is defined as a judgment about another person (cf. Schrader, 2010). In the context of instruction, it is generally deemed crucial that an instructor can *accurately assess* a learner. This is because accurate assessments of a learner's prerequisites for learning, a learner's learning processes, and a learner's learning outcomes are regarded as being mandatory to adapt instruction to the individual learner (Klug et al., 2013; Schrader, 2010; Van de Pol et al., 2010; Vogt & Rogalla, 2009; Wittwer & Renkl, 2008; see Weinert & Schrader, 1986 for a deviating view on the necessity of accurate assessments). Instructional measures have to be adapted to the needs and prerequisites of the individual learner to optimally foster learning processes (Corno & Snow, 1986; Vogt & Rogalla, 2009; Van de Pol et al., 2010; Wittwer & Renkl, 2008; see also Schrader, 2010).

The significance that research attaches to an instructor's assessments is reflected in the prominent role assessment skills play in several models of teachers' knowledge and skills (e.g., Baumert & Kunter, 2006; Borko, Mayfield, Marion, Flexer, & Cumbro, 1997; Borko & Putnam, 1996; Magnusson, Krajcik, & Borko, 1999; Weinert, Helmke, & Schrader, 1992). Moreover, both knowledge about assessment and assessment skills have recently become part of the German standards for teacher education (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004). This demand to train teacher students and preservice teachers in accurate

assessments of students emphasizes the practical need for accurate assessments in instruction.

Forms of Assessment

Research on assessment distinguishes two distinctive but potentially interacting forms of assessment that are relevant to instruction (e.g., Birenbaum et al., 2006; Black, 1993; Shavelson et al., 2008). First, *formative assessment* takes place mainly during an instructional unit (Black & Wiliam, 2009; Ruiz-Primo & Furtak, 2007). In tutoring, a tutor's assessments of the tutee's current understanding on a moment-to-moment basis can be described as formative assessment (e.g., Black & Wiliam, 2009; Ruiz-Primo & Furtak, 2007). Formative assessment is used to infer information about a learner with the aim of adapting instruction to this learner. Thus, it is meant to foster learning (e.g., Bennett, 2011; Black, 1993; Black & Willliam, 1998, 2009).

Second, *summative assessment* is usually carried out at the end of an instructional unit to document a learner's learning outcomes (e.g., Bennett, 2011; Shavelson et al., 2008). A tutor's comprehensive assessment of a tutee's understanding after tutoring can, thus, be characterized as summative assessment (e.g., Shepard, 2001).

As described in the section *Important Structures and Processes in a Tutoring Session*, research assumes that both forms of assessment are interrelated with each other. Instructors can, for instance, use information gathered during formative assessment to receive summative assessments of a learner's learning outcomes (e.g., Birenbaum et al., 2006; Black, 1993). It should be noted that definitions of formative and summative assessment still vary across research. In research on classroom instruction, for example, summative assessment is sometimes equated with external standardized assessments (e.g., Perie et al., 2009; but cf. Shepard, 2001; for other differentiations see Black, 1993; Shavelson et al., 2008). It is unlikely, however, that tutors in practical applications of tutoring such as private tutoring (Haag, 2010) employ, for example, standardized assessments. Tutors' assessments can therefore be studied best within the framework of the definitions outlined above.

Research on Tutors' Assessments

Tutors' Assessments in General. Research on tutoring is mostly interested in processes that explain the effectiveness of tutoring. In this vein, a few studies have examined a tutor's formative assessments. Intriguingly, these studies have found that tutors seldom

deliberately assess a tutee's understanding. Nor do the tutors usually adapt their teaching contents to a tutee's particular needs. Instead, the selection of contents and the progression through the contents of a tutoring session were largely determined by the tutor's internal curriculum script of what a tutee was to learn (Chi et al., 2004, 2008; Graesser et al., 1995; McArthur et al., 1990; Putnam et al., 1987; see also Cromley & Azevedo, 2005).

These results are critical given that Bloom (1984) has documented the particular effectiveness of tutoring with embedded formative tests as compared with traditional classroom instruction. As an exception to the practice of only studying process measures, Chi et al. (2004) also measured a tutor's assessment accuracy after half a tutoring session and after the tutoring session. Thus, they measured assessment accuracy from a more summative perspective. Chi et al. (2004) found that tutors generally overestimated a tutee's correct understanding of the subject matter. None of these studies, however, has related formative and summative measures of assessment to each other. Thus, so far, interrelations between tutors' assessments at varying moments in the tutoring process have not been examined.

Tutors' Assessments of a Tutee's Expressed Knowledge Deficits. Besides the relevance of a tutor's assessment accuracy in general, research has paid interest to the assessments of a tutee's expressed knowledge deficits (e.g., Chi et al., 2004; Cromley & Azevedo, 2005; Graesser et al., 1995; Putnam, 1987). A tutee's knowledge deficits comprise simple missing knowledge pieces but also complex misconceptions (Chi et al., 2004; misconceptions are naïve normatively incorrect beliefs about a subject matter that are overall resistant to change, cf., Chi, 2005; Vosniadou, 1999).

Research has shown that knowledge deficits can seriously hamper learning (Vosniadou, 1999; Vosniadou, Vamvakoussi & Skopeliti, 2008). Given the detrimental effects knowledge deficits can have, the accurate assessment of a tutee's knowledge deficits seems necessary. Furthermore, a tutee's expressed knowledge deficits are diagnostically informative because they indicate what a tutee does not know (Chi et al., 2004; Graesser et al., 1995). Research has shown, however, that assessing knowledge deficits seems to be a particular challenge for tutors (e.g., Chi et al., 2004; Graesser et al., 1995). Nevertheless, studies have also found that tutors sometimes respond with specific strategies to a tutee's expressed knowledge deficits that can be regarded as strategies of formative assessment. These strategies comprise a tutor scaffolding or giving feedback (Chi et al., 2004; Cromley & Azevedo, 2005; Graesser et al., 1995; McArthur et al., 1990).

Research on Assessments Outside the Tutoring Context

Instructors Strengths and Limits in Assessments. Studies from outside the tutoring context corroborate the observation that instructors have difficulty in assessing a learners understanding. These studies most often examined the accuracy of an instructor's summative assessment with regard to a formal test the learners took (for overviews, see Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012). Whereas instructors examined in these studies quite accurately knew how a learner performs relative to other learners in a (fictitious) class, they often largely overestimated a learner's absolute learning outcome (e.g., Feinberg & Shapiro, 2009; Südkamp et al., 2012; Südkamp, Möller, & Pohlmann, 2008). However, studies also have shown that variance of assessment accuracy among instructors is large (for overviews, see Hoge & Coladarci, 1989; Südkamp et al., 2012). Moreover, some studies have demonstrated that formative assessments benefit learning (Furtak et al., 2008; Wiliam et al., 2004; see also Black & Wiliam, 1998).

Systematic Influences on Assessment Accuracy. With regard to influences on assessment accuracy, two studies from outside the tutoring context have suggested that experienced instructors (i.e., classroom teachers) can more accurately assess a learner's understanding than inexperienced instructors (i.e., university students; Dünnebier, Gräsel, & Krolak-Schwerdt, 2009; Mulholland & Berliner, 1992; cf. also Krolak-Schwerdt, Böhmer, & Gräsel, 2009). The assessment accuracy of tutors might similarly be influenced by the level of the tutor's experience. Differences in tutors' assessment accuracy might parallel the finding that tutoring of experienced tutors seems to be more effective than tutoring of inexperienced tutors (Graesser et al., 2011; cf. the section *Evidence for the Effectiveness of Tutoring*). However, direct comparisons are scarce even outside the tutoring context. Moreover, evidence of the impact of experience on assessment accuracy is inconclusive (see also Hinds, 1999; Nathan & Petrosino, 2003). Furthermore, there is no research that has directly compared the formative and summative assessment accuracy of inexperienced and experienced tutors in one-on-one tutoring with each other. Generally, studies from outside the tutoring context are valuable to inform research on tutoring. Nevertheless, it is yet to be shown whether results that are valid in, for example, classroom contexts hold for the one-on-one situation in tutoring as well.

Overall, research curiously suggests that tutors' assessment accuracy is less than optimal. However, other findings emphasize the relevance of assessments in instruction. Moreover, a few studies have suggested that tutors have at least some assessment skills.

Studies have also implied that tutors differ in their assessment skills. Thus, results are inconclusive with respect to the quality of tutors' assessments. Finally, research relating different forms of assessments to each other is missing.

Training Tutors

Given the suboptimal assessment accuracy of tutors and the significance of accurate assessments for instruction, it seems obvious to think about teaching tutors instructional strategies that help them to assess a tutee's understanding. Nonetheless, so far, no explicit training of assessment has been conducted. In one study, Wittwer, Nückles, Landmann, and Renkl (2010) provided a group of tutors with information about their respective tutee's prior knowledge. These tutors were more adaptive to their tutee's level of knowledge than tutors who were not provided with information. Moreover, these tutors elicited deeper learning in their tutee than the uninformed tutors. Yet, the tutors in this study did not have to assess a tutee's understanding themselves.

In another study, Casey and Williamson (2011) trained parents to tutor their child. The training method included direct instruction of the tutoring method, role play with feedback, and provision of written instructions. During tutoring, parents were able to detect their child's errors. Furthermore, the children's performance increased from pretest to posttest. The aim of the study was, however, not to train assessment accuracy. Consequently, the relationship between assessments and learning remains unclear. As there also was no control group, neither the parents' accurate assessments nor the children's learning can unambiguously be attributed to the training method.

The evidence, thus, is inconclusive. Nevertheless, the two studies (Casey & Wiliamson, 2011; Wittwer, Nückles, Landmann et al., 2010) provide first hints that it is possible to train a tutor's assessment accuracy. They also imply that enhanced assessments support learning. They indicate that (short) interventions can change a tutor's instructional strategies and that they can, thereby, enhance effectiveness of tutoring. However, the studies (Casey & Wiliamson, 2011; Wittwer, Nückles, Landmann et al., 2010) leave the question completely unanswered as to what a tutor should learn to do to enhance assessment accuracy. That is, they do not provide information about possible contents of a training of assessment accuracy.

Evidence from Training Tutors' Effectiveness – Laboratory Studies

Moreover, direction on how to train assessment accuracy best is missing. Recommendations may come from research that seeks to enhance the effectiveness of inexperienced tutors. This research has already tested methods for changing a tutor's instructional strategies. Few laboratory studies provide insight (Chi et al., 2001; Randell, Hall, Bizo, & Remington, 2007). In these studies, trained tutors were informed about the nature and advantages of the strategies to be trained (cf., Brown, Campione & Day, 1981). In addition, they received worked-out examples of strategy use (cf. Renkl. 2005; Schworm & Renkl, 2007) and written instructions (Chi et al., 2001). Alternatively, trained tutors interacted ten to twenty minutes with a computer simulation applying the strategies to be trained, or they watched a playback version of the computer simulation that modeled correct strategy use (i.e., presented worked-out examples of strategy use; Randell et al., 2007). Compared with untrained controls, training resulted in enhanced application of the trained strategies (Chi et al., 2001) and in enhanced declarative and procedural knowledge about the strategies (Randell et al., 2007). However, effects of training on a tutee's learning were either not found (Chi et al., 2001) or not even measured (Randell et al., 2007).

Evidence from Training Tutors' Effectiveness – Field Studies

More evidence on how to change tutors' interactional strategies comes from field studies on training parents to tutor their child. In particular, the tutoring method Dialogic Reading (Whitehurst et al., 1988) was implemented in several training studies. Dialogic Reading is a structured method of joint book reading that aims at fostering children's language development. The training studies experimentally or quasi-experimentally compared parents, who were trained to apply Dialogic Reading, with parents, who were instructed to read to their child as usual (for an overview, see Mol, Bus, de Jong, & Smeets, 2008).

As compared with the untrained controls, trained parents implemented the trained instructional strategies (but see Arnold, Lonigan, Whitehurst, & Epstein, 1994 for mixed results). Children of trained parents improved their oral language skills (e.g., Whitehurst et al., 1988; Blom-Hoffman et al., 2006). The original training method comprises verbal information about the strategies, modeled application of the strategies (i.e., worked-out examples, cf. Renkl, 2005), and role play with feedback, provided during two half-hour training sessions (e.g., Whitehurst et al., 1988). Nevertheless, training was also effective when it was provided via short videos, lasting 15 to 45 minutes, which included verbal descriptions and modeled application of the strategies to be trained (Arnold et al., 1994;

Blom-Hoffman et al., 2006). Furthermore, training was effective, when it was provided as 20 minutes verbal telephone training (Chow & McBride-Chang, 2003). With any training method, parents received written instructions. These studies provide further evidence that a tutor's knowledge and instructional strategies can be changed by interventions, even by very short interventions.

Some implications for training tutors' assessment accuracy can be derived from the literature reviewed. First, training tutors to apply certain instructional strategies during tutoring seems possible. Applying these strategies can lead to enhanced effectiveness of tutoring. Second, effective training methods often included worked-out examples that modeled the strategies to be trained. Impact of these interventions on a tutee's learning was found in field trials but not in laboratory studies. Third, applying trained instructional strategies might also lead to enhanced assessment accuracy and, *consequently*, to enhanced effectiveness. Forth, studies have shown that even short interventions can have the power to change the instructional strategies of tutors and the outcomes of tutoring. Effectiveness of short interventions would add practical relevance to training tutors' assessment accuracy. This is because it has been criticized that training tutors may be too costly and too laborious to meet the challenges of real world conditions (e.g., Baker, Gersten, & Keating, 2000; Belzer, 2006; Graesser et al., 2011).

Aims and Contents of the Doctoral Thesis

The aim of this doctoral thesis is to provide a more comprehensive picture of tutors' assessments of a tutee's understanding in one-on-one tutoring. To this end, two studies of one-on-one tutoring were conducted (for overviews of the studies, see Table 1 and Table 2). *Chapter 1* summarizes the results of the first study. The chapter comprises two articles. These articles are *Does it Make a Difference? Investigating the Assessment Accuracy of Teacher Tutors and Student Tutors* (Herppich, Wittwer, Nückles, & Renkl, 2013b; i.e., article 1) and *Addressing Knowledge Deficits in Tutoring and the Role of Teaching Experience: Benefits for Learning and Summative Assessment* (Herppich, Wittwer, Nückles, & Renkl, 2013a; i.e., article 2). The two articles report findings on the assessment accuracy of tutors with varying levels of experience. In this regard, the analyses described in the articles also attend to the relationships between a tutor's assessments at different moments in the tutoring process and to the relationship of assessments with the effectiveness of tutoring. *Chapter 2* summarizes results of the second study. An article is presented that reports results on whether a tutor's assessments can be enhanced via a short

training intervention. The article is titled *Benefits for Processes Cause Decrements in Outcomes: Training Improves Tutors' Interactivity at the Expense of Assessment Accuracy* (Herppich, Wittwer, Nückles, & Renkl, in press; i.e., article 3).

In both studies, tutors taught the structure and the function of the human circulatory system. All tutors were knowledgeable about the topic. By having tutors teach the human circulatory system, this doctoral thesis employed a conceptual content domain for the study of tutors' assessments. Hence, tutees were to learn about certain concepts and their interrelations rather than they were to learn problem-solving skills (cf., Chi et al., 2004). Most studies in tutoring research use a procedural domain. That is, they study tutoring of problem-solving (e.g., quantitative kinematics, Chi et al., 2008; decoding, Cromley & Azevedo, 2005; algebra, McArthur et al., 1990).

The human circulatory system, however, was chosen as content domain for several reasons. First, the human circulatory system is generally a well-researched content domain. It is adequately described in many textbooks. It is taught at school. People are familiar with it, at least on a superficial level. Moreover, it is widely accepted as a relevant topic, for example, with regard to the issue of cardiovascular diseases. Consequently, second, models of a learner's understanding about the human circulatory system have been developed and common misconceptions have been documented (e.g., Azevedo, Cromley, & Seibert, 2004; Chi, De Leeuw, Chiu, & LaVancher, 1994; Chi et al., 2001; Michael et al., 2002; Riemeier et al., 2010). A learner's understanding of such a well-researched content can be described very precisely. Therefore, the human circulatory system can optimally serve as a basis for determining a tutor's assessment accuracy. Third, the content domain has previously been used to study the assessment accuracy of tutors (Chi et al., 2001, 2004). This fact comes in handy for comparing the results obtained in this doctoral thesis with the results of previous research. Fourth, the human circulatory system is a well-structured content domain. Structure is a prerequisite for tutoring to be maximally effective (Graesser et al., 2011).

In both studies the tutees were seventh-grade students. According to relevant German curricula of biology, the human circulatory system is taught in the fifth or sixth grade for the first time. Hence, seventh-grade students should have acquired some knowledge about the content domain to work with not too long ago. However, they should not possess a perfect understanding of the topic.

Comparing the Assessments of Experienced and Inexperienced Tutors (Chapter 1: Articles 1 and 2)

Chapter 1 documents the results of a study that drew on a contrastive approach (Chi, 2006) to examine the assessments of classroom teachers of biology (i.e., teacher tutors) as experienced tutors and of university students of biology (i.e., student tutors) as inexperienced tutors (see Table 1). The study refers to the first approach and the second approach to studying tutoring sensu Graesser et al. (2011). According to the first approach, it tested the impact of teaching experience on a tutor's assessments. According to the second approach, at the same time, it analyzed tutoring processes and their relationships with tutoring outcomes. This was done to obtain more comprehensive evidence than by applying only one approach.

In practical applications of tutoring, such as private tutoring, mainly inexperienced tutors provide instruction but sometimes also experienced tutors are employed (e.g., Chi et al., 2001; Graesser et al., 2011). Classroom teachers likely possess comprehensive knowledge about students and about assessments in classroom environments (e.g., Krolak-Schwerdt et al., 2009; Martínez, Stecher, & Borko, 2009). Hence, they can be thought of as experienced tutors. University students, on the contrary, likely lack this knowledge. Therefore, they can be thought of as inexperienced tutors (cf. also Chi et al., 2001; Graesser et al., 2011).

Article 1

Description of Variables. Article 1 takes a predominantly summative perspective to compare teacher tutors and student tutors in their assessments of a tutee's understanding. To this end, a method introduced by Chi et al. (2004) was adapted. Accordingly, assessment accuracy was measured at two levels of a tutee's understanding about the human circulatory system. It was measured at the level of single concepts and at the level of mental models. Concepts refer to proposition type pieces of knowledge like *the heart pumps blood*. Mental models refer to a tutee's understanding about the human circulatory system in terms of an integrated knowledge. Assessment accuracy at the level of concepts was measured after the tutoring session. Assessment accuracy at the level of mental models was measured twice, a first time after half of the tutoring session and a second time after the tutoring session. For both the level of concepts and the level of mental models two measures of assessment accuracy were determined. The *relative assessment accuracy* was measured to determine if a tutor could assess whether the tutee's learning outcome was

relatively low or relatively high. The *absolute assessment accuracy* was measured to identify the extent to which a tutor was able to assess the absolute learning outcomes of the tutee.

Tested Hypotheses. It was hypothesized, first, that all tutors would overestimate a tutee's understanding at the level of concepts and at the level of mental models (Chi et al., 2004; Feinberg & Shapiro, 2009; Südkamp et al., 2008). However, it was assumed that teacher tutors would assess a tutee's understanding more accurately than student tutors would assess a tutee's understanding (Dünnebier, et al., 2009; Mulholland & Berliner, 1992). Superiority of the teacher tutors' assessments as compared to the student tutors' assessments was hypothesized for both levels of understanding and for both measures of assessment accuracy.

Furthermore, it was assumed that tutors formatively assess a tutee's understanding and thereby accumulate individual information about the tutee's understanding (Birenbaum et al., 2006; Black, 1993; Snow & Swanson, 1992). Accordingly, tutors' assessment accuracy at the level of mental models, second, should improve in the course of tutoring. Again, the teacher tutors' assessment accuracy should improve more strongly than the student tutors' assessment accuracy (cf. Dünnebier et al., 2009).

After tutoring, a tutor's self-ratings of assessment accuracy at the level of mental models were measured. Little is known about tutors' monitoring of their assessments. Thus, it is an open question as to whether tutors are aware of their difficulty in assessing a tutee's understanding. Teachers are familiar with assessments and assessment difficulties in classroom situations (Martínez et al., 2009). Therefore, they should have a general understanding of the difficulties of assessment. Teachers should, furthermore, be able to spend some cognitive resources on self-monitoring processes. This is because they are experienced instructors (Feldon, 2007; Wittwer, Nückles, & Renkl, 2010; Zimmerman, 2006). As said before, university students likely lack these experiences. Consequently, it was, third, hypothesized that teacher tutors should be more accurate than student tutors in self-rating the accuracy of their assessments after half of the tutoring session and after the completed tutoring session.

Article 2

Description of Variables. As compared with article 1, article 2 focuses on tutors' formative assessments. More precisely, article 2 focuses on teacher tutors' and student tutors' use of

strategies to formatively assess a tutee's expressed knowledge deficits (Chi et al., 2004; Cromley & Azevedo, 2005; Graesser et al., 1995; McArthur et al., 1990). The article reports on the tutors' use of scaffolding and feedback relative to the tutors' use of explaining comments (i.e., *correct answers*) in response to a tutee's expressed knowledge deficits. Two types of expressed knowledge deficits were differentiated. First, an expressed knowledge deficit could originate from a tutor's deliberate elicitation of the tutee's understanding (i.e., *tutor-initiated* expressed knowledge deficits). Second, a knowledge deficit could be spontaneously expressed by the tutee (i.e., *tutee-initiated* expressed knowledge deficits). Two more variables were examined in article 2. These were the tutors' absolute summative assessments of a tutee's understanding at the level of single concepts and the tutee's learning.

Tested Hypotheses. The first hypothesis was based on previous research investigating the relationship between formative assessments and learning (Furtak et al., 2008; Wiliam et al., 2004; see also Black & Wiliam, 1998). It was hypothesized that tutees of tutors who more often engage in strategies of formative assessment in response to the tutee's expressed knowledge deficits should learn more than tutees of tutors who less often engage in strategies of formative assessment in response to the tutee's expressed knowledge deficits.

The second hypothesis was based on assumptions about the relationship between formative assessments and summative assessments (e.g., Birenbaum et al., 2006; Black, 1993). It was hypothesized that tutors who more often engage in strategies of formative assessment in response to the tutee's expressed knowledge deficits should summatively assess a tutee's understanding after tutoring more accurately than tutors who less often engage in these strategies.

Moreover, it was assumed that tutors should be better prepared to respond to a tutor-initiated expressed knowledge deficit than to respond to a tutee-initiated expressed knowledge deficit. This is because tutee-initiated expressed knowledge deficits are not expected by tutors (Shavelson, 2006). Thus, the third hypothesis was that tutors should more often engage in strategies of formative assessment in response to tutor-initiated expressed knowledge deficits than in response to tutee-initiated expressed knowledge deficits.

Furthermore, research suggests that experienced and inexperienced tutors differ in their use of instructional strategies that are discussed in the context of formative assessment (Chi et al., 2004; Cromley & Azevedo, 2005; Graesser et al., 1995; McArthur

et al., 1990; see also Black & Wiliam, 2009; Chi, 2009; Hmelo-Silver & Barrows, 2006). Experienced tutors regularly make use of strategies of formative assessment such as scaffolding and giving feedback. Inexperienced tutors, on the contrary, are more prone to giving lengthy explanations (Cade et al., 2008; Chae, Kim, & Glass, 2005; Chi et al., 2001, 2008; Cromley & Azevedo, 2005). Based on this research, the fourth hypothesis stated that teacher tutors should more often cause their tutees to express knowledge deficits than should student tutors. In addition, the fifth hypothesis was that teacher tutors should more often engage in strategies of formative assessment in response to a tutee's expressed knowledge deficits than should student tutors.

To examine the relationships between a tutor's experience, a tutor's formative assessments and a tutee's learning a mediation hypothesis was put forward. Sixth, it was hypothesized that teacher tutors would support a tutee's learning more strongly than student tutors. This result should be attributable to the difference in the extent to which teacher tutors and student tutors engage in strategies of formative assessment in response to a tutee's expressed knowledge deficits.

A second mediation hypothesis was put forward to examine the relationships between a tutor's experience, a tutor's formative assessments and a tutor's summative assessments. Seventh, previous research has suggested and results presented in article 1 have confirmed that teacher tutors more accurately summatively assess a tutee's understanding at the level of concepts after tutoring. It was hypothesized that this effect should, again, be explained by the difference in the extent to which teacher tutors and student tutors engage in strategies of formative assessment in response to a tutee's expressed knowledge deficits.

Table 1
Overview of Study 1 (Article 1 and 2)

Approaches	First approach: Influences of general characteristics on tutoring Second approach: Structures and processes of tutoring
Independent Variable	Tutors' teaching experience with two levels (teacher tutors vs. student tutors)
Dependent Variables: Process Measures	*Tutees' expressed knowledge deficits* *Tutor-initiated* *Tutee-initiated* *Extent of tutors' formative assessment*
Dependent Variables: Effect Measures	Tutor' summative assessment accuracy at the level of mental models (measured twice) Absolute Relative *Tutor' summative assessment accuracy at the level of concepts (measured once)* *Absolute* Relative Tutors' self-ratings of assessment accuracy for mental models (measured twice) *Tutees' learning at the level of concepts*

Note. Displays the approaches to the study of tutoring that have been adopted (cf. Graesser et al., 2011), the independent variable, and the dependent variables measured. Underlined dependent variables pertain to article 1. *Italicized* dependent variables pertain to article 2. *Italicized underlined* dependent variables pertain to article 1 and to article 2.

Training Tutors to Enhance Their Assessments

(Chapter 2: Article 3)

Chapter 2 documents results of a study that was designed to enhance a tutor's assessments via changing the tutor's instructional strategies (see Table 2). The study is in line with the third approach to studying tutoring sensu Graesser et al. (2011). More precisely, a training experiment was conducted to contrast trained student tutors (i.e., *trained tutors*) with untrained student tutors (i.e., *untrained tutors*). As a consequence of the experimental design, differences in instructional strategies and differences in assessments between trained tutors and untrained tutors can be attributed directly to the training. Alternative explanations are largely ruled out by this approach.

Contents of the Training Method

Contents of a training method for fostering a tutor's assessments could not be drawn from previous training studies. This is because no study has attempted to train a tutor's assessments. Nevertheless, it is possible to identify certain instructional strategies that may be useful in such an attempt. Tutoring research has discussed scaffolding and giving feedback as interactive instructional strategies. These interactive strategies are seen as one key to the effectiveness of tutoring because they elicit constructive responses from the tutee (e.g., Chi, 2009; Chi et al., 2001; VanLehn, 2011; see also the section *Important Structures and Processes in a Tutoring Session*). Tutoring research, moreover, has documented that experienced tutors make use of these interactive instructional strategies more often than inexperienced tutors (e.g., Cade et al., 2008; Chae et al., 2005; Cromley & Azevedo, 2005; see also the section *Comparing the Assessments of Experienced and Inexperienced Tutors*). Research on assessments in classroom situations has discussed the same instructional strategies as examples of an instructor's activities in formative assessment (Black & Wiliam, 2009; Ruiz-Primo & Furtak, 2007). As outlined before, formative assessment has been found to yield learning, as well (e.g., Wiliam et al., 2004; cf. the section *Research on Assessments Outside the Tutoring Context*).

The first study reported in this doctoral thesis (see *Chapter 1*) integrated both lines of research with a focus on formative assessment. The results of article 2 have highlighted that experienced tutors more often engage in scaffolding and feedback relative to correct answers in response to a tutee's expressed knowledge deficits than do inexperienced tutors. A mediation analysis revealed that this difference in the use of strategies of formative assessment accounted for another result. That is, due to this difference, the experienced tutors were more accurate in summatively assessing a tutee's understanding at the level of concepts than the inexperienced tutors.

In utilizing these findings, the training method was designed to foster an interactive style of tutoring based on strategies of formative assessment. A tutor's use of these strategies during tutoring should, consequently, enhance the tutor's assessments. As inexperienced tutors seem to be less able to assess a tutee's understanding than experienced tutors, only university students of biology participated as (student) tutors in the tutoring sessions conducted for the study.

How Tutors' Assessments Could be Trained

Direction on how to train tutors to improve their assessments has been obtained from research on training tutors to enhance the effectiveness of tutoring. Further direction has been obtained from research on training learning strategies and from research on training cognitive skills in general. Research on training tutors' effectiveness has found that rather short interventions have been adequate to enhance a tutor's knowledge about a tutoring method (Randell et al., 2007). These short interventions have also been adequate to change a tutor's instructional strategies (Arnold et al., 1994; Blom-Hoffman et al., 2006; Chi et al., 2001), and to enhance the effectiveness of tutoring (Arnold et al., 1994; Blom-Hoffman et al., 2006; Chow & McBride-Chang, 2003). These findings were informative for the design of the study presented in this doctoral thesis. This is because training tutors has been criticized for being often too costly and too laborious for practical use (e.g., Baker et al., 2000; Belzer, 2006; Graesser et al., 2011). Designing a short and economic intervention, therefore, seemed appropriate.

All of the interventions cited included some kind of verbal information about the strategies to be learned. A key feature in many of the interventions, furthermore, was the use of worked-out examples (cf. Renkl, 2005) that modeled strategy use (e.g., Chi et al., 2001; Randell et al., 2007; Arnold et al., 1994; Blom-Hoffman et al., 2006). Two studies provided training via short videos (Arnold et al., 1994; Blom-Hoffman et al., 2006). These studies reported particular promising results regarding the effect of the training on tutors' instructional strategies and on tutees' learning. Research on training cognitive skills has presented further evidence on the effectiveness of videos that model a skill. More precisely, video-based examples that model a skill were particularly effective when learners were prompted to self-explain the content of the video (Schworm & Renkl, 2007). That is, accompanying a video-based example, learners were asked to analyze the video. To do so, the learners had to answer questions about the given example of the skill to be learned.

Finally, research on training learning strategies advises training methods to include several principles (Klauer, 1988; Friedrich & Mandl, 1992; Pressley, Goodchild, Fleet, Zajchowski, & Evans, 1989). First, training methods should inform about the advantages associated with the strategies targeted by the training method (cf., Brown et al., 1981). This principle is meant to enhance the motivation for strategy use. Second, training methods should provide comprehensive information about the strategies. This principle is meant to support the learner in constructing declarative knowledge about the strategies. Third,

training methods should help to practice the strategies to be learned. This principle is meant to foster procedural knowledge about the strategies. When learning strategies were trained according to these principles, results were particularly positive (e.g., Dignath, Buettner, & Langfeldt, 2008; Leutner, Leopold, & Elzen-Rump, 2007).

Based on these findings, a training method was set up that instructed the trained tutors to implement an interactive tutoring style to enhance their assessments. First, the training method informed the trained tutors about the advantages of interactive instructional strategies for assessing a tutee's understanding. Second, the trained tutors were informed about interactive instructional strategies of formative assessment such as scaffolding and giving feedback. The presentation of each strategy was accompanied by video examples. Third, the trained tutors were confronted with more video examples of strategy use and were prompted to self-explain these videos. Different kinds of video-examples and self-explanation prompts were used to stimulate the application of the trained strategies (cf. Klauer, 1988; Schworm & Renkl, 2007).

Article 3 presents results of the training study. It was, first, hypothesized that trained tutors should engage in interactive instructional strategies to a larger extent than untrained tutors. When tutors use interactive instructional strategies to a larger extent they are also assumed to show more formative assessment activity than inexperienced tutors (Black & Wiliam, 2009; Ruiz-Primo & Furtak, 2007). The second hypothesis stated that trained tutors should more accurately summatively assess a tutee's understanding after tutoring than untrained tutors. Moreover, the first study (see *Chapter 1*, article 2) has established that the extent of a tutor's formative assessment accounts for the difference in summative assessment accuracy between experienced and inexperienced tutors. Based on this finding, the third hypothesis was put forward. It was hypothesized that the more interactive tutoring style of trained tutors should explain why trained tutors are more accurate than untrained tutors in summatively assessing a tutee's understanding after tutoring.

Table 2
Overview of Study 2 (Article 3)

Approach	Third approach: Experimental manipulation of tutors' activities
Independent Variable	Training of interactive tutoring style with two levels (trained tutors vs. untrained tutors)
Dependent Variable: Process Measure	Interactivity of tutors' tutoring style: Tutees' tutor-initiated expressed knowledge deficits
Dependent Variable: Effect Measure	Tutor' absolute summative assessment accuracy at the level of concepts (measured once)

Note. Displays the approach to the study of tutoring that has been adopted (cf. Graesser et al., 2011), the independent variable, and the dependent variables measured.

Chapter 1

Article 1:

Does it Make a Difference? Investigating the Assessment Accuracy of Teacher Tutors and Student Tutors

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013b). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*, 242-260. doi: 10.1080/00220973.2012.699900

THE JOURNAL OF EXPERIMENTAL EDUCATION, *81*(2), 242–260, 2013

ISSN: 0022-0973 print/1940-0683 online
DOI: 10.1080/00220973.2012.699900

Routledge
Taylor & Francis Group

LEARNING, INSTRUCTION, AND COGNITION

Does it Make a Difference? Investigating the Assessment Accuracy of Teacher Tutors and Student Tutors

Stephanie Herppich and Jörg Wittwer
University of Göttingen, Germany

Matthias Nückles and Alexander Renkl
University of Freiburg, Germany

Tutors often have difficulty with accurately assessing a tutee's understanding. However, little is known about whether the professional expertise of tutors influences their assessment accuracy. In this study, the authors examined the accuracy with which 21 teacher tutors and 25 student tutors assessed a tutee's understanding of the human circulatory system in the course of tutoring. The authors found that the teacher tutors were more accurate than were the student tutors in assessing whether a tutee had a low or high level of knowledge about concepts relevant to the human circulatory system. In addition, in comparison with the student teachers, the teacher tutors more accurately assessed the number of concepts that a tutee would know. However, the teacher tutors and the student tutors did poorly in assessing a tutee's mental model of the human circulatory system even though the teacher tutors were more aware of their assessment difficulties than were the student tutors.

Keywords *assessment accuracy, expertise in teaching, expert-novice comparison, human circulatory system, human tutoring*

IT IS WIDELY ACKNOWLEDGED THAT instruction should best be adapted to the individual learner (e.g., Corno & Snow, 1986; Jonassen & Grabowski, 1993; Kalyuga, 2007). Human one-to-one tutoring is a method particularly suitable for providing adaptive instruction because tutors have the opportunity to be responsive to a tutee's current understanding on a moment-to-moment

Parts of this article are based on a paper presented at the 2011 conference of the Cognitive Science Society in Boston, Massachusetts, USA. This research was supported by grants from the German Science Foundation (DFG, WI 3348/2–1). The authors thank Julian Etzel, Imme Husmeier, Tatjana Scharping, Anika Schoneville, and Raoul Zimmermann for their help with many practical aspects of the project.

Address correspondence to Stephanie Herppich, Educational Institute, University of Göttingen, Waldweg 26, D-37073 Göttingen, Germany. E-mail: stephanie.herppich@sowi.uni-goettingen.de

basis (e.g., Chi & Roy, 2010; Katz, Allbritton, & Connelly, 2003; Lehman, Matthews, D'Mello, & Person, 2008; Snow & Swanson, 1992). However, to do so, tutors must be able to assess a tutee's understanding accurately. Prior research has shown that tutors often have difficulties in collecting diagnostically relevant information. This seems to be true irrespective of whether teachers or students serve as tutors (for an overview, see Chi, Siler, & Jeong, 2004).

Nevertheless, even though teachers and students often provide tutoring, no previous study has directly examined the ability to accurately assess a tutee's understanding as a function of a tutor's level of expertise in teaching. In this article, we present a study in which we used a contrastive approach from research on expertise (Chi, 2006) to compare classroom teachers as experts in teaching with university students as novices in teaching. In general, experts have been shown to excel in comparison to novices (for an overview, see Ericsson, Charness, Feltovich, & Hoffman, 2006). Therefore, it seems intuitively obvious that classroom teachers are more accurate than are university students in assessing a tutee's understanding. However, a common weakness of experts is that they have difficulty with assessing the understanding of people with less expertise (for a review, see Chi, 2006). In this study, we will show under which circumstances classroom teachers provide more accurate assessments than university students and under which circumstances they fail to do so.

Tutors' Assessment of a Tutee's Understanding

Previous studies on the assessment accuracy of tutors can be roughly divided into two types: (a) studies that examined the assessment skills of classroom teachers who served as tutors (i.e., teacher tutors) and (b) studies that examined the assessment skills of university students who served as tutors (i.e., student tutors).

Teacher Tutors

Putnam (1987) examined whether mathematics teachers who served as tutors would form a mental model of a second-grade tutee's individual understanding in the course of tutoring. He found that the teacher tutors rarely took into account a tutee's specific needs. Instead, the tutorial actions were based on a curriculum script that largely determined which problems were to be provided to a tutee. Similarly, Chi, Roy, and Hausmann (2008) selected an experienced physics teacher to serve as a tutor for undergraduate university students as tutees. The teacher tutor did not adapt the difficulty level of the presented problems to a tutee's level of understanding. Thus, irrespective of whether the tutees were good problem solvers, the teacher tutor always provided the tutees with a similar rate of easy and difficult problems to be learned. Also, McArthur, Stasz, and Zmuidzinas (1990) found that tutors who were experienced mathematics teachers failed to take into account their ninth-grade or tenth-grade tutee's comprehension problems. Instead, the teacher tutors mainly asked questions such as *Do you understand?* which are not really diagnostically informative (for more details, see Chi et al., 2004).

Student Tutors

In Graesser, Person, and Magliano (1995), advanced university students provided tutoring on research methods for undergraduate university students as tutees. The student tutors rarely

attempted to correct the tutees' misconceptions. Chi et al. (2004) presented an in-depth analysis of a tutor's ability to accurately assess an eighth-grade tutee's understanding of the human circulatory system. They developed a methodology to directly measure a tutor's assessment skills. The tutors who were university students were asked to draw and explain what they thought the tutees would know about the blood path. In the same way, the tutees were asked to draw and explain the blood path as they knew it. Chi et al. (2004) analyzed a tutor's assessment accuracy at two levels. At the level of propositions, they compared the concepts that the student tutors assumed the tutees to mention in their explanations with the concepts that the tutees actually mentioned in their explanations. It was found that the student tutors overestimated the number of the tutees' correct concepts (e.g., "The aorta is an artery"; Chi et al., 2004, p. 374). At the level of mental models, Chi et al. (2004) compared the drawings of the blood path that the student tutors assumed the tutees to make with the drawings of the blood path that the tutees actually made. Likewise, it turned out that the student tutors overestimated the number of correctly drawn blood paths. Hence, the results showed that the student tutors overestimated a tutee's correct understanding. Chi et al. (2004) attributed the results to the tutors' bias to use their own normative understanding as a basis for assessing a tutee's understanding.

In sum, the findings of the studies with teachers and university students as tutors suggest that tutors have difficulty with accurately assessing a tutee's understanding. Thus, irrespective of the level of expertise in teaching, tutors seem to fall short when assessing a tutee's understanding. However, it is important to note that none of the five studies reported (Chi et al., 2004, 2008; Graesser et al., 1995; McArthur et al., 1990; Putnam, 1987) directly compared the assessment skills of teacher tutors with the assessment skills of student tutors. Therefore, strictly speaking, it remains open as to whether there are differences in the assessment accuracy between teacher tutors and student tutors. To elucidate possible differences, research on the accuracy of judgments about learners and novices outside the tutoring context is particularly instructive.

Outside the Tutoring Context: Accuracy of Judgments About Learners and Novices

The accuracy of judgments about learners in the context of classroom teaching has been intensively investigated (for an overview, see Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012). Most studies look at classroom teachers whereas few studies are concerned with university students or compare classroom teachers with university students. In addition, there are studies that examine the assessment accuracy of experts and novices outside the educational context.

Classroom Teachers

It is well documented that classroom teachers are accurate in knowing how a learner performs relative to other learners in a class, as reflected in correlations between the classroom teachers' estimates of the learners' performance and the learners' actual performance (Hoge & Coladarci, 1989; Südkamp et al., 2012). Moreover, when looking at the absolute level of the classroom teachers' estimates of the learners' performance, as reflected in the agreements between the teachers' estimates of the learners' performance and the learners' actual performance, classroom teachers often overestimate the learners' performance (e.g., Bates & Nettelbeck, 2001). In particular, classroom teachers have difficulty with accurately assessing the performance of low-performing

learners (e.g., Feinberg & Shapiro, 2009; Leinhardt, 1983; Lin & Chiu, 2010; Madelaine & Wheldall, 2005).

University Students

In several studies, Südkamp and colleagues (e.g., Südkamp & Möller, 2009; Südkamp, Möller, & Pohlmann, 2008) investigated how accurately university students assessed the performance of fictitious K–12 learners in a virtual computer-simulated classroom. In this virtual classroom, university students were instructed to ask questions to the learners and to use the learners' answers to assess their performance. The university students were found to be fairly accurate in assessing a learner's performance relative to the performance of the other learners. However, they overestimated the learners' absolute performance. This was particularly true when assessing low-performing learners. In sum, the results are in line with the findings obtained for classroom teachers (e.g., Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989).

Comparison of Classroom Teachers and University Students

In addition to studies that examine classroom teachers or university students, there are some studies that compare classroom teachers' with university students' assessments of learners. For example, Dünnebier, Gräsel, and Krolak-Schwerdt (2009) showed that classroom teachers accurately graded the performance of a learner in a German test. They were not strongly influenced by a grade believed to be provided by an experienced colleague. In contrast, university students uncritically used the grade as an anchor (cf. Tversky & Kahneman, 1974) for their judgments. Similarly, Krolak-Schwerdt, Böhmer, and Gräsel (2009) found that classroom teachers changed between different types of processing when assessing a learner's performance. More specifically, when the classroom teachers were asked to form an initial impression of the learner they paid most attention to stereotypical information about the learner. In contrast, when the classroom teachers were asked to give an important educational recommendation they paid most attention to individual information about the learner. University students, however, failed to display such different types of processing. Last, in a study by Mulholland and Berliner (1992), classroom teachers, on average, assessed the performance of learners more accurately than university students. However, there was a large overlap in the individual assessment accuracy between classroom teachers and university students. In addition, a limitation of this study was that the classroom teachers, in contrast with the university students, had regularly taught the learners. Therefore, it cannot be ruled out that this experience might have improved the classroom teachers' assessment accuracy.

Taken together, the results suggest that classroom teachers assess a learner's understanding more accurately than university students. The difference in assessment accuracy might be explained by the fact that classroom teachers usually possess more knowledge about learners than university students (e.g., Krolak-Schwerdt et al., 2009). In addition, as opposed to university students, classroom teachers routinely dedicate a lot of their professional time to engage in assessment activities (e.g., Martínez, Stecher, & Borko, 2009; Stiggins & Conklin, 1992). Therefore, they have more experience in assessment activities that might help them to assess a learner's understanding accurately (e.g., Mulholland & Berliner, 1992).

Comparison of Experts and Novices

Nevertheless, research in the area of expertise (Chi, 2006) suggests that experts are not necessarily more successful than novices in assessing what people with less expertise know. Instead, they might be systematically inaccurate. For example, Hinds (1999) found that experts with task-specific knowledge overestimated a novice's performance on a task more strongly than non-experts. Similarly, in the field of teaching, Nathan and Petrosino (2003) observed that preservice teachers with more subject-matter expertise judged the difficulty of algebra problems for learners more inaccurately than preservice teachers with less subject-matter expertise. Last, by using the think-aloud methodology, Nückles et al. (2006) found that computer experts who gave advice to novices rarely made an attempt to take the perspective of the novices to provide adaptive advice.

Overall, the findings show that experts have more difficulties in assessing people with less expertise. This phenomenon has been coined *expert blind spot* (Nathan & Petrosino, 2003, p. 906) or *curse of expertise* (Hinds, 1999, p. 205). Such difficulties might be caused by the ready availability of an expert's domain knowledge (Hinds, 1999). As a result, experts are likely to fail to take the perspective of novices accurately (Nathan & Petrosino, 2003; Nückles et al., 2006). Hence, in contrast with studies that show that teachers as experts are more accurate than university students as novices in assessing a learner's understanding (e.g., Mulholland & Berliner, 1992), the research in the area of expertise suggests that an expert's domain knowledge might be an obstacle to providing accurate assessments.

Present Study and Hypotheses

We analyzed the accuracy with which more knowledgeable tutors (teachers, university students) assessed a less knowledgeable tutee's (K–12 student) understanding of the human circulatory system. Thus, according to Topping's (1996) typology of tutoring, we implemented one-to-one cross-ability tutoring situations with fixed roles. Using a contrastive approach (Chi, 2006), we were interested in whether the assessment accuracy of classroom teachers who served as tutors (i.e., teacher tutors) would differ from the assessment accuracy of university students who served as tutors (i.e., student tutors). To rule out that differences in the topic-specific knowledge between teacher tutors and student tutors would account for differences in assessment accuracy, as research on expertise suggests (e.g., Hinds, 1999; Nathan & Petrosino, 2003), the classroom teachers and the university students who served as tutors in this study possessed comparable levels of topic-specific knowledge. Similar to the methodology used by Chi et al. (2004), we examined a tutor's assessment accuracy at two levels. At the level of propositions, we looked at a tutor's assessment of a tutee's knowledge in terms of single concepts about the human circulatory system. At the level of mental models, we looked at a tutor's assessment of a tutee's understanding in terms of an integrated knowledge about the human circulatory system.

Research has provided converging evidence that it is difficult for both tutors and teachers to assess a learner's understanding accurately (e.g., Chi et al., 2004; Feinberg & Shapiro, 2009). As a result, tutors and teachers often overestimate a learner's understanding (2009). Therefore, we expected that both teacher tutors and student tutors would overestimate a tutee's understanding at the level of propositions and at the level of mental models. However, given their expertise in teaching (e.g., professional experience with learners, learning, and assessment activities), teacher

tutors should more accurately assess a tutee's understanding at the level of propositions (*level-of-propositions hypothesis*) and at the level of mental models (*level-of-mental-models hypothesis*) than student tutors. This should be particularly true when the topic-specific knowledge of teacher tutors and student tutors did not differ from each other (Hinds, 1999; Nathan & Petrosino, 2003).

In the course of tutoring, tutors can normally accumulate individual information about a tutee and use this information to assess a tutee's understanding more accurately (Snow & Swanson, 1992). Hence, we expected that tutors would become more accurate in assessing a tutee's understanding in the course of tutoring. However, prior research suggests that university students might have more difficulty with processing information about learners than classroom teachers (Dünnebier et al., 2009; Krolak-Schwerdt et al., 2009). Therefore, we expected that teacher tutors would become more accurate in their assessments in the course of tutoring than student tutors (*improvement hypothesis*).

Even though prior research has shown that tutors have difficulties with assessing a learner's understanding, little is known about whether tutors are aware of their assessment difficulties. It can be assumed that, given their professional experience with assessment activities, classroom teachers have a better awareness of their assessment skills than university students. As classroom teachers routinely assess a learner's understanding in the context of teaching (Martínez et al., 2009), they are likely to know the conditions under which it is difficult to assess a learner's understanding accurately. In addition, it can be assumed that, due to their expertise, classroom teachers, as opposed to university students, can save more cognitive resources that can be devoted to self-monitor their assessment accuracy (Feldon, 2007; Wittwer, Nückles, & Renkl, 2010; Zimmerman, 2006). Therefore, we expected that teacher tutors would more accurately self-rate the accuracy with which they assessed a tutee's understanding than student tutors (*self-rating hypothesis*).

METHOD

Sample and Design

Participants were 46 tutor–tutee dyads. Using a contrastive approach, we selected teacher tutors (i.e., experts in teaching) and student tutors (i.e., novices in teaching) on the basis of the academic qualification of the participants. As *teacher tutors*, we had 21 biology teachers with a mean age of 44.05 years ($SD = 11.76$). Of these teacher tutors, 11 tutors were female and 10 tutors were male. On average, the teacher tutors had 13.00 years ($SD = 12.30$) of professional experience. As *student tutors*, we had 25 university students majoring in biology with a mean age of 22.24 years ($SD = 2.83$). Of the student tutors, 21 tutors were female and 4 tutors were male.

The tutees were seventh-grade students. Of the tutees, 19 were female and 27 were male. Their mean age was 12.65 years ($SD = 0.53$). The tutees were randomly assigned to one of the two tutor groups. Therefore, the tutors and the tutees did not know each other before tutoring. The main dependent variable was the accuracy with which the tutors assessed a tutee's understanding of the human circulatory system at the level of propositions and at the level of mental models.

Materials

Textbook passage (tutees and tutors)

In the tutoring session, the tutor and the tutee engaged in a dialogue on the basis of a passage about the human circulatory system. The passage was provided to tutors and tutees in previous studies by Chi et al. (2001) and taken from a textbook used in junior high schools. We adapted the passage for the present study by deleting and reformulating some sentences. Each of the remaining 59 sentences of the passage was printed on a separate sheet of paper. The sentences were presented to the tutor and the tutee in a ring binder.

Concepts test (tutees and tutors)

The test with 25 multiple-choice items (see Figure 1 [left] for an example) measured a tutee's knowledge of the human circulatory system at the local level of propositions. The items were either adapted from tests developed by Sungur and Tekkaya (2003) and by Michael et al. (2002) or constructed on the basis of literature on misconceptions of the human circulatory system (e.g., Pelaez, Boyd, Rojas, & Hoover, 2005). A correct answer to an item indicated a scientifically correct understanding. Each of the incorrect answers to an item indicated a scientifically incorrect understanding. The original 74 items were pretested with a sample of 60 eighth graders. On the basis of the results of this pretest, the resultant 25 items were selected according to the following criteria: (a) they should cover a wide range of topics related to the human circulatory system (e.g., concepts related to gas exchange, the heart, blood vessels, blood circuits) and relationships between the concepts, and (b) the concepts should be explicitly or implicitly mentioned in the passage of the textbook.

The tutee was administered the test at the beginning and at the end of tutoring. Each item that a tutee answered correctly in the concepts test was assigned 1 point. Hence, a tutee could achieve a maximum number of 25 points in the concepts test. Internal consistency of the test (administered at the beginning of tutoring) was satisfying, $\alpha = .85$.

To measure the accuracy with which the tutors assessed the tutees' knowledge of the human circulatory system at the level of propositions, the tutors were also administered the test at the end of tutoring and asked to indicate how the tutee would answer each of the test items.

Drawings of the human circulatory system (tutees and tutors)

On a sheet of paper, the outline of a human body was displayed. The tutees were asked to draw the blood path of the circulatory system into the human body and to explain the blood path orally. The explanations were audio-taped. By adopting this methodology, which was previously used by Chi et al. (2004), we assessed a tutee's understanding of the human circulatory system at the global level of mental models. The tutees were asked to accomplish the drawing and explanation task at the beginning, in the midst, and at the end of the tutoring. Their drawings and explanations of the human circulatory system were coded by using a coding scheme adapted from Azevedo, Cromley, and Seibert (2004). The coding scheme consists of 12 mental models that reflect different types of understanding about the human circulatory system ranging from 0 (*no understanding*) to 11 (*complete understanding*; for more details, see Azevedo et al., 2004).

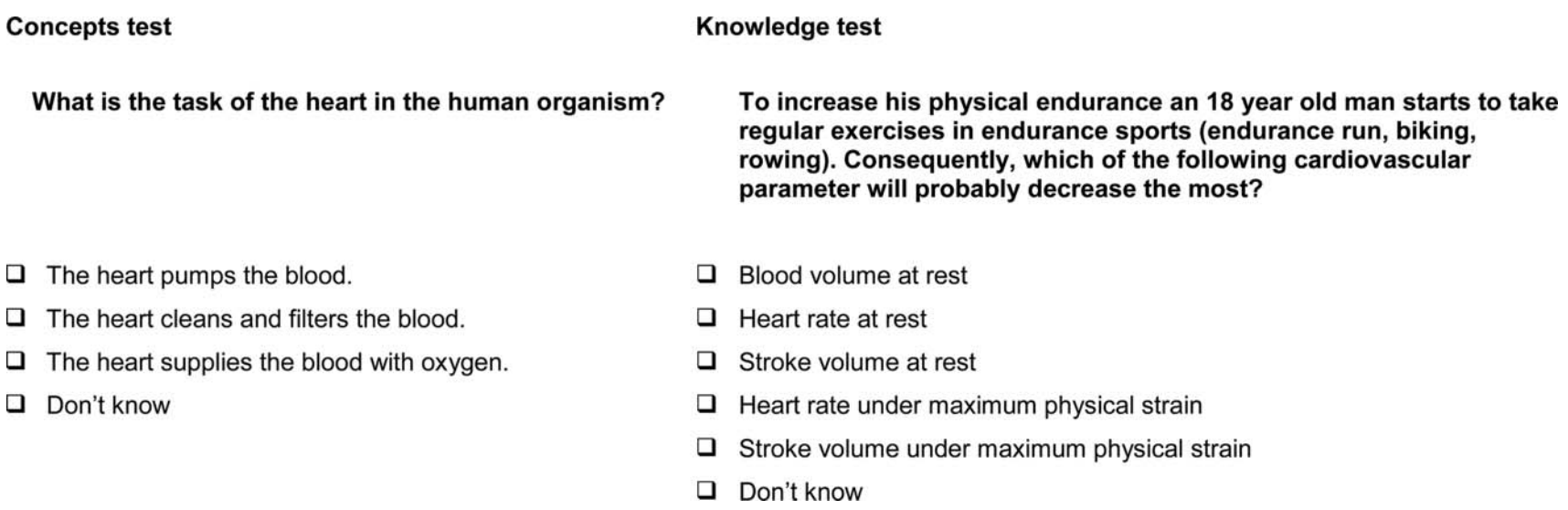


FIGURE 1 Examples of items from the tutees' concepts test (left) and the tutors' knowledge test (right).

This detailed coding scheme is an expansion of earlier work by Chi and colleagues (e.g., Chi, de Leeuw, Chiu, & LaVancher, 1994). We used it instead of the coding scheme employed by Chi et al. (1994) to analyze the tutee's mental models on an even more fine-grained level.

To measure the accuracy with which the tutors assessed a tutee's mental model of the human circulatory system, the tutors were also administered the drawing and explanation task in the midst of tutoring and at the end of tutoring. They were asked to draw and explain the blood path as they assumed the tutee to draw and explain the blood path.

The tutees' and the tutors' drawings and explanations were scored independently by two raters. Drawings and explanations were coded simultaneously by using the same assignment criteria to allow for both measures to complement one another. To standardize coding, both raters used a written code book. The code book consisted of general coding rules, descriptions of the mental models, and assignment criteria for each mental model (for the original coding scheme, see Azevedo et al., 2004). The intraclass correlations measuring absolute agreement between the two coders were satisfying both for the tutees (range of ICC[2, 1] = .73 to .81) and the tutors (range of ICC[2, 1] = .53 to .85).[1] After testing interrater agreement, discrepancies between coders were discussed until a consensus about codes was reached.

Self-ratings (tutors)

At the end of tutoring, the tutors were asked to self-rate the accuracy with which they had assessed the correctness of a tutee's mental model in the midst of tutoring and at the end of tutoring. The tutors indicated their assessment accuracy on a 4-point rating scale ranging from 1 (*very imprecisely*) to 4 (*very precisely*).

Knowledge test (tutors)

To assess the tutors' knowledge about the human circulatory system, we developed a test with 18 multiple-choice items (see Figure 1 [right] for an example). The test measured knowledge not only about basic concepts to be discussed in tutoring but also about advanced concepts of the human circulatory system, about the relationships among these concepts, and about the relevance of these concepts for life processes. Hence, answering the test correctly required different levels of knowledge. Each correct answer was assigned 1 point. The reliability of the test was satisfying, $\alpha = .76$.

Procedure

Each tutoring session was divided into three phases: pretest phase, tutoring phase, and posttest phase. On average, a tutoring session lasted about 3 hr. In the pretest phase, the tutees completed the concepts test. In addition, the tutees were asked to draw the blood path of the human circulatory system and to explain the blood path as they knew it. Afterwards, the tutees individually read the

[1]Note that the relatively low intraclass correlation obtained for the codings of the tutors' mental models at the end of tutoring is not produced by a low interrater agreement per se but by restricted variance between the codings. The restricted variance is due to the fact that the coders mainly assigned the highest codes because all tutors assumed the tutees to have a correct mental model at the end of tutoring.

passage about the human circulatory system. The tutors completed the knowledge test and were also asked to individually read the passage about the human circulatory system.

In the tutoring phase, the dyads of tutors and tutees read each sentence of the passage about the human circulatory system and engaged in a dialogue about each sentence. After the 33th sentence, tutoring was interrupted and the dyads were separated. The tutees were asked to draw and explain the blood path of the human circulatory system. Also, the tutors were required to draw and explain the tutees' mental model of the human circulatory system in order to measure what the tutors thought that the tutees would know about the blood path. Once this task was accomplished, tutoring was continued.

In the posttest phase, the dyads of tutors and tutees were separated again and asked to draw and explain the blood path of the human circulatory system. Afterwards, the tutees completed the concepts test once again. The tutors also received the items of the concepts test and were asked to indicate how the tutee would answer each of the items of the concepts test. In addition, the tutors self-rated the accuracy with which they had assessed the correctness of a tutee's mental model in the midst of tutoring and at the end of tutoring.

Analysis of Assessment Accuracy

To measure the accuracy with which the tutors assessed a tutee's understanding at the level of propositions and at the level of mental models, we compared a tutee's performance in the concepts test and in the drawing and explanation task with a tutor's estimate of a tutee's performance in the concepts test and in the drawing and explanation task. To do so, we used two measures of assessment accuracy. The *relative assessment accuracy* was calculated by computing correlations between a tutee's performance and a tutor's estimate of a tutee's performance (see Hoge & Coladarci, 1989). This measure indicates whether a tutor is able to assess the relative performance of a tutee (e.g., whether a tutee's performance is relatively low or relatively high). The *absolute assessment accuracy* was calculated by computing differences between a tutee's performance and a tutor's estimate of a tutee's performance (see Südkamp & Möller, 2009). In addition to the relative assessment accuracy, this measure indicates the extent to which a tutor is able to assess the absolute performance of a tutee (e.g., the number of items a tutee answers correctly in the concepts test).

RESULTS

For all analyses, we used an alpha level of .05. In the cases of directional hypotheses, we used one-tailed tests. We used eta square as effect size measure. Eta square estimates the proportion of variance in the dependent variable that is explained by an independent variable. Cohen (1988) suggested the following interpretation: $\eta^2 = .01$ is a small effect, $\eta^2 = .06$ is a medium effect, and $\eta^2 = .14$ is a large effect.

Preanalyses

In a first step, we examined the tutors' knowledge of the human circulatory system. This analysis was performed in order to exclude the possibility that a difference in the assessment accuracy

between the teacher tutors and the student tutors resulted from different levels of knowledge about the human circulatory system. On average, the teacher tutors correctly answered 12.43 ($SD = 3.43$) of the 18 items in the knowledge test. The student tutors correctly answered 11.56 ($SD = 3.86$) of the 18 items. Hence, the teacher tutors and the student tutors had sufficient and comparable knowledge about the human circulatory system, $F(1, 44) = 0.63$, $p = .43$, $\eta^2 = .01$ (small effect).

In a second step, we analyzed the tutees' knowledge before tutoring to exclude the possibility that a difference in the assessment accuracy between teacher tutors and student tutors resulted from a difference in the level of a tutee's prior understanding of the human circulatory system. At the level of propositions, there was no significant difference between the knowledge level of both tutee groups, $F(1, 44) = 1.58$, $p = .22$, $\eta^2 = .04$ (small effect). On average, the tutees of teacher tutors correctly answered 9.38 ($SD = 2.67$) items of the concepts test. The tutees of student tutors correctly answered 10.48 ($SD = 3.16$) items. At the level of mental models, there was also no significant difference between the tutee groups, $F(1, 44) = 0.04$, $p = .85$, $\eta^2 < .01$ (small effect). On average, the level of the mental model held by tutees of teacher tutors was 3.95 ($SD = 2.96$). The level of the mental model held by tutees of student tutors was 3.80 ($SD = 2.57$). In sum, the results showed that all tutees already had a basic understanding of the human circulatory system prior to tutoring.

Level-of-Propositions Hypothesis

The level-of-propositions hypothesis stated that the teacher tutors assessed more accurately a tutee's understanding of the human circulatory system at the level of propositions than the student tutors. To test this hypothesis, we analyzed the relative and the absolute accuracy of both tutor groups. To examine the relative assessment accuracy, we computed correlations between the items that the tutors assumed to be correctly answered by the tutees in the concepts test and the items that the tutees actually answered correctly. We found a significant correlation for the teacher tutors, $r = .38$, $p = .04$, and a correlation near zero for the student tutors, $r = .08$, $p = .70$. Hence, the teacher tutors were more accurate in assessing whether a tutee had a low or a high level of understanding in the concepts test.

To examine the absolute assessment accuracy, we compared the items that the tutees actually answered correctly in the concepts test with the items that the tutors assumed that the tutees answered correctly. On average, the tutees of teacher tutors answered 14.09 ($SD = 3.13$) items correctly and the tutees of student tutors answered 12.72 ($SD = 3.16$) items correctly. The teacher tutors assumed on average that the tutees answered 14.19 ($SD = 3.04$) items correctly and the student tutors assumed on average 14.92 ($SD = 3.37$) correctly answered items. We performed a repeated-measures analysis with the estimated number of correctly answered items and the actual number of correctly answered items as repeated-measures factor and with the type of tutor as between-subjects factor. The tutors in general significantly overestimated the number of correctly answered items, as indicated by a significant main effect for the repeated-measures factor, $F(1, 44) = 3.73$, $p = .03$, $\eta^2 = .07$ (medium effect). As expected, there was no significant main effect of the between-subjects factor, $F(1, 44) = 0.20$, $p = .67$, $\eta^2 < .01$ (small effect). However, there was a significant interaction effect between the repeated-measures factor (i.e., assumed and actual number of correctly answered items) and the type of tutor, $F(1, 44) = 3.13$, $p = .04$, $\eta^2 =$

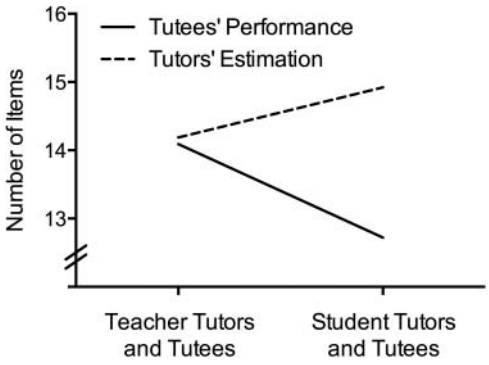


FIGURE 2 Interaction effect between number of items (correctly answered vs. assumed to be correctly answered) and type of tutor (teacher tutors, student tutors) on assessment accuracy.

.06 (medium effect). As shown in Figure 2, the teacher tutors were significantly more accurate in knowing the number of correctly answered items than the student tutors.

Level-of-Mental-Models Hypothesis

The level-of-mental-models hypothesis stated that the teacher tutors more accurately assessed a tutee's understanding of the human circulatory system at the level of mental models than the student tutors. To test this hypothesis, we analyzed the relative and the absolute accuracy with which tutors assessed a tutee's understanding at the level of mental models. To examine the relative assessment accuracy, we computed correlations between the score that was assigned to the tutors' drawings of the assumed tutees' mental model and the score that was assigned to the tutees' actual drawings. Remember that we had the tutees and the tutors draw the human circulatory system in the midst of tutoring and at the end of tutoring. The correlation for the teacher tutors in the midst of tutoring was significant, $r = .50$, $p = .01$, whereas the correlation for the teacher tutors at the end of tutoring failed to reach the level of statistical significance, $r = .20$, $p = .19$. The correlations for the student tutors in the midst of tutoring, $r = .17$, $p = .41$, and at the end of tutoring, $r = .16$, $p = .46$, both were not statistically significant. The teacher tutors were evidently more accurate in assessing whether a tutee had a relatively incorrect or a relatively correct mental model in the midst of tutoring than the student tutors. The difference in the relative assessment accuracy between the teacher tutors and the student tutors, however, was no longer observable at the end of tutoring.

To examine the absolute assessment accuracy, we compared the score that was assigned to the tutees' actual drawings of the human circulatory system with the score that was assigned to the teacher tutors' and the student tutors' drawings of the tutee's mental model of the human circulatory system. The tutees of teacher tutors achieved on average for their mental model a score of 6.10 ($SD = 2.74$) in the midst of tutoring and of 8.05 ($SD = 2.67$) at the end of tutoring. The tutees of student tutors achieved, on average, a mental model score of 6.88 ($SD = 2.58$) in the midst of tutoring and of 7.84 ($SD = 2.75$) at the end of tutoring. The teacher tutors assumed the tutees to have a mental model to be scored on average with 8.29 ($SD = 2.59$) in the midst of

tutoring and with 10.00 ($SD = 0.84$) at the end of tutoring. The student tutors assumed the tutees to have a mental model to be scored with 7.92 ($SD = 2.41$) in the midst of tutoring and with 10.12 ($SD = 0.93$) at the end of tutoring. Two repeated-measures analyses with the estimated mental model and the actual mental model as repeated-measures factor and with the type of tutor as between-subjects factor showed that the tutors in general overestimated the correctness of the tutees' mental models in the midst of tutoring, as indicated by the main effect for the repeated-measures factor in the first repeated-measures analysis, $F(1, 44) = 13.50$, $p = .001$, $\eta^2 = .23$ (large effect) and at the end of tutoring, as indicated by the main effect for the repeated-measures factor in the second repeated-measures analysis, $F(1, 44) = 27.94$, $p < .001$, $\eta^2 = .39$ (large effect). There was no main effect of the between-subjects factor in both repeated measures analyses, in the midst of tutoring, $F(1, 44) = 0.11$, $p = .74$, $\eta^2 < .01$ (small effect), at the end of tutoring, $F(1, 44) = 0.01$, $p = .92$, $\eta^2 < .01$ (small effect). More interesting to note is that there was also no significant interaction effect between the repeated-measures factor (i.e., assumed and actual mental model) and the type of tutor, neither in the midst of tutoring, $F(1, 44) = 1.72$, $p = .10$, $\eta^2 = .03$ (small effect), nor at the end of tutoring, $F(1, 44) = 0.17$, $p = .34$, $\eta^2 < .01$ (small effect). It is obvious that teacher tutors and student tutors had similar difficulties with accurately assessing the correctness of a tutee's mental model of the human circulatory system.

Improvement Hypothesis

The improvement hypothesis stated that the teacher tutors became more accurate in assessing a tutee's understanding in the course of tutoring than the student tutors. We performed a repeated-measures analysis with the difference scores reflecting the difference between the mental model as estimated by the tutor and the tutee's actual mental model in the midst of tutoring and at the end of tutoring as repeated-measures factor and with the type of tutor as between-subjects factor. We calculated the difference scores by subtracting the scores assigned to the tutees' drawings and explanations from the scores assigned to the tutors' drawings and explanations. Lower difference scores indicated more accurate assessments.

The main effect for the repeated-measures factor was not significant, $F(1, 45) = 2.08$, $p = .16$, $\eta^2 = .05$ (small effect). Thus, the tutors overall failed to provide more accurate assessments from the midst of tutoring to the end of tutoring. In addition, there was, as expected, no significant effect for the between-subjects factor, $F(1, 45) = 0.29$, $p = .59$, $\eta^2 = .01$ (small effect). However, we found a significant interaction between measurement point (i.e., midst of tutoring and end of tutoring) and type of tutor, $F(1, 45) = 4.53$, $p = .02$, $\eta^2 = .09$ (medium effect). As shown in Figure 3, the absolute assessment accuracy of the teacher tutors slightly increased from the midst of tutoring (difference score: $M = 2.19$, $SD = 2.66$) to the end of tutoring (difference score: $M = 1.95$, $SD = 2.64$), whereas the absolute assessment accuracy of student tutors decreased from the midst of tutoring (difference score: $M = 1.04$, $SD = 3.21$) to the end of tutoring (difference score: $M = 2.28$, $SD = 2.76$).

Self-Rating Hypothesis

The self-rating hypothesis stated that the teacher tutors self-rated more accurately their assessment accuracy than the student tutors. Remember that the tutors were asked to indicate their accuracy with which they had assessed the correctness of a tutee's mental model in the midst of

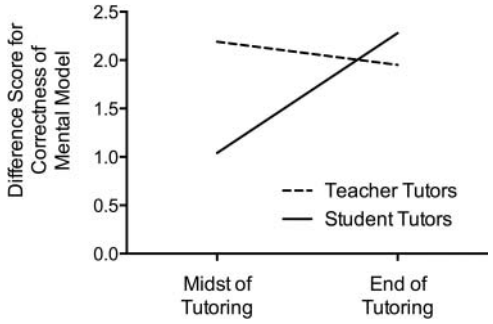


FIGURE 3 Interaction effect between measurement point (midst of tutoring, end of tutoring) and type of tutor (teacher tutors, student tutors) on assessment accuracy (lower difference scores represent a higher assessment accuracy).

tutoring and at the end of tutoring on a 4-point rating scale. The mean score of the teacher tutors' self-ratings of their assessment accuracy was 2.10 ($SD = 0.55$) in the midst of tutoring and 2.85 ($SD = 0.49$) at the end of tutoring. The mean score of the student tutors' self-ratings of their assessment accuracy was 2.52 ($SD = 0.59$) in the midst of tutoring and 3.00 ($SD = 0.58$) at the end of tutoring. We conducted a repeated-measures analysis with the tutors' self-ratings in the midst of tutoring and at the end of tutoring as repeated-measures factor and with the type of tutor as between-subjects factor. The results showed a significant main effect for the repeated-measure factor, $F(1, 43) = 30.13$, $p < .01$, $\eta^2 = .40$ (large effect). That is, the tutors in general self-rated their assessment accuracy as being significantly more accurate at the end of tutoring than in the midst of tutoring. Also, the results showed a significant main effect for the type of tutor. In other words, the self-ratings of the student tutors were significantly higher than the self-ratings of the teacher tutors, $F(1, 43) = 5.32$, $p = .01$, $\eta^2 < .11$ (medium effect). Hence, although the results of the previous analyses of the tutors' absolute assessment accuracy at the level of mental models revealed no significant difference between the teacher tutors and the student tutors, the student tutors had the impression of being more accurate than the teacher tutors. There was no significant interaction effect between the repeated-measures factor and the between-subjects factor, $F(1, 43) = 1.45$, $p = .23$, $\eta^2 = .03$ (small effect).

In addition, we compared the tutors' self-ratings with the absolute accuracy with which they assessed the correctness of the tutees' mental model in the midst of tutoring and at the end of tutoring. To do so, we computed correlations between the tutors' self-ratings and their absolute assessment accuracy. Remember that the absolute assessment accuracy was calculated by subtracting a tutee's actual mental model from a tutor's assumed mental model. Thus, more accurate assessments were indicated by lower difference scores. The correlations for the self-ratings of the teacher tutors were statistically significant: in the midst of tutoring, $r = -.47$, $p = .02$, at the end of tutoring, $r = -.56$, $p = .01$. Hence, the more the teacher tutors assumed their assessments to be accurate, the more accurate their assessments actually were. In contrast, the correlations for the self-ratings of the student tutors were low and did not reach the level of statistical significance: in the midst of tutoring, $r = .14$, $p = .49$, at the end of tutoring, $r = .00$, $p = .99$. Thus, the results showed that the teacher tutors were fairly accurate in knowing the extent to which they accurately assessed a tutee's understanding of the human circulatory system at the level of mental models whereas the student tutors were not.

DISCUSSION

In this study, we examined the extent to which teacher tutors and student tutors accurately assessed a tutee's understanding of the human circulatory system. Following the methodology by Chi et al. (2004), we analyzed a tutor's assessment accuracy at the level of propositions and at the level of mental models.

First, the study demonstrated that both the teacher tutors and the student tutors had difficulties with accurately assessing a tutee's understanding. At the level of propositions, we found that the tutors overestimated the number of items that a tutee would answer correctly in the concepts test. At the level of mental models, we observed that the tutors overestimated the correctness of a tutee's mental model of the human circulatory system. Thus, irrespective of the level of expertise in teaching, classroom teachers and university students misjudged a tutee's understanding. This finding is in line with the study by Chi et al. (2004) and with research in the area of classroom teaching showing that not only university students but also classroom teachers misjudge a learner's absolute level of performance with a bias toward overestimation (e.g., Feinberg & Shapiro, 2009; Lin & Chiu, 2010; Südkamp & Möller, 2009). It is plausible to assume that tutors rely too heavily on their own understanding as a basis for assessing a tutee's understanding. As a result, they overestimate what tutees know (see also Chi et al., 2004; Nickerson, 1999).

Second, we provided initial evidence for differences in the assessment accuracy between teacher tutors and student tutors. The teacher tutors were more accurate than the student tutors in assessing a tutee's relative and absolute understanding at the level of propositions. This finding is in line with the study by Mulholland and Berliner (1992) who showed that classroom teachers assessed learners in a classroom setting more accurately than university students. An explanation for this finding is that teacher tutors, in contrast with student tutors, have more experience with learners with different levels of understanding. In addition, they may be more experienced in assessment activities, which might help them to assess a learner's understanding accurately.

Third, at the level of mental models, the results partly confirmed our hypotheses. Even though we found that the absolute assessment accuracy of the student tutors decreased in the course of tutoring whereas the absolute assessment accuracy of the teacher tutors increased in the course of tutoring, the teacher tutors were not significantly more accurate than the student tutors in the midst of tutoring and at the end of tutoring. Hence, it was obviously difficult for teacher tutors to assess a tutee's understanding at the level of mental models. This might have been due to the complexity of the assessment task. Assessing a mental model required tutors to retrieve all pieces of information (e.g., concepts and relations among concepts) that make up a tutee's mental model from their episodic memory and to integrate these pieces into their working memory to form a complete picture of the tutee's understanding. In contrast, assessing a tutee's understanding at the level of propositions might have been relatively easy because the number of pieces of information that make up a correct proposition was rather low. In addition, the task of drawing and explaining a tutee's mental model of the human circulatory system might have been unfamiliar to both the classroom teachers and the university students in this study. Even though classroom teachers often make use of multiple-choice tests, such as the concepts test that was used in this study, in order to collect information about a learner, classroom teachers might not routinely require learners to draw and verbally explain their thoughts (Martínez et al., 2009). Therefore, it is plausible to assume that the classroom teachers (and the university students) did not know how

to use the drawing and explanation task to gain diagnostically relevant information about a tutee. This interpretation is line with findings obtained by Yin et al. (2008) who showed that forms of formative assessments in classroom teaching are not useful unless teachers are instructed in how to extract relevant information provided by such forms of assessment.

Fourth, we found that the teacher tutors more accurately knew the relative correctness of the tutee's mental model than the student tutors in the midst of tutoring. At the end of tutoring, the difference in the relative assessment accuracy between the teacher tutors and the student tutors was not observable any longer. An explanation for this finding is that the teacher tutors might have erroneously assumed that all tutees would have a more or less complete understanding of the human circulatory system at the end of tutoring even though this was not the case. This might have been because all (normatively correct) contents of the textbook passage about the human circulatory system had been discussed at the end of tutoring and, thus, had possibly been learned by the tutees. This interpretation is statistically corroborated by the relatively low variance in the teacher tutors' assessments of the tutees' mental models at the end of tutoring. Remember that whereas the variance in the teacher tutors' assessments in the midst of tutoring was relatively high ($SD = 2.59$), the variance in the teacher tutors' assessments at the end of tutoring was relatively low ($SD = 0.84$). It is obvious that at the end of tutoring, the teacher tutors made the group of tutees more homogeneous than they actually were because the variance in the correctness of the tutees' mental models was still relatively high ($SD = 2.67$).

Fifth, the results showed that the student tutors were more confident in accurately assessing a tutee's understanding than the teacher tutors. However, in reality, the teacher tutors were more accurate in judging their assessment accuracy than the student tutors. Thus, in contrast with the student tutors, the teacher tutors were quite aware of whether or not they were able to accurately assess a tutee's mental model of the human circulatory system. The ability to correctly self-rate their assessment skills might have been acquired by the teacher tutors in their profession as a classroom teacher, whereas the student tutors likely lack this professional experience (Zimmerman, 2006). In addition, it can be assumed that their expertise helped the classroom teachers to free up cognitive resources that they could use to engage in self-monitoring processes (Wittwer et al., 2010). The results also showed that both the teacher tutors and the student tutors were more confident in the accuracy of their assessments at the end of tutoring than in the midst of tutoring. However, the analysis testing the improvement hypothesis demonstrated that the student tutors became more inaccurate and the teachers became more accurate in the course of tutoring. Thus, the finding corroborates the assumption that university students were less capable of self-monitoring their assessment skills than classroom teachers.

In general, the shortcomings in the assessments of tutors observed in this study suggest that tutors, irrespective of their level of expertise in teaching, need to improve their assessment skills. Therefore, tutors might benefit from explicit training in assessment strategies that could help them to make tutoring more effective (e.g., Chi & Roy, 2010; Snow & Swanson, 1992). The results of this study suggest (a) that all tutors need to overcome their tendency to overestimate a tutee's understanding (particularly at the end of tutoring), (b) that all tutors need to be trained in applying innovative forms of assessments, such as the use of drawings, to get a more complete picture of a tutee's understanding (cf. Yin et al., 2008), (c) that university students need to receive more intensive training than classroom teachers, and (d) that university students need to be trained in self-monitoring their assessment skills (cf. Zimmerman, 2006). Future research should

empirically test whether trainings that target such factors in fact improve a tutor's assessment skills and affect, in the long run, the effectiveness of tutoring.

Up to now, little is known about the factors influencing the assessment accuracy of teachers and tutors. In particular, there is scarce information about the characteristics of teachers and tutors that affect their assessment accuracy (Südkamp et al., 2012). In this study, we examined the role of expertise in teaching as a factor accounting for assessment accuracy. Using expertise in teaching as a broad category to compare classroom teachers (i.e., experts in teaching) with university students (i.e., novices in teaching), we were successful at revealing differences in the assessment accuracy between the two groups of tutors. Of course, however, expertise in teaching is a rather coarse-grained measure comprised of many components (e.g., views of learning, Staub & Stern, 2002; pedagogical knowledge, Shulman, 1986) that are theoretically related to the ability to accurately assess a learner's understanding. Therefore, to improve our understanding about the factors that influence assessment accuracy, future research should examine in more detail which aspects of expertise in teaching are most relevant for assessing a learner's understanding accurately.

AUTHOR NOTES

Stephanie Herppich is a doctoral student and research associate in the Educational Institute at the University of Göttingen, Germany. Her research interests focus on tutoring, diagnostic expertise of instructors, and family literacy. **Jörg Wittwer** is a professor of empirical educational research on instruction and learning at the University of Göttingen, Germany. His research interests include tutoring, diagnostic expertise of instructors, learning from instructional explanations, learning from worked-out examples, understanding and assessing explanations, and competency-oriented testing. **Matthias Nückles** is a professor of empirical research on instruction and schools at the University of Freiburg, Germany. His research interests pertain to communication between experts and novices, self-regulated learning, writing-to-learn, and teaching skills in secondary and higher education. **Alexander Renkl** is a professor of educational and developmental psychology at the University of Freiburg, Germany. His research investigates example-based learning, instructional explanations and self-explanations, learning from multiple representations, learning by journal writing, and concept mapping as learning method.

REFERENCES

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, *29*, 344–370.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology*, *21*, 177–187.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 21–30). Cambridge, England: Cambridge University Press.

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Chi, M. T. H., & Roy, M. (2010). How adaptive is an expert human tutor? In J. Kay, V. Aleven, & J. Mostow (Eds.), *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems International* (pp. 401–412). Berlin, Germany: Springer-Verlag.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning, *Cognitive Science*, *32*, 301–341.

Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*, 363–387.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471–533.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Corno, L., & Snow, R. E. (1986). Adapting teaching to individual differences among learners. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 605–629). New York, NY: Macmillan.

Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung. Eine experimentelle Studie zu Ankereffekten. [Biases in teachers' assessments of student performance: An experimental study of anchoring effects.]. *Zeitschrift für Pädagogische Psychologie*, *23*, 187–195.

Ericsson, K. A., Charness, N., Feltovich, P., & Hoffman, R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge, England: Cambridge University Press.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research*, *102*, 453–462.

Feldon, D. F. (2007). Cognitive load in the classroom: The double-edged sword of automaticity. *Educational Psychologist*, *42*, 123–137.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, *9*, 495–522.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology*, *5*, 205–221.

Hoge, R., & Coladarci, T. (1989). Teacher-based judgments of academic achievement. *Review of Educational Research*, *59*, 297–313.

Jonassen, D., & Grabowski, B. L. (1993). *Handbook of individual differences, learning and instruction*. Hillsdale, NJ: Erlbaum.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*, 509–539.

Katz, S., Allbritton, D., & Connelly, J. (2003). Going beyond the problem given: How human tutors use postsolution discussions to support transfer. *International Journal of Artificial Intelligence and Education*, *13*, 79–116.

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess. Der Lehrer als flexibler Denker. [Goal-directed processing of students' attributes: The teacher as "flexible thinker".]. *Zeitschrift für Pädagogische Psychologie*, *23*, 175–186.

Lehman, B. A., Matthews, M., D'Mello, S. K., & Person, N. (2008). What are you feeling? Investigating student affective states during expert human tutoring sessions. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems* (pp. 50–59). Berlin, Germany: Springer-Verlag.

Leinhardt, G. (1983). Novice and expert knowledge of individual students' achievement. *Educational Psychologist*, *18*, 165–179.

Lin, J.-W., & Chiu, M.-H. (2010). The mismatch between students' mental models of acids/bases and their sources and their teacher's anticipations thereof. *International Journal of Science Education*, *32*, 1617–1646.

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development and Education*, *52*, 33–43.

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, *14*, 78–102.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction*, *7*, 197–224.

Michael, J. A., Wenderoth, M. P., Modell, H. I., Cliff, W., Horwitz, B., McHale, P., . . . Whitescarver, S. (2002). Undergraduates' understanding of cardiovascular phenomena. *Advances in Physiology Education*, *26*, 72–84.

Mulholland, L. A., & Berliner, D. C. (1992, April). *Teacher experience and the estimation of student achievement*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California.

Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal*, *40*, 905–928.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*, 737–759.

Nückles, M., Winter, A., Wittwer, J., Herbert, M., & Hübner, S. (2006). How do experts adapt their explanations to a layperson's knowledge in asynchronous communication? An experimental study. *User Modeling and User-Adapted Interaction*, *16*, 87–127.

Pelaez, N. J., Boyd, D. D., Rojas, J. B., & Hoover, M. A. (2005). Prevalence of blood circulation misconceptions among prospective elementary teachers. *Advances in Physiology Education*, *29*, 172–181.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal*, *24*, 13–48.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*, 4–14.

Snow, R. E., & Swanson, J. (1992). Instructional psychology: Aptitude, adaptation, and assessment. *Annual Review of Psychology*, *43*, 583–626.

Staub, F., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology*, *93*, 144–155.

Stiggins, R. J., & Conklin, N. F. (1992). *In teacher's hands: Investigating the practices of classroom assessment.* Albany: State University of New York Press.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*, 743–762.

Südkamp, A., & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum: Direkte und indirekte Einschätzungen von Schülerleistungen. [Reference-group-effects in a simulated classroom: Direct and indirect judgments.]. *Zeitschrift für Pädagogische Psychologie*, *23*, 161–174.

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der simulierte Klassenraum. Eine experimentelle Untersuchung zur diagnostischen Kompetenz. [The simulated classroom: An experimental study on diagnostic competence]. *Zeitschrift für Pädagogische Psychologie*, *22*, 261–276.

Sungur, S., & Tekkaya, C. (2003). Students' achievement in human circulatory system unit: The effect of reasoning ability and gender. *Journal of Science Education and Technology*, *12*, 59–64.

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*, *32*, 321–345.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Wittwer, J., Nückles, M., & Renkl, A. (2010). Using a diagnosis-based approach to individualize instructional explanations in computer-mediated communication. *Educational Psychology Review*, *22*, 9–23.

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., & Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education*, *21*, 335–359.

Zimmerman, B. J. (2006). Development and adaption of expertise: The role of self-regulatory processes and beliefs. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 705–722). Cambridge, England: Cambridge University Press.

Article 2:

Addressing Knowledge Deficits in Tutoring and the Role of Teaching Experience: Benefits for Learning and Summative Assessment

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013a). *Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment.* Manuscript submitted for publication.

This article has been submitted to the *Journal of Educational Psychology*.

Addressing Knowledge Deficits in Tutoring and the Role of Teaching Experience: Benefits for Learning and Summative Assessment

Date of submission: March 21, 2013

Abstract

In the course of tutoring, tutors have the opportunity to formatively assess a tutee's understanding. The information gathered by engaging in formative assessment can be used by tutors not only to adapt instruction in order to enhance learning but also to form a summative judgment in order to document a tutee's learning after tutoring. We report about an empirical study with 46 tutor-tutee dyads that examined a tutor's formative assessment in response to a tutee's knowledge deficits. The results showed that formative assessment during tutoring supported learning and improved the accuracy with which tutors summatively assessed a tutee's understanding after tutoring. At the same time, formative assessment was more pronounced in response to knowledge deficits that resulted from a tutor's deliberate elicitation of a tutee's understanding than in response to knowledge deficits that tutees spontaneously expressed on their own initiative. In addition, tutors with teaching experience not only caused tutees to express more knowledge deficits but they also more often engaged in formative assessment in response to knowledge deficits than tutors without teaching experience. This difference also explained why tutors with teaching experience were more accurate than tutors without teaching experience in summatively assessing a tutee's understanding after tutoring. Our findings suggest that the learning potential of knowledge deficits that tutees express largely depends on a tutor's formative assessment. In addition, when tutors engage in formative assessment they are able to form a more accurate picture of what a tutee has learned after tutoring.

Keywords: formative assessment, knowledge deficits, learning, summative assessment, tutoring

Addressing Knowledge Deficits in Tutoring and the Role of Teaching Experience: Benefits for Learning and Summative Assessment

Human tutoring provides tutors with the opportunity to formatively assess a tutee's individual understanding (e.g., Chi & Roy, 2010; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Katz, Allbritton, & Connelly, 2003; Snow & Swanson, 1992). For example, when a tutee answers a question, a tutor can learn what a tutee does and does not know (Chi, Siler, & Jeong, 2004; Cromley & Azevedo, 2005). The information gathered by engaging in such formative assessment can be used by tutors not only to adapt instruction in order to enhance learning but also to summatively assess a tutee's understanding at the end of tutoring (e.g., Black, 1993; Black & Wiliam, 2009).

In this article, we are interested in a tutor's formative assessment in response to a tutee's knowledge deficits. Knowledge deficits such as incomplete beliefs or misconceptions often occur in the course of tutoring and provide a tutor with the opportunity to engage in formative assessment to diagnose a tutee's misunderstandings in more detail (e.g., Chi et al., 2004). We report a study in which we examined whether such formative assessment would improve a tutee's learning and a tutor's summative assessment of a tutee's understanding at the end of tutoring. Given tutors' varying experience in tutoring (e.g., Graesser, D'Mello, & Cade, 2011; VanLehn, 2011), we also studied how the experience of tutors would come into play when assessing a tutee's knowledge deficits in the course of a tutoring session.

Formative and Summative Assessment

The concepts of *formative assessment* and *summative assessment* have been widely used to describe types of assessment procedures that take place in classroom teaching (see, e.g., Black & William, 1998, 2009; Shavelson et al., 2008). Formative assessment refers to an assessment practice that is undertaken in the course of teaching with the aim to gather information that can be used to adapt instruction and, thus, to support learning (e.g., Black & Willliam, 1998; Shepard, Hammerness, Darling-Hammond, & Rust, 2005). Therefore,

formative assessment is often called *assessment for learning* (Birenbaum et al., 2006; Shepard, 2005; Stiggins, 2006). In contrast, the primarily goal associated with summative assessment is not to improve learning but to document a person's learning at the end of instruction (e.g., Shavelson et al., 2008). Thus, summative assessment is frequently termed *assessment of learning* (Birenbaum et al., 2006). Formative assessment and summative assessment complement each other because the information gathered by engaging in formative assessment can be used to summatively assess a person's learning (e.g., Birenbaum et al., 2006; Black, 1993). Conversely, summative assessment can be used to inform subsequent instruction (for more details, see Black & Wiliam, 2009). For example, the results of summative assessment might help to prepare the next teaching unit by selecting learning material that is specifically suited to a learner's assessed level of understanding (e.g., Perie, Marion, & Gong, 2009).

Formative Assessment and Learning

Research has shown that formative assessment benefits learning. For example, Wiliam, Lee, Harrison, and Black (2004) found that students learned more when teachers integrated procedures of formative assessment into their classroom teaching. Similarly, Furtak et al. (2008) demonstrated that the extent to which a teacher engaged in formative assessment during classroom teaching was positively related to learning. Likewise, in the context of tutoring, Bloom (1984) showed that tutoring in which formative tests were embedded was more effective than traditional classroom teaching. Despite the benefits of formative assessment for learning, it is not undisputed what counts as formative assessment (see, e.g., Black & Wiliam, 2009). For example, Shavelson et al. (2008) proposed three types of formative assessment, namely, (1) on-the-fly formative assessment where an instructor unintentionally receives information about a learner's understanding, (2) planned-for-interaction formative assessment where an instructor, for example, asks a question to deliberately assess a learner's understanding, and (3) embedded-in-the-curriculum formative

assessment where the formative assessment procedures are an integral part of the curriculum. In addition, Black and Wiliam (2009) argued that there are some distinctive activities in formative assessment such as eliciting a learner's understanding, providing feedback, and activating a learner (see also Birenbaum et al., 2006; Shepard, 2005).

Formative Assessment as a Basis for Summative Assessment

Apart from supporting the learning process on a moment-to-moment basis, formative assessment can be used to summatively assess a learner's understanding (e.g., Black, 1993). For example, when evaluating a learner's understanding at the end of a learning unit, all information gathered in the course of instruction by engaging in formative assessment can be collected to form a summative judgment (e.g., Perie et al., 2009). However, very little attention has been paid to whether formative assessment is a reliable method to inform summative assessment. Thus, it is not clear whether instructors are able to aggregate the information resulting from engaging in formative assessment into a summative judgment. Research in the field of classroom teaching has mainly focused on the accuracy with which teachers summatively judge a learner's academic achievement as displayed in performance tests (for overviews, see Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012). In contrast, research on human tutoring is usually interested in a tutor's activities that are responsible for learning (e.g., Graesser et al., 2011). Therefore, this research often focuses on activities associated with formative assessment such as giving feedback or providing scaffolding. However, learning measures are rarely collected in research on tutoring. Therefore, it is often not possible to study a tutor's summative assessment (for an exception, see, e.g., Chi et al., 2001; Herppich, Wittwer, Nückles, & Renkl, 2013).

Formative Assessment of a Learner's Understanding in Tutoring: The Case of Knowledge Deficits

Human tutoring provides a tutor with the opportunity to formatively assess a tutee's understanding on a moment-to-moment basis (Graesser et al., 2011; Graesser, Person, &

Magliano, 1995). In the context of this study, we are interested in the way a tutor formatively assesses knowledge deficits that a tutee expresses in the course of tutoring. According to Chi et al. (2004), knowledge deficits refer, for example, to contradictory beliefs (e.g., a tutee assumes: "Blood goes to the various body parts after it leaves the right ventricle." but a passage from a textbook says: "Blood goes to the lungs after it leaves the right ventricle."; Chi et al., 2004, p. 366) or incomplete beliefs (e.g., a tutee says "a valve" but the correct answer would be "a semilunar valve"; Chi et al., 2004, p. 379). Such expressed knowledge deficits are diagnostically informative because they indicate what a tutee does not know (Chi et al., 2004; Graesser et al., 1995). Following Chi et al. (2004), knowledge deficits can be further differentiated as either *tutee-initiated* or *tutor-initiated*. A tutee-initiated knowledge deficit occurs when a tutee expresses a knowledge deficit while, for example, asking a question to the tutor. In contrast, a tutor-initiated knowledge deficit occurs when a tutee expresses a knowledge deficit, for example, in response to a tutor's question. In line with the typology proposed by Shavelson et al. (2008), tutee-initiated knowledge deficits form the basis for on-the-fly formative assessment because such knowledge deficits occur unexpectedly for tutors. Tutor-initiated knowledge deficits, in contrast, are part of planned-for-interaction formative assessment because tutors deliberately ask questions to elicit a tutee's understanding including knowledge deficits.

When a tutee expresses a knowledge deficit, regardless of whether it is tutee-initiated or tutor-initiated, it is important to consider how a tutor addresses a tutee's knowledge deficit (Black & Wiliam, 2009; Chi et al., 2004). At one extreme, a tutor might ask questions in response to a tutee's knowledge deficit in order to gain more information about a tutee's understanding. At the other extreme, a tutor might provide a correct answer without addressing a tutee's understanding in more detail. Research on human tutoring has identified mainly three different types of responses to knowledge deficits, namely, providing negative feedback, scaffolding, and giving a correct answer (Chi et al., 2004; Graesser et al., 1995). All

three types of responses address a tutee's knowledge deficit. However, whereas providing negative feedback and scaffolding in response to knowledge deficits can be regarded as formative assessment procedures because they elicit answers from tutees, giving a correct answer in response to knowledge deficits fails to provide a tutor with more information about a tutee's understanding (Black & Wiliam, 2009).

Providing Negative Feedback

Feedback is normally used to comment on a tutee's contribution (Cromley & Azevedo, 2005; VanLehn, 2011). In particular, negative feedback is assumed to benefit learning because it provides a tutee with the information what has not been understood yet (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995; Merrill, Reiser, Ranney, & Trafton, 1992). From a tutee's response to negative feedback, a tutor can derive information that can be used to assess a tutee's level of understanding.

Providing Scaffolding

Scaffolding refers to a tutor's moves, such as asking questions or providing hints, that are undertaken to support a tutee in proceeding further in a line of reasoning or in a task that a tutee would be not able to accomplish alone (Chi et al., 2001; Van de Pol et al., 2010). Hence, scaffolding serves the function to interactively guide a tutee to a more complete and correct understanding (Chi, 2009; Graesser et al., 1995, 2011). When tutors provide scaffolding, they receive information that can be used for assessing a tutee's understanding.

Giving a Correct Answer

Giving a correct answer is an example of an activity that is undertaken by a tutor to provide a tutee with content-related information (D'Mello, Lehman, & Person, 2010; Graesser et al., 1995). However, in many instances, correct answers are not fully adapted to a tutee's specific understanding (Chi et al., 2001; Wittwer & Renkl, 2008, Wittwer, Nückles, Landmann, & Renkl, 2010). Therefore, correct answers often do not enhance learning. In addition, giving a correct answer is per se a non-interactive activity because as long as a tutor

talks a tutee usually does not talk (e.g., VanLehn et al., 2003). Therefore, this activity usually fails to provide a tutor with the opportunity to receive further information from a tutee. Hence, even though giving a correct answer indicates that a tutor obviously recognizes a tutee's knowledge deficit, it does not necessarily help a tutor to further assess a tutee's individual understanding.

Evidence for Formative Assessment in Response to Knowledge Deficits

There are only a few studies that examined a tutor's formative assessment in response to a tutee's knowledge deficits. Chi et al. (2004) examined exclusively tutee-initiated knowledge deficits and assumed that when tutors provided negative feedback, scaffolded a tutee, or gave a correct answer, they unambiguously addressed a tutee's knowledge deficit. Moreover, when a tutor vaguely signaled the detection of a knowledge deficit but did not unambiguously address and repair it, the tutor was assumed to "attempt to repair" the knowledge deficit (Chi et al., 2004, p. 382). In addition, when a tutor continued the discussion without referring to a knowledge deficit, acknowledged it as being correct or repeated a knowledge deficit, it was assumed that a tutor accepted the knowledge deficit. It was found that the tutors unambiguously addressed less than half of all tutee-initiated knowledge deficits and accepted about one third. Hence, the tutee-initiated knowledge deficits were not always used by tutors to engage in formative assessment. In contrast to the results reported by Chi et al. (2004), Cromley and Azevedo (2005) observed in their study that tutors addressed a tutee's knowledge deficits in the great majority of cases. The tutors most often provided scaffolding, gave negative feedback, or presented a correct answer. Similarly, Graesser et al. (1995) showed that tutors frequently provided a correct answer to a tutee's knowledge deficit. Moreover, the tutors provided hints, gave lengthy explanations, provided direct negative feedback, or asked a question to direct a tutee towards the correct answer.

Overall, the reported studies (Chi et al., 2004; Cromley & Azevedo, 2005; Graesser et al., 1995) suggest that a tutor frequently reacts to a tutee's knowledge deficits with providing

scaffolding, giving negative feedback, and conveying a correct answer. Of these reactions, only scaffolding and giving negative feedback count as formative assessment procedures, whereas giving a correct answer does not provide a tutor with further information about a tutee's understanding. Although the reported studies (Chi et al., 2004; Cromley & Azevedo, 2005; Graesser et al., 1995) provide important insights into how tutors formatively assess a tutee's knowledge deficits, several questions remain unanswered.

First, no study has examined whether formative assessment in response to knowledge deficits in the course of tutoring supports a tutee's learning and a tutor's summative assessment of a tutee's understanding at the end of tutoring. This question is, however, important because knowledge deficits are regarded as a central learning opportunity for tutees (Chi et al., 2004; Graesser et al., 1995). In addition, when the information collected by engaging in formative assessment is not useful for summatively assessing a tutee's understanding, assessment procedures other than a tutor's assessment such as standardized tests would be necessary to document a tutee's level of understanding at the end of a learning unit (e.g., Harlen & James, 1997).

Second, it is not clear how different types of formative assessment are used in human tutoring. More specifically, it is plausible to assume that formative assessment in response to tutee-initiated knowledge deficits (on-the-fly formative assessment; Shavelson et al., 2008) is designed differently than formative assessment in response to tutor-initiated knowledge deficits (planned-for-interaction formative assessment; Shavelson et al., 2008). For example, Shavelson (2006) argues that, due the unexpectedness of getting insight into a learner's understanding in on-the-fly formative assessment, instructors might not know how to respond to the learner appropriately. In contrast, when formative assessment is planned and knowledge deficits, for example, are deliberately elicited, instructors might be better equipped with responding to such knowledge deficits. Hence, it might be that tutors differ in the way they

respond to knowledge deficits depending on whether such knowledge deficits are tutee-initiated or tutor-initiated.

Third, it remains open as to whether the way a tutor formatively assesses a tutee's knowledge deficits depends on a tutor's experience. Persons such as university students or parents who possess relevant knowledge but who are not trained in teaching (i.e., *normal* tutors) often serve as tutors (Chi et al., 2001; Graesser et al., 2011; VanLehn, 2011). Sometimes, however, tutors are classroom teachers or graduate students who are more experienced because they are trained in tutoring or possess teaching experience (e.g., Chi, Roy, & Hausmann, 2008; McArthur, Stasz, & Zmuidzinas, 1990; Putnam, 1987). Research shows that inexperienced tutors have the propensity to provide information during tutoring (e.g., Chi et al., 2001), whereas experienced tutors more often engage in scaffolding (Cade, Copeland, Person, & D'Mello, 2008; Chae, Kim, & Glass, 2005; Chi et al., 2008; Cromley & Azevedo, 2005). In light of this finding, it can be assumed that, depending on their experience, tutors also differ in the way they formatively assess a tutee's knowledge deficits. More concretely, when a tutee's expresses a knowledge deficit, experienced tutors might be more likely to provide scaffolding whereas inexperienced tutors might more often respond with providing information.

Present Study and Hypotheses

We examined the extent to which tutors engaged in formative assessment in response to tutees' knowledge deficits. In particular, we were interested in whether a tutor's formative assessment would benefit a tutee's learning and a tutor's summative assessment of a tutee's understanding at the end of tutoring. The analyses reported in this study are based on protocol data that were collected but not reported in the study by Herppich et al. (2013), who examined tutoring in the domain of the human circulatory system. The study by Herppich et al. (2013) already showed that experienced tutors, that is, classroom teachers (i.e., teacher tutors) summatively assessed a tutee's conceptual understanding about the human circulatory system

more accurately than did inexperienced tutors, that is, university students (i.e., student tutors). In the present study, we addressed the following hypotheses:

- Learning-Effect hypothesis: Tutors who more often engage in formative assessment in response to knowledge deficits more strongly support a tutee's learning than tutors who less often engage in formative assessment in response to knowledge deficits.
- Assessment-Effect hypothesis: Tutors summatively assess a tutee's understanding at the end of tutoring more accurately when they more often engage in formative assessment in response to knowledge deficits during tutoring.
- Knowledge-Deficit-Type hypothesis: Tutors more often engage in formative assessment in response to tutor-initiated knowledge deficits than in response to tutee-initiated knowledge deficits.
- Tutor-Type-Knowledge-Deficit hypothesis: Teacher tutors more often cause tutees to produce knowledge deficits than student teachers.
- Tutor-Type-Formative-Assessment hypothesis: Teacher tutors more often engage in formative assessment in response to knowledge deficits than student teachers.
- Tutor-Type-Learning-Effect hypothesis: Teacher tutors support a tutee's learning more strongly than student tutors. This effect can be explained by a difference in the extent to which teacher tutors and student tutors engage in formative assessment in response to knowledge deficits.
- Tutor-Type-Assessment-Effect hypothesis: That teacher tutors are more accurate than student tutors in summatively assessing a tutee's understanding at the end of tutoring (see Herppich et al., 2013) is attributable to a difference in the extent to which teacher tutors and student tutors engage in formative assessment in response to knowledge deficits.

Method

The materials and the procedure applied in this study were thoroughly described in Herppich et al. (2013). Please refer to this article for more detailed information.

Sample and Design

A total of *N* = 46 tutor-tutee dyads participated in the study (Herppich et al., 2013). As *teacher tutors,* we had 21 biology teachers with a mean age of 44.05 years (*SD* = 11.76). Of them, 11 tutors were female and 10 tutors were male. As *student tutors,* we had 25 university students majoring in biology with a mean age of 22.24 years (*SD* = 2.83). Of them, 21 tutors were female and 4 tutors were male. The tutees were seventh-grade students. Of them, 19 were female and 27 were male. Their mean age was 12.65 years (*SD* = 0.53). The tutees were randomly assigned to one of the two tutor groups. Therefore, tutors and tutees did not know each other before tutoring.

The dependent variables were (1) the number of knowledge deficits that a tutee expressed during tutoring, (2) a tutor's formative assessment operationalized as the ratio of a tutor's interactive responses to knowledge deficits to a tutor's non-interactive responses to knowledge deficits during tutoring, (3) a tutor's item-by-item summative assessment accuracy at the end of tutoring (*item responses* sensu Hoge & Coladarci, 1989), and (4) a tutee's learning gain at the end of tutoring.

Materials

Textbook passage (tutees and tutors). In the tutoring session, the dyads of tutor and the tutee engaged in a dialogue on the basis of a passage about the human circulatory system.

Concepts test (tutees and tutors). The test consisted of 25 multiple-choice items that assessed a tutee's understanding of concepts related to the human circulatory system (for an example, see Figure 1). A correct answer to an item indicated a correct understanding. Each of the incorrect answers to an item indicated an incorrect understanding. The tutee was administered the test at the beginning and at the end of tutoring. Each item that a tutee answered correctly in the concepts test was assigned 1 point. Hence, a tutee could achieve a maximum number of 25 points.

To examine the accuracy with which the tutors assessed a tutee's knowledge of the human circulatory system, the tutors also received the concepts test at the end of tutoring but they were asked to indicate how their tutee would answer each of the 25 items.

Procedure

Each tutoring session was divided into three phases: pretest phase, tutoring phase, and posttest phase. On average, a tutoring session lasted about 3 hours. In the pretest phase, the tutees completed the concepts test. In addition, each tutee and each tutor individually read the passage about the human circulatory system. In the tutoring phase, tutor-tutee dyads jointly read the passage about the human circulatory system sentence-by-sentence and engaged in a dialogue about each sentence. All tutoring phases were videotaped. In the posttest phase, the tutees completed the concepts test once again. The tutors also received the items of the concepts test and were asked to indicate how the tutee would answer each of the items of the concepts test.

Coding

To code a tutee's knowledge deficits and a tutor's responses, we used a coding scheme adapted from Chi et al. (2004, for original coding scheme).

Tutees' knowledge deficits. The recordings of the tutoring sessions were coded for a tutee's knowledge deficits by using an event sampling procedure. An event was coded as a knowledge deficit when the tutee uttered a belief that (1) contradicted a piece of knowledge explicitly or implicitly stated in the textbook passage, that (2) was incomplete (e.g., *Oxygen is what we breathe in* as an answer to a tutor's question *What kind of substances are oxygen and carbon dioxide?*), that (3) was vague (e.g., *So we can live* as an answer to a tutor's question *Why do we need oxygen?*), that (4) was incorrect and not explicitly or implicitly addressed in the content of the textbook passage, that is, a false belief, or when the tutee (5) did not utter a certain piece of information at all, that is, the tutee obviously missed this piece of information. The coding differentiated between knowledge deficits initiated by the tutor (i.e., tutor-

initiated) that followed a tutor's question or prompt and knowledge deficits initiated by the tutee (i.e., tutee-initiated), which were unprompted questions, remarks, or self-explanations. To standardize coding, the coder used a written instruction. To test for the reliability of the coding scheme, a second coder independently coded tutees' knowledge deficits and tutors' responses for 10 dyads (22 %). We calculated interrater agreement at the level of coding intervals. Cohen's kappa (Cohen, 1960) for coding a tutor-initiated knowledge deficit (κ = .85) and for coding a tutee-initiated knowledge deficit (κ = .74) was good (Fleiss & Cohen, 1973).

Formative assessment: Tutors' responses to tutees' knowledge deficits. The recordings of the tutoring sessions were also coded for a tutor's responses to a tutee's knowledge deficits. Generally, a response was defined as a tutor's first reaction to a knowledge deficit uttered by a tutee. The following types of responses were coded: (1) direct feedback, that is, a tutor gave a short negative feedback that could include a short correction of a tutee's utterance, like *No* or *Like that it's not correct* or *No, it's the other way round*, (2) scaffolding, that is, a tutor gave a hint or prompt (cf. Chi et al., 2001; Van de Pol et al., 2010) that pushed the tutee towards discovering the correct information on her or his own, (3) correct answer, that is, a tutor comprehensively supplied the correct piece of information (4) attempt to repair, that is, a tutor's response indicated that the knowledge deficit was detected but the tutor did not try to correct it (e.g., a tutor acknowledged that the tutee could not know a piece of information or the tutor instructed the tutee to stop talking and concentrated on the actual content of the text passage again), (5) ignoring, that is, the tutor did neither obviously detect nor correct a tutee's knowledge deficit (e.g., a tutor gave positive feedback or repeated a tutee's utterance verbatim; termed *accept* in the coding scheme by Chi et al., 2004). Due to the low incidence of recognition without repair and ignoring (in total 9 %), we confined our analyses to a tutor's responses that comprised (1) direct feedback, (2) scaffolding, and (3) correct answers.

The following episodes from a videotaped tutoring session provide examples of the different types of a tutor's responses to a tutee's knowledge deficits.

TUTOR:	Okay. And, now, what's a circuit?	
TUTEE:	A circuit is… {taps with one hand on the other; grimaces} oh, I don't know it. Circuit… Circuit…	[Tutor-initiated]
TUTOR:	Can you imagine if something goes in circles?	[Scaffolding]
TUTEE:	Yes.	

TUTOR:	Why does the blood need to go to the lung? What does the lung do? …{pauses} Why is the blood supposed to be there?	
TUTEE:	Yes, um, yes, the lung filters the blood doesn't it?	[Tutor-initiated]
TUTOR:	No.	[Direct Feedback]

TUTEE:	[Reads a sentence] 'These arteries first branch out to arterioles and then branch out to capillaries.' Arterioles were the very small ones. And capillaries…{pauses}?	[Tutee-initiated]
TUTOR:	And the capillaries are the even smaller ones.	[Correct Answer]

To test our hypotheses, we proceeded in the following way: First, we summed up all direct feedback responses, scaffolding responses, and correct answer responses to a tutee's knowledge deficits separately for each tutor and separately for tutor-initiated knowledge deficits and tutee-initiated knowledge deficits, respectively. Second, we computed a combined measure of formative assessment by adding the number of direct feedback responses to the

number of scaffolding responses for each tutor. Third, we followed the suggestion by Chi et al. (2008) and divided the resulting sum by the number of correct answer responses for each tutor in order to calculate the ratio of interactive responses to non-interactive responses. A higher ratio indicated a larger amount of formative assessment in response to a tutee's knowledge deficits. The interrater reliability for interactive responses to tutor-initiated knowledge deficits was $\kappa = .80$ (Cohen, 1960). For interactive responses to tutee-initiated knowledge deficits, the interrater reliability was $\kappa = .25$. For non-interactive responses to tutor-initiated knowledge deficits, the interrater reliability was $\kappa = .63$. Finally, for non-interactive responses to tutee-initiated knowledge deficits, the interrater reliability was $\kappa = .74$. With the exception of the interrater reliability for interactive responses to tutee-initiated knowledge deficits, the interrater reliability for all responses was good (Fleiss & Cohen, 1973). The low value of $\kappa = .25$, however, is not produced by a low interrater agreement per se but due to the scarcity of interactive responses to tutee-initiated knowledge deficits (Wirtz & Caspar, 2002; cf. Table 2).

Summative assessment: Tutors' assessment accuracy. To measure the accuracy with which a tutor summatively assessed a tutee's conceptual understanding of the human circulatory system at the end of tutoring, we compared a tutee's answers in the concepts test administered in the posttest phase with a tutor's estimations of the tutee's answers in this concepts test on an item-by-item basis (called *item responses* sensu Hoge & Coladarci, 1989). To do so, we assigned a tutor 1 point for every correct prediction. Thus, if a tutor estimated that the tutee would choose response, for example, option 2 and a tutee indeed chose response option 2, the tutor was assigned 1 point. The concepts test comprised 25 items. Thus, a tutor could achieve a maximum score of 25 points. The number of points was used as indicator of a tutor's assessment accuracy.

Tutees' learning gain. We measured a tutee's learning gain by subtracting the pretest score in the concepts test from the posttest score in the concepts test. For example, a tutee

who had 4 points in the pretest and 18 points in the posttest achieved a learning gain of 14 points.

Results

For all analyses, we used an alpha level of .05. In the cases of directional hypotheses, we used one-tailed tests. Depending on the statistical analysis, we employed differing effect sizes. We report η^2 as effect-size measure for ANOVAs (Cohen, 1988), φ as effect-size measure for Fisher's exact test (Cohen, 1988), R^2 and standardized regression coefficients (β) as effect-size measures for simple linear OLS regressions (Cohen, 1988), and κ^2 as effect-size measure for indirect effects in simple mediation analyses (Preacher & Kelley, 2011).

Note that not in all tutor-tutee dyads a tutee expressed knowledge deficits and not in all tutor-tutee dyads a tutor responded interactively *and* non-interactively to a tutee's knowledge deficits. Therefore, the analysis of a tutor's formative assessment in response to a tutee's knowledge deficit can only apply to those dyads in which (1) a tutee expressed at least one knowledge deficit and (2) in which a tutor responded at least once non-interactively to a tutee's knowledge deficit. Given these constraints, the analysis of formative assessment in response to tutor-initiated knowledge deficits is based on a sample with $N = 42$ tutor-tutee dyads. In the analysis of formative assessment in response to tutee-initiated knowledge deficits, the sample size is $N = 38$ tutor-tutee dyads. The analysis of formative assessment in response to both types of knowledge deficits is restricted to $N = 35$ tutor-tutee-dyads. All analyses were performed using Excel 2010, SPSS 20.0.0, AMOS 20.0.0, and the PROCESS macro for SPSS introduced by Hayes (2012).

Learning-Effect Hypothesis

The learning-effect hypothesis predicted that a larger amount of formative assessment in response to knowledge deficits benefits learning more than does a smaller amount of formative assessment in response to knowledge deficits. To test this hypothesis, we performed two simple linear regression analyses with a tutee's learning gain as the criterion and the

amount of formative assessment in response to tutor-initiated knowledge deficits and in response to tutee-initiated knowledge deficits, respectively, as the predictor (see Table 1 and Table 2 for means and standard deviations). The results showed that a larger amount of formative assessment positively and significantly accounted for a tutee's learning gain when formative assessment was undertaken in response to tutor-initiated knowledge deficits, R^2 = .07 (medium effect), $F(1, 40) = 3.06$, $\beta = .27$, $p < .05$, 95% CI [.01, .52], and in response to tutee-initiated knowledge deficits, R^2 = .09 (medium effect), $F(1, 36) = 3.61$, $\beta = .30$, $p < .05$, 95% CI [.03, .58]. Thus, tutors who more often engaged in formative assessment in response to knowledge deficits more strongly supported a tutee's learning.

Assessment-Effect Hypothesis

According to the assessment-effect hypothesis, tutors should summatively assess a tutee's understanding at the end of tutoring more accurately when they more often engage in formative assessment in response to knowledge deficits in the course of tutoring. To test this hypothesis, we performed two simple linear regression analyses with the amount of formative assessment in response to a tutee's knowledge deficits as predictor and with the accuracy of a tutor's assessment at the end of tutoring as the criterion (see Table 1 and Table 2 for means and standard deviations). We found that a larger amount of formative assessment positively and significantly predicted more accurate assessments when tutors engaged in formative assessment in response to tutor-initiated knowledge deficits, R^2 = .10 (medium effect), $F(1, 40) = 4.42$, $\beta = .32$, $p < .05$, 95% CI [.06, .57], and in response to tutee-initiated knowledge deficits, R^2 = .31 (large effect), $F(1, 36) = 16.40$, $\beta = .56$, $p < .05$, 95% CI [.33, .79]. Hence, tutors were more accurate at assessing a tutee's understanding at the end of tutoring when they more often engaged in formative assessment in response to a tutee's knowledge deficits in the course of tutoring.

Knowledge-Deficit-Type Hypothesis

The knowledge-deficit-type hypothesis stated that tutors more often respond with formative assessment to tutor-initiated knowledge deficits than to tutee-initiated knowledge deficits. To test this hypothesis, we computed a repeated-measures analysis with the amount of formative assessment in response to tutor-initiated knowledge deficits and the amount of formative assessment in response to tutee-initiated knowledge deficits as levels of the repeated-measures factor. In the reduced sample of $N = 35$ dyads with valid responses to tutor-initiated knowledge deficits and to tutee-initiated knowledge deficits, a tutor's amount of formative assessment in response to tutor-initiated knowledge deficits was, on average, 1.34 ($SD = 1.26$) and a tutor's amount of formative assessment in response to tutee-initiated knowledge deficits was, on average, 0.22 ($SD = 0.35$). The correlation between formative assessment in response to tutor-initiated knowledge deficits and in response to tutee-initiated knowledge deficits was $r = .35$, $p < .05$. Thus, tutors who more often engaged in formative assessment in response to tutor-initiated knowledge deficits also more often engaged in formative assessment in response to tutee-initiated knowledge deficits. The repeated-measures analysis showed that tutor-initiated knowledge deficits were indeed followed by a larger amount of formative assessment than tutee-initiated knowledge-deficits, $F(1, 34) = 31.38$, $p < .05$, $\eta^2 = .48$ (large effect). Thus, tutors responded to tutor-initiated knowledge deficits more often with formative assessment than to tutee-initiated knowledge deficits.

Tutor-Type-Knowledge-Deficit Hypothesis

The tutor-type-knowledge-deficit hypothesis predicted that teacher tutors cause their tutees more often to produce knowledge deficits than do student tutors. In a first step, we counted the number of dyads in which a tutee did not express a knowledge deficit at all. Of the $n = 21$ dyads with teacher tutors, there was 1 dyad in which the tutee did not express a tutor-initiated knowledge deficit and there was also 1 dyad in which the tutee did not express a tutee-initiated knowledge deficit. Of the $n = 25$ dyads with student tutors, there were 3 dyads in which the tutee did not express a tutor-initiated knowledge deficit and there were 6

dyads in which the tutee did not express a tutee-initiated knowledge deficit. Two Fisher's exact tests showed that there were more dyads with student tutors whose tutees did not express a knowledge deficit than dyads with teacher tutors whose tutees did not express a knowledge deficit. This difference, however, failed to reach the level of statistical significance for tutor-initiated knowledge deficits, $p = .37$, $\varphi = .13$ (small effect), and for tutee-initiated knowledge deficits, $p = .08$, $\varphi = .27$ (medium effect).

In a second step, we performed two one-way ANOVAs including the type of tutor (i.e., teacher tutor vs. student tutor) as the independent variable and the number of tutor-initiated knowledge deficits and the number of tutee-initiated knowledge deficits, respectively, as dependent variable. Overall, the $n = 21$ teacher tutors initiated, on average, 40.90 ($SD = 27.16$) knowledge deficits and their tutees, on average, expressed 10.81 ($SD = 12.75$) tutee-initiated knowledge deficits. The $n = 25$ student tutors initiated, on average, 22.68 ($SD = 25.33$) knowledge deficits and their tutees, on average, expressed 4.36 ($SD = 5.84$) tutee-initiated knowledge deficits. As predicted, teacher tutors elicited significantly more tutor-initiated knowledge deficits than did student tutors, $F(1, 44) = 5.53$, $p < .05$, $\eta^2 = .11$ (medium effect). Similarly, teacher tutors also made their tutees express significantly more tutee-initiated knowledge deficits than did student tutors, $F(1, 44) = 5.13$, $p < .05$, $\eta^2 = .07$ (medium effect). Thus, teacher tutors caused their tutees to produce more knowledge deficits than did student tutors.

Tutor-Type-Formative-Assessment Hypothesis

The tutor-type-formative-assessment hypothesis stated that teacher tutors more often engage in formative assessment in response to knowledge deficits than student tutors. We tested this hypothesis with two one-way ANOVAs. The ANOVAs included the type of tutor as the independent variable and the amount of formative assessment in response to tutor-initiated knowledge deficits and to tutee-initiated knowledge deficits, respectively, as dependent variable (see Table 1 and Table 2 for means and standard deviations). The results

showed that teacher tutors more often engaged in formative assessment in response to tutor-initiated knowledge deficits, $F(1, 40) = 7.09$, $p < .05$, $\eta^2 = .15$ (large effect), and to tutee-initiated knowledge-deficits, $F(1, 36) = 3.93$, $p < .05$, $\eta^2 = .10$ (medium effect), than did student tutors.

Tutor-Type-Learning-Effect Hypothesis

According to the tutor-type-learning-effect hypothesis, tutees of teacher tutors should achieve higher learning gains than tutees of student tutors. Moreover, the fact that teacher tutors more often engaged in formative assessment in response to knowledge deficits than did student tutors should explain why tutees of teacher tutors achieved higher learning gains than tutees of student tutors. To test this mediation hypothesis, we calculated total, direct, and indirect effects applying OLS regression-based path analysis. In this analysis, teacher tutors were coded as 1 and student tutors were coded as 0. To test the indirect effect, we applied the bootstrapping procedure suggested by Preacher and Hayes (2008; see also Hayes, 2009, 2012). We used 10.000 resamples with replacement to derive a 95% bias-corrected confidence interval for the indirect effect of the type of tutor on a tutee's learning gain via the amount of a tutor's formative assessment. We performed two separate analyses for tutor-initiated knowledge deficits and tutee-initiated knowledge deficits, respectively. These mediation analyses are illustrated in Figure 2 (see Table 1 and Table 2 for means, standard deviations and correlations). For tutor-initiated knowledge deficits we, first, found a significant total effect of the type of tutor on a tutee's learning gain (cf. Figure 2 left) indicating that tutees of teacher tutors achieved significantly larger learning gains than tutees of student tutors. Second, we found a standardized indirect effect of .07 ($\kappa^2 = .07$, small effect) with a standardized 95% confidence interval ranging from -.02 to .21. As the interval includes zero, the indirect effect was not significant (Preacher & Hayes, 2008). For the tutee-initiated knowledge deficits, the analysis also revealed a significant total effect of the type of tutor on a tutee's learning gain (cf. Figure 2 right) indicating that tutees of teacher tutors

achieved significantly larger learning gains than tutees of student tutors. Moreover, there was a standardized indirect effect of .07 (κ^2 = .07, small effect) with a standardized 95% confidence interval ranging from -.01 to .18. Thus, the indirect effect was again not significant. To sum up, teacher tutors supported learning of their tutees more strongly than did student tutors. Yet, this effect is not explained by the difference in the extent to which teacher tutors and student tutors engaged in formative assessment in response to knowledge deficits.

Tutor-Type-Assessment-Effect Hypothesis

The tutor-type-assessment-effect hypothesis stated that the difference in the extent to which teacher tutors and student tutors engaged in formative assessment explains why teacher tutors were more accurate than student tutors in summatively assessing a tutees' understanding at the end of tutoring (see Herppich et al., 2013). To test this mediation hypothesis, we again calculated total, direct, and indirect effects applying OLS regression-based path analysis as described before. We tested the indirect effect of the type of tutor on a tutor's accuracy of summative assessment via the amount of a tutor's formative assessment with the bootstrapping procedure and used 10.000 resamples with replacement to derive a 95% bias-corrected confidence interval for the indirect effect. We performed two separate analyses for tutor-initiated knowledge deficits and tutee-initiated knowledge deficits, respectively. The mediation analyses are illustrated in Figure 3 (see Table 1 and Table 2 for means, standard deviations, and correlations). For tutor-initiated knowledge deficits, we found a standardized indirect effect of .11 (κ^2 = .11, medium effect) with a standardized 95% confidence interval ranging from .03 to .23. Because the interval does not include zero, the indirect effect was significant (Preacher & Hayes, 2008). For tutee-initiated knowledge deficits, the analysis revealed a standardized indirect effect of .16 (κ^2 = .17, medium effect) with a standardized 95% confidence interval ranging from .06 to .30. Again, the indirect effect was statistically significant. Hence, the fact that teacher tutors were more accurate than student tutors in summatively assessing a tutee's understanding after tutoring could be

explained by the difference in the extent to which teacher tutors and student tutors engaged in formative assessment.

Discussion

We examined whether engaging in formative assessment in response to a tutee's knowledge deficits supported a tutee's learning and improved the accuracy with which tutors assessed a tutee's understanding at the end of tutoring. In addition, we were interested in whether the teaching experience of tutors made a difference in formative assessment.

First, we found that a larger amount of formative assessment enhanced a tutee's learning. This result is in line with previous studies that demonstrate the benefits of formative assessment for learning (e.g., Bloom, 1984; Furtak et al., 2008; Wiliam et al., 2004). However, our results extend prior research because we examined formative assessment in a tutoring setting where we did not train the tutors in engaging in formative assessment, unlike other studies (e.g., Wiliam et al., 2004). In addition, we showed that benefits for learning materialized regardless of whether tutors intensified formative assessment in response to tutee-initiated knowledge deficits or in response to tutor-initiated knowledge deficits. Given that tutee-initiated knowledge deficits, in contrast to tutor-initiated knowledge deficits, occurred unexpectedly for tutors (on-the-fly formative assessment; Shavelson et al., 2008), it is remarkable that formative assessment supported learning even under these circumstances.

Second, this study showed that formative assessment reliably informed summative assessment at the end of tutoring. In other words, when tutors engaged in formative assessment in the course of tutoring they had a more accurate picture of what has been learned by a tutee after tutoring. In research on educational assessment, a central question is how information collected by engaging in formative assessment can be aggregated for summative purposes (e.g., Perie et al., 2009). Admittedly, our findings do not uncover the processes involved in aggregating the information gathered by formative assessment in order to form a summative judgment. However, this study suggests that engaging in formative assessment

increases the accuracy of summative assessments. Clearly, further research is needed to examine in more detail the inferential processes that occur when information gained by formative assessment is used to summatively assess a learner's understanding.

Third, the results yielded that the amount of formative assessment that tutors engaged in depended on the type of knowledge deficits that a tutee expressed. More concretely, when tutees expressed knowledge deficits on their own initiative, that is, when they uttered tutee-initiated knowledge deficits, tutors rarely engaged in formative assessment. Conversely, tutor-initiated knowledge deficits that were elicited by a tutor were more often followed by formative assessment. In line with the reasoning by Shavelson (2006), it can be assumed that it is fairly difficult to formatively assess a tutee-initiated knowledge deficit because it occurs unexpectedly for a tutor. In contrast, when tutors actively elicit a tutee's understanding, they might expect that a knowledge deficit is likely to occur and, thus, are better prepared to respond to such a knowledge deficit.

Fourth, we studied formative assessment as a function of a tutor's teaching experience. We found that teacher tutors caused tutees to express more knowledge deficits than did student tutors. This was true both for tutor-initiated knowledge deficits and tutee-initiated knowledge deficits. The finding suggests that, in contrast to student tutors, teacher tutors provided a more interactive style of tutoring, which helped to elicit a tutee's knowledge deficits. At the same time, we observed that teacher tutors not only let tutees express more knowledge deficits but they also more frequently responded to such knowledge deficits with formative assessment than did student tutors. This result is in line with previous research showing that more experienced tutors more often provide scaffolding whereas less experienced tutors more often tend to give a correct answer (e.g., Cade et al., 2008; Chae et al., 2005; Chi et al., 2008; Cromley & Azevedo, 2005).

Fifth, the results revealed that tutees learned more when being tutored by teacher tutors than being tutored by student tutors. This finding is not trivial because there is not

strong evidence that tutors with more expertise are in fact more effective than tutors with less expertise: "The question is still unsettled on the impact of tutoring expertise on learning gains" (Graesser et al., 2011, p. 411; see also VanLehn, 2011). Hence, our results add to the picture that tutors' expertise – in this study conceptualized as teaching experience – makes a difference in learning.

Sixth, we examined whether the difference in a tutee's learning between teacher tutors and student tutors could be explained by the fact that teacher tutors more often engaged in formative assessment than student tutors. The results of the mediation analysis, however, failed to show a significant result. Hence, teacher tutors and student tutors might have engaged in activities other than formative assessment that enhanced a tutee's learning. These activities, however, were not examined in this study.

Seventh, the mediation analysis that examined a tutor's summative assessment showed that the difference in the accuracy with which teacher tutors and student tutors assessed a tutee's understanding at the end of tutoring (Herppich et al., 2013) was attributable to the fact that teacher tutors more often engaged in formative assessment than student tutors. This finding again suggests a close relationship between formative and summative assessment.

In this study, our analysis was confined to a tutor's formative assessment in response to a tutee's knowledge deficits. Of course, formative assessment need not refer only to knowledge deficits but can also address, for example, what a learner has already understood correctly (Black & Wiliam, 2009). Hence, future research is encouraged to examine all events that can be formatively assessed in the course of instruction. This might reveal which kind of formative assessment is particularly beneficial for learning and summative assessment. In addition, we found that teacher tutors more often elicited knowledge deficits and more often engaged in formative assessment than did student tutors. As persons without teaching experience act as tutors in the majority of cases, such tutors might be trained in eliciting knowledge deficits from tutees and in responding to such knowledge deficits appropriately.

Whether trainings that target such factors in fact improve tutoring should be examined in future studies.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*, 167-207.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., …. Nickmans, G. (2006). A learning integrated assessment system. *Educational Research Review, 1*, 61-67.

Black, P. (1993). Formative and summative assessment by teachers. *Studies in Science Education, 21*, 49-97.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7-68.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability, 21*, 5-31.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

Cade, W. L., Copeland, J. L., Person, N. K., & D'Mello, S. K. (2008). *Dialogue modes in expert tutoring.* Paper presented at the Ninth International Conference on Intelligent Tutoring Systems, Montreal, Canada.

Chae, H. M., Kim, J. H., & Glass, M. (2005). *Effective behaviors in a comparison between novice and expert algebra tutors*. Paper presented at the Sixteenth Midwest AI and Cognitive Science Conference (MAICS), Dayton, OH.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73-105.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning, *Cognitive Science, 32*, 301-341.

Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22*, 363-387.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Cromley, J. G., & Azevedo, R. (2005). What do reading tutors do? A naturalistic study of more and less experienced tutors in reading. *Discourse Processes, 40*, 83-113.

D'Mello, S., Lehman, B. A., & Person, N. K. (2010). *Expert tutors feedback is immediate, direct, and discriminating.* Paper presented at the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), Key West, FL.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*, 360-389.

Graesser, A. C., D'Mello, S., & Cade, W. L. (2011). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 408-426). New York: Routledge.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9*, 495-522.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*, 365-379.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*, 408-420.

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*, 242-260.

Hoge, R., & Coladarci, T. (1989). Teacher-based judgments of academic achievement. *Review of Educational Research, 59*, 297-313.

Katz, S., Allbritton, D., & Connelly, J. (2003). Going beyond the problem given: How human tutors use postsolution discussions to support transfer. *International Journal of Artificial Intelligence and Education, 13*, 79–116.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). *Tutoring techniques in algebra. Cognition and Instruction, 7*, 197-224.

Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences, 2*, 277-305.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*, 5-13.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879-891.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16*, 93-115.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*, 13-48.

Shavelson, R. J. (2006). *On the integration of formative assessment in teaching and learning with implications for teacher education.* Paper prepared for the Stanford Education Assessment Laboratory and the University of Hawaii Curriculum Research and Development Group.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., …Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*, 295-314.

Shepard, L. A. (2005). *Formative assessment: Caveat emptor.* Paper presented at the ETS Invitational Conference 2005, New York.

Shepard, L. A., Hammerness, K., Darling-Hammond, L., & Rust, F. (with Baratz-Snowden, J., Gordon, E., Gutierrez, C., & Pacheco, A.) (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275-326). San Francisco: Jossey-Bass.

Snow, R. E., & Swanson, J. (1992). Instructional psychology: Aptitude, adaptation, and assessment. *Annual Review of Psychology, 43*, 583-626.

Stiggins, R. J. (2006). Assessment for learning: A key to motivation and achievement. *Edge, 2*, 3-19.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762.

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review, 22*, 271-296.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197-221.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*, 209-249.

Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*, 49-65.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität* [Rater agreement and rater reliability]. Göttingen, Germany: Hogrefe.

Wittwer, J., Nückles, M., Landmann, N., & Renkl, A. (2010). Can tutors be supported in giving effective explanations? *Journal of Educational Psychology, 102*, 74-89.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist, 43*, 49-64.

Table 1

Means, Standard Deviations, and Correlations of Study Variables for the N =42 Dyads with Tutor-Initiated Knowledge Deficits

Measure	Teacher tutors *M* (*SD*)	Student tutors *M* (*SD*)	All tutors *M* (*SD*)	1	2	3	4	5
1. Type of tutor	—	—	—	—				
2. Knowledge deficits	42.95 (26.15)	25.77 (25.49)	33.95 (26.93)	.32*	—			
3. Formative assessment	1.86 (1.41)	0.92 (0.82)	1.37 (1.22)	.39*	.59*	—		
4. Tutee's learning gain	4.60 (4.32)	2.32 (3.47)	3.40 (4.02)	.29*	.01	.27*	—	
5. Summative assessment	12.40 (2.28)	11.45 (2.94)	11.90 (2.66)	.18	.17	.32*	.35*	—

Note. All statistics refer to the sample of those dyads in which a tutee expressed a least one tutor-initiated knowledge deficit.

*$p < .05$.

Table 2

Means, Standard Deviations, and Correlations of Study Variables for the N = 38 Dyads with Tutee-Initiated Knowledge Deficits

	Teacher tutors	Student tutors	All tutors					
Measure	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	1	2	3	4	5
1. Type of tutor	—	—	—	—				
2. Knowledge deficits	11.35 (12.83)	5.94 (6.21)	8.79 (10.48)	.26	—			
3. Formative assessment	0.30 (0.41)	0.09 (0.17)	0.20 (0.34)	.31*	-.04	—		
4. Tutee's learning gain	4.85 (4.31)	2.06 (3.40)	3.53 (4.10)	.34*	.32*	.30*	—	
5. Summative assessment	12.45 (2.31)	10.72 (2.82)	11.63(2.68)	.33*	.02	.56*	.36*	—

Note. All statistics refer to the sample of those dyads in which a tutee expressed a least one tutee-initiated knowledge deficit.

*$p < .05$.

Figure 1. Example item from the concepts test. The test was used to measure a tutee's conceptual understanding about the human circulatory system and a tutor's assessment accuracy with regard to a tutee's conceptual understanding.

What is the task of the heart in the human organism?

- ☐ The heart pumps the blood.
- ☐ The heart cleans and filters the blood.
- ☐ The heart supplies the blood with oxygen.
- ☐ Don't know

Figure 2. A tutor's formative assessment in response to tutor-initiated knowledge deficits (left) and to tutee-initiated knowledge deficits (right) failed to mediate the relationship between type of tutor (i.e., teacher tutors and student tutors) and a tutee's learning gain. Numbers represent standardized path coefficients for direct and, in parentheses, total effects.

*$p < .05$.

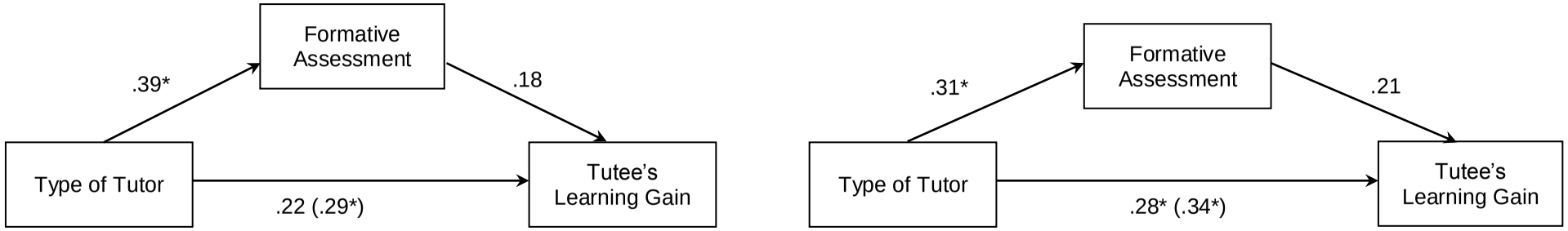
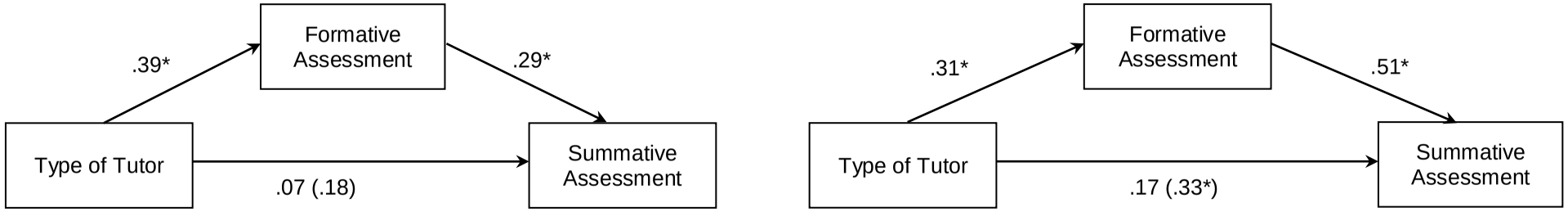


Figure 3. A tutor's formative assessment in response to tutor-initiated knowledge deficits (left) and to tutee-initiated knowledge deficits (right) mediated the relationship between type of tutor (i.e., teacher tutors and student tutors) and a tutor's summative assessment. Numbers represent standardized path coefficients for direct and, in parentheses, total effects.

*$p < .05$.



Chapter 2

Article 3:

Benefits for Processes Cause Decrements in Outcomes: Training Improves Tutors' Interactivity at the Expense of Assessment Accuracy

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (in press). Benefits for processes cause decrements in outcomes: Training improves tutors' interactivity at the expense of assessment accuracy. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

This article has been accepted for presentation as poster at the 35th Annual Conference of the Cognitive Science Society and for publication in the *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Acceptance was based on a strict review process.

Note. At the time this doctoral thesis was published, article 3 had been published in the *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. The official citation that should be used in referencing this material is Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Benefits for Processes Cause Decrements in Outcomes: Training Improves Tutors' Interactivity at the Expense of Assessment Accuracy. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2530-2535). Austin, TX: Cognitive Science Society. The version of the article printed in this doctoral thesis might not exactly replicate the final version published in the proceedings. The article is available online at http://mindmodeling.org/cogsci2013/papers/0458/paper0458.pdf

Benefits for Processes Cause Decrements in Outcomes: Training Improves Tutors' Interactivity at the Expense of Assessment Accuracy

Stephanie Herppich (stephanie.herppich@sowi.uni-goettingen.de)
University of Göttingen, Educational Institute
Waldweg 26, 37073 Göttingen, Germany

Jörg Wittwer (joerg.wittwer@sowi.uni-goettingen.de)
University of Göttingen, Educational Institute
Waldweg 26, 37073 Göttingen, Germany

Matthias Nückles (matthias.nueckles@ezw.uni-freiburg.de)
University of Freiburg, Department of Educational Science, Instructional and School Research
Rempartstrasse 11, 79098 Freiburg, Germany

Alexander Renkl (renkl@psychologie.uni-freiburg.de)
University of Freiburg, Department of Psychology, Developmental and Educational Psychology
Engelbergerstrasse 41, 79085 Freiburg, Germany

Abstract

Tutoring gives tutors the opportunity to engage in interactive strategies that help them to assess a tutee's understanding. However, tutors without teaching experience often do not engage in interactive strategies and, thus, have difficulty with accurately assessing a tutee's understanding. We conducted an experiment with 39 tutor-tutee dyads to test whether tutors who received training in interactive strategies would become more interactive and more accurate in assessing a tutee's understanding. Results showed that trained tutors provided a more interactive style of tutoring than untrained tutors. However, due to being more interactive, trained tutors produced less accurate assessments than untrained tutors. This suggests that changing the style of tutoring to implement interactive strategies puts a high burden on a tutor's cognitive capacity. Hence, there is obviously little cognitive capacity left that could be used to assess a tutee's understanding. Training methods that automate strategy use might enhance a tutor's assessment accuracy.

Keywords: one-on-one human tutoring; training; tutoring interactions; assessment accuracy

Introduction

In one-on-one tutoring, tutors have the possibility to engage in interactive tutoring strategies such as asking questions or providing hints. When a tutee responds to a tutor's interactive tutoring strategies, for example, by answering a question, a tutor can learn what a tutee does and does not know (Chi, 2009; Hmelo-Silver & Barrows, 2006). Thus, in the course of tutoring, a tutor has the opportunity to collect a multitude of information that can be used to summatively assess a tutee's understanding after tutoring session. This summative assessment may also help a tutor to prepare the next tutoring session by choosing material that is suited to a tutee's individual level of understanding (e.g., Chi, Jeong, & Siler, 2004; Kalyuga, 2007; cf. also the discussion of the concept of interim assessments for the school context by Perie, Marion, & Gong, 2009).

However, research has shown that inexperienced tutors, that is, tutors who are not trained in teaching (Chi et al., 2001; Graesser, D'Mello, & Cade, 2011), often do not engage in interactive tutoring strategies. Instead, they frequently dominate tutoring by providing lengthy explanations (e.g., Chi et al., 2001; Cromley & Azevedo, 2005). In addition, inexperienced tutors regularly fail to assess a tutee's understanding accurately (Chi et al., 2004; Herppich et al., 2013b).

Against this background, we conducted an experimental study to test whether inexperienced tutors who received training in interactive tutoring strategies would be able to implement an interactive style of tutoring. We were interested in whether a more interactive style of tutoring would benefit a tutor's assessment of a tutee's understanding after tutoring.

Tutoring Strategies of Experienced and Inexperienced Tutors and Their Influence on Assessment

In contrast to inexperienced tutors, experienced tutors are trained or experienced in teaching (cf. Cromley & Azevedo, 2005; D'Mello et al., 2010; McArthur, Stasz, & Zmuidzinas, 1990). Research shows that experienced tutors tend to provide a different style of tutoring than do inexperienced tutors. More specifically, experienced tutors more often engage in interactive tutoring strategies than inexperienced tutors. For example, they frequently scaffold a tutee by providing hints or asking questions (Cade et al., 2008; Chi, Roy, & Hausmann, 2008; Cromley & Azevedo, 2005). Scaffolding is a genuinely interactive tutoring strategy because it elicits constructive responses from a tutee (Hmelo-Silver & Barrows, 2006). In this vein,

Herppich et al. (2013a, 2013b) found that experienced tutors caused tutees to utter more knowledge deficits, that is, incomplete beliefs, incorrect beliefs, or misconceptions, in the course of tutoring than inexperienced tutors. In addition, experienced tutors were more accurate in assessing a tutee's understanding after tutoring than inexperienced tutors. The results suggest that a tutee's uttered knowledge deficits are diagnostically informative because they indicate what a tutee does not know (cf. Chi, et al., 2004; Cromley & Azevedo, 2005). Thus, tutors might derive information from these knowledge deficits that can be used to assess a tutee's understanding after tutoring.

Training Inexperienced Tutors

To test whether training inexperienced tutors in interactive tutoring strategies would improve their style of tutoring, we developed a training method that aimed at prompting inexperienced tutors to abstain from giving lengthy explanations and, instead, to engage in more interactive tutoring strategies such as scaffolding (cf. Chi, et al., 2008). As a result of implementing more interactive tutoring strategies in the course of tutoring, tutors were assumed to more intensively engage in collecting diagnostically relevant information that could be used to assess a tutee's understanding after tutoring.

Based on what is known about effective training methods in the domain of learning strategies (Mandl & Friedrich, 1992), the development of our training method was guided by several principles. First, training methods should inform about the advantages associated with the strategies targeted in the training. Second, training methods should directly convey knowledge about the strategies that need to be trained. Third, training methods should help to practice the targeted strategies (Klauer, 1988; Mandl & Friedrich, 1992). Research has shown that training methods that are in accordance with these principles are particularly effective (Dignath, Buettner, & Langfeldt, 2008; Leutner, Leopold, & Elzen-Rump, 2007).

By now, little attention has been given to training methods that aim at fostering an interactive tutoring style in the service of improving assessment accuracy. However, existing research on training tutors with the aim of enhancing a tutee's learning has well documented that tutors are often able to spontaneously implement the strategies that are targeted in training. Yet, tutors have difficulty with changing their style of tutoring in the long run (King, Staffieri, & Adelgais, 1998). Moreover, even though tutors are able to change their tutoring strategies, this might not necessarily increase the effectiveness of tutoring (Chi et al., 2001). In their review on tutoring-based instruction, Graesser et al. (2011) summarized research on tutor training in the following way:

> …it is difficult to train tutors to adopt particular strategies. They rely on their normal conversational and pedagogical styles.… it is difficult to force the human tutors to adopt changes in their language and discourse, particularly those levels that are unconscious and involuntary. (p. 422).

Hypotheses

In this study, we tested the effectiveness of a training method that aimed at helping tutors to implement a more interactive style of tutoring. We addressed the following hypotheses:

1) Trained tutors engage in more interactive tutoring strategies in the course of tutoring than untrained tutors.
2) Trained tutors are more accurate in assessing a tutee's understanding after tutoring than untrained tutors.
3) The more interactive style of tutoring explains why trained tutors are more accurate than untrained tutors in assessing a tutee's understanding after tutoring.

Method

Sample and Design

A total of $N = 39$ dyads of tutors and tutees participated in the experiment. The topic of tutoring was the human circulatory system. All tutors were university students majoring in biology with a mean age of 22.38 years (SD = 2.47). Thirty-five tutors were female and 4 tutors were male. Twenty tutors received training in interactive tutoring strategies (= *trained tutors*), whereas 19 tutors received no training (= *untrained tutors*). As indicated by a multiple-choice test, all tutors had sufficient knowledge about the human circulatory system. There was no significant difference in knowledge between trained tutors ($M = 8.45$, $SD = 2.26$) and untrained tutors ($M = 8.26$, $SD = 1.78$), $F(1, 37) = 0.81$, $p > .05$, $\eta^2 < .01$ (small effect). Moreover, trained (mean rank = 18.88) and untrained tutors (mean rank = 21.18) did not differ in their previous experience in providing tutoring, coded as 1 = *no experience*, 2 = *sporadic tutoring*, 3 = *regular tutoring*, $U = 167.50$, $z = -0.69$, $p > .05$, $r = -.11$ (small effect). Tutees were seventh-grade students from the middle track of the German school system (i.e., from *Realschulen*). Of the tutees, 9 were female and 29 were male; one tutee did not indicate gender.

Tutors were randomly assigned to the two experimental conditions (training vs. no training) and tutees were randomly assigned to tutors. The dependent variables in this experiment were the extent to which a tutor elicited knowledge deficits from a tutee in the course of tutoring and the accuracy with which a tutor assessed a tutee's understanding after tutoring.

Materials

Textbook Passage (Tutees and Tutors) In the tutoring session, the tutor-tutee dyads engaged in a dialogue based on a passage about the human circulatory system. We adapted this passage from the study by Chi et al. (2001). The passage consisted of 59 sentences and each sentence was printed on a separate sheet of paper. The sentences were presented to the tutor and the tutee in a ring binder.

Concepts Test (Tutees and Tutors) We used a shortened version of a test that was employed by Herppich et al. (2013b). This shortened version consisted of 16 multiple-choice items that assessed a tutee's understanding of concepts about the human circulatory system. For example, it included the following item: What is the task of the heart in the human organism? (1) The heart pumps the blood. (2) The heart cleans and filters the blood. (3) The heart supplies the blood with oxygen. (4) Don't know. The items of the original test were adapted from tests developed by Sungur and Tekkaya (2003) and by Michael et al. (2002) or constructed on the basis of the literature on misconceptions of the human circulatory system (e.g., Pelaez et al., 2005). A correct answer indicated a scientifically correct understanding of the concept. Each of the incorrect answers indicated a specific type of incorrect understanding of the concept. Hence, a tutee could achieve a maximum number of 16 points in the concepts test.

To examine the accuracy with which the tutors assessed a tutee's understanding of the human circulatory system after tutoring the tutors were also administered the test.

Training in Interactive Tutoring Strategies (Trained Tutors) The trained tutors received training in interactive tutoring strategies. The training took about 45 minutes and was presented on a computer screen. The training aimed at helping the trained tutors to adopt interactive tutoring strategies that would enable them to elicit knowledge deficits from a tutee. The training consisted of two building blocks. In the first building block, the trained tutors were informed about the problem that tutors often are not interactive and, thus, cannot accurately assess a tutee's understanding (Brown, Campione, & Day, 1981). Subsequently, the trained tutors were provided with information about three strategies, namely, (1) abstaining from giving lengthy explanations, (2) intensifying question asking, and (3) increasing scaffolding in response to a tutee's contribution (Cade et al., 2008, Chi et al., 2008; Herppich et al., 2013a). To learn about the three strategies, the trained tutors first read an explanatory text and then watched two videos of fictitious tutoring sessions. The first video presented a tutor who failed to engage in interactive tutoring strategies and, thus, to receive information about a tutee's understanding. The second video, in contrast, presented the same tutor who did engage in interactive tutoring strategies, which helped the tutor to receive information about a tutee's understanding (cf. Renkl, 2005). In the second building block, trained the tutors also watched videos that presented positive and negative examples of tutoring strategies. This time, however, the tutoring strategies were not explained to the trained tutors. Instead, the trained tutors were prompted to self-explain what constituted the difference between the positive and negative examples. More specifically, the trained tutors were asked to provide information about the tutoring strategies that they saw in the videos and about the effects of such tutoring strategies for assessing a tutee's understanding (cf. Renkl, 2005). Finally, the trained tutors were required to indicate what they would do in order to change the tutoring strategies that they saw in a negative example. This was done to actively stimulate the application of the to-be-learned strategies (cf. Klauer, 1988).

Introductory Text (Untrained Tutors) Instead of receiving training in interactive tutoring strategies, the untrained tutors read a short text. The text provided information about the effectiveness of tutoring and about problems associated with assessing a tutee's understanding. However, the untrained tutors did not receive any instruction on how to solve these problems. Instead, they were asked to provide tutoring in whatever manner they assumed appropriate.

Procedure

Each tutoring session was divided into three phases: pretest phase, tutoring phase, and posttest phase. On average, a tutoring session lasted about 3 hours.

In the pretest phase, each tutee and each tutor individually read the passage about the human circulatory system. Afterwards, the trained tutors received training and the untrained tutors read the text.

In the tutoring phase, tutor-tutee dyads jointly read the passage about the human circulatory system sentence-by-sentence and engaged in a dialogue about each sentence. All tutoring phases were videotaped.

In the posttest phase, the tutees completed the concepts test. The tutors also received the items of the concepts test and were asked to indicate for each item which of the given response options the tutee would choose.

Codings and Analyses

Elicitation of Knowledge Deficits (Tutors) As an indicator of engaging in interactive tutoring strategies, we coded the knowledge deficits that a tutor elicited from a tutee. To do so, we used a coding scheme adapted from Chi et al. (2004). Every knowledge deficit that a tutee uttered was coded from its beginning to its end (event sampling procedure).

We coded a knowledge deficit whenever a tutor elicited from a tutee an utterance that (1) contradicted a piece of knowledge stated in the textbook passage, that (2) was incomplete, that (3) was vague, that (4) was incorrect and not addressed by the textbook passage, or when the tutee (5) did not utter a certain piece of information at all, that is, the tutee obviously missed this piece of information. In one tutoring session, for example, the tutor asked: "Why does the blood need to go to the lung? What does the lung do?" And the tutee answered: "Yes, um, yes, the lung filters the blood." This answer was coded as utterance of a knowledge deficit because it represents a normatively incorrect understanding. To standardize coding, the coder used a written instruction. For each tutor-tutee dyad, we summed up the number of elicited knowledge deficits.

Summative Assessment (Tutors) To examine the accuracy with which a tutor assessed a tutee's understanding of the human circulatory system after tutoring, we compared a tutee's responses in the concepts test with a tutor's estimations of a tutee's responses in the concepts test. To do so, we made the comparison on an item-by-item basis (cf. Hoge & Coladarci, 1989). Hence, a tutor could achieve a maximum score of 16 points. Higher scores indicated a higher assessment accuracy.

Mediation Analysis To test our hypotheses, we performed a mediation analysis. We calculated total, direct, and indirect effects in accordance with our hypotheses by applying regression-based path analysis. To test for the statistical significance of an indirect effect, we derived 95% confidence intervals for indirect effects as well as standard errors for indirect effects via bias-corrected bootstrap (for guidelines, see, e.g., Hayes, 2009, 2012). This approach resolves some methodological problems associated with the Sobel test (Hayes, 2009).

Results

For all analyses, we used an alpha level of .05. For directional hypotheses, we used one-tailed tests. In the analyses, trained tutors were coded as 1 and untrained tutors were coded as 0. As effect size for indirect effects in the mediation analysis, we report κ^2. According to Preacher and Kelley (2011), effects are small when $\kappa^2 = .01$, medium when $\kappa^2 = .09$, and large when $\kappa^2 = .25$. All analyses were performed using SPSS 20.0.0, the PROCESS macro for SPSS introduced in Hayes (2012; to perform the mediation analysis), and AMOS 20.0.0 (to receive standardized path coefficients for the mediation analysis). Table 1 shows the means and standard deviations of the dependent variables.

Table 1: Means and standard deviations (in parentheses) of the experiment's dependent variables

Variable	Trained Tutors *M* (*SD*)	Untrained Tutors *M* (*SD*)	All Tutors *M* (*SD*)
Elicited Knowledge Deficits	71.30 (40.46)	32.11 (28.63)	52.21 (40.01)
Assessment Accuracy	8.05 (2.54)	8.21 (2.30)	8.13 (2.40)

Impact of Training on Implementing Interactive Tutoring Strategies

Our first hypothesis stated that trained tutors would more often engage in interactive tutoring strategies than untrained tutors. Thus, trained tutors should elicit more knowledge deficits from their tutees than untrained tutors. As can be seen in Figure 1, trained tutors elicited more utterances of knowledge deficits from their tutee than did untrained tutors, $R^2 = .25$, $F(1, 37) = 12.08$, $p < .05$, 95% CI [.26, .74]. Hence, the trained tutors in fact engaged in more interactive tutoring strategies than the untrained tutors.

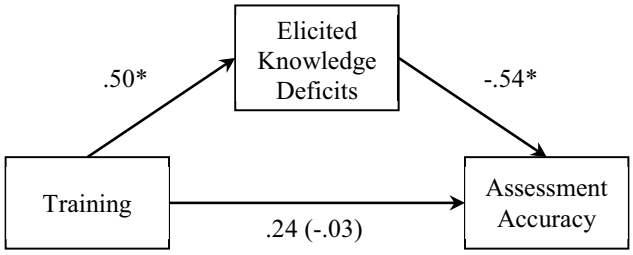


Figure 1: Mediation model for the effect of tutor training on assessment accuracy explained by the number of expressed knowledge deficits a tutor elicited from a tutee. Numbers represent standardized path coefficients for direct effects and, in parentheses, the total effect of the independent variable on the dependent variable. *$p < .05$.

Impact of Training on Summative Assessment

Our second hypothesis stated that trained tutors would more accurately assess a tutee's understanding after tutoring than untrained tutors. However, as the total effect depicted in Figure 1 shows, there was no significant difference in assessment accuracy between trained tutors and untrained tutors, $R^2 < .01$, $F(1, 37) = 0.04$, $p > .05$, 95% CI [-.31, .24]. Hence, if only zero-order relations are taken into account, training tutors to implement interactive tutoring strategies failed to exert an influence on assessment accuracy.

Interactive Tutoring Strategies as Mediator

Our third hypothesis stated that the higher number of a tutee's elicited knowledge deficits would explain why trained tutors assessed a tutee's understanding after tutoring more accurately than untrained tutors. To statistically test this hypothesis, we computed the indirect effect even though the total effect (i.e., the effect of training on assessment accuracy) was not significant (cf. Hayes, 2009; Shrout & Bolger, 2002). To test the indirect effect, we constructed a bias corrected 95% bootstrap confidence interval as well as bootstrap standard errors from 10000 bootstrap samples. We found a significant negative indirect effect indicating that implementing interactive tutoring strategies as a result of receiving training decreased assessment accuracy with a standardized point estimate of -.27 (*SE* = .10), 95% CI [-.46, -.12], $\kappa^2 = .26$ (zero-order correlation between elicited knowledge deficits and assessment accuracy: $r = -.43$, $p < .05$). Translated to unstandardized estimates, the number of items correctly estimated by trained tutors was 1.28 points (*SE* = 0.54) lower (and not higher) than the number of items correctly estimated by untrained tutors as mediated by the number of elicited knowledge deficits.

Discussion

This study examined the effectiveness of a training method that aimed at helping tutors to engage in interactive tutoring strategies in the course of tutoring. It was assumed that engaging in interactive tutoring strategies would benefit a tutor's assessment of a tutee's understanding after tutoring.

First, we found that trained tutors in fact showed a more interactive style of tutoring than untrained tutors. Hence, even though the duration of our training was rather short, it was obviously sufficient to help the tutors to implement more interactive tutoring strategies. As a result, tutees tutored by trained tutors more often uttered knowledge deficits than tutees tutored by untrained tutors. This finding is consistent with the results obtained by Herppich et al. (2013a).

Second, however, the trained tutors failed to assess a tutee's understanding more accurately than the untrained tutors. The trained tutors were even less accurate than the untrained tutors. As show by the mediation analysis, this result was explained by the greater extent to which trained tutors engaged in interactive tutoring strategies as a result of receiving training. This effect was probably not observable in the zero-order analysis because the two paths making up the indirect effect were opposite in sign (cf. Hayes, 2009).

An explanation for why trained tutors and untrained tutors did not differ in assessment accuracy, as indicated by the total effect in the mediation analysis, is that the changes in the tutoring strategies due to receiving training might not have been sufficient to produce changes in assessment accuracy. This explanation would be in accordance with the results obtained by Roscoe and Chi (2007), who found that strategies of tutors can only be influenced to a certain extent. Hence, in the context of the present study, the information gained from being more interactive might not have been enough to generate more accurate assessments (cf. Graesser et al., 2011).

However, it still remains an open question as to why the elicitation of knowledge deficits was detrimental for assessing a tutee's understanding after tutoring, as indicated by the indirect effect in the statistical analysis. First, it might be that trained tutors and untrained tutors differed in the types of knowledge deficits they elicited from a tutee. Eliciting a larger number of scientifically incorrect utterances as compared to missing knowledge pieces, for example, might have been more informative for the summative assessment. This is because the incorrect response options in the concepts test were based on common types of incorrect understanding of a concept (e.g., Pelaez et al., 2005). However, the relative number of knowledge deficits elicited per category did not differ significantly between trained tutors and untrained tutors for any of the five categories of knowledge deficits coded.

Second, the detrimental effect of eliciting knowledge deficits on summative assessment might be related to our measure of summative assessment accuracy. During the training, the tutors were repeatedly informed that a tutor should *get a picture of a tutee's understanding*. As a consequence, the trained tutors might have conceived a tutee's understanding on a more global level than on the level of conceptual understanding. Thus, after having completed the training, being more interactive and receiving more information from the tutees could have drawn the tutors' attention away from the knowledge they were to assess in the concepts test. This conjecture could be tested in future research that uses measures of assessment accuracy that are as manageable for tutors as a multiple-choice test on conceptual knowledge but that would tap different levels of a tutee's understanding.

Third, another explanation refers to the fact that the tutors in this study did not possess teaching experience. Hence, the interactive tutoring strategies targeted in the training might have been quite unfamiliar to the tutors. As a result, implementing interactive tutoring strategies during tutoring might have put a fairly high burden on a tutor's cognitive capacity (Feldon, 2007). Thus, there might not have been enough cognitive capacity left to derive information from a tutee's utterances of knowledge deficits as a basis for assessing a tutee's understanding after tutoring.

This interpretation is in accordance with results from research on the acquisition of memory strategies. Often, learners can spontaneously implement a newly learned memory strategy but experience a so-called *utilization deficiency* (Miller, 1990). That is, implementing the strategy does not immediately improve recall or even hinders it. It is argued that using a newly learned strategy, which is not yet automated, demands most of the cognitive capacity of a learner. Thus, there is little capacity left to spend on processing the material to be recalled (e.g., Miller & Seier, 1994).

Given this interpretation, it seems to be important to develop training methods that increase the automaticity with which interactive tutoring strategies are executed (Klauer, 1988). When interactive tutoring strategies occur more automatically, there might be more cognitive capacity available that can be used by tutors to assess a tutee's understanding (Feldon, 2007). Future research is encouraged to test whether training methods that target the automaticity of interactive tutoring strategies in fact improve assessment accuracy.

Acknowledgments

We thank Hannah Bartels, Victoria Denise Claes, Julian Etzel, Sophia Kammer, Rico Krieger, Amelie Krug, Annette Lehmann, Karina Meyer, Bosse Nietsch, Anne-Kristin Rückert, Tatjana Scharping, Eva Wiemers, Lisa Zimmer, and Raoul Zimmermann for their help with many practical aspects of the project. This research was supported by grants from the German Science Foundation DFG (WI 3348/2-1).

References

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher, 10*, 14-21.

Cade, W. L., Copeland, J. L., Person, N. K., & D'Mello, S. K. (2008). *Dialogue modes in expert tutoring.* Paper presented at the Ninth International Conference on Intelligent Tutoring Systems, Montreal, Canada.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73-105.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32*, 301-341.

Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately?. *Cognition and Instruction, 22*, 363-387.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. M. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533.

Cromley, J. G., & Azevedo, R. (2005). What do reading tutors do? A naturalistic study of more and less experienced tutors in reading. *Discourse Processes, 40*, 83-113.

Dignath, C., Buettner, G., & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review, 3*, 101-129.

D'Mello, S., Lehman, B. A., & Person, N. K. (2010). *Expert tutors feedback is immediate, direct, and discriminating.* Paper presented at the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), Key West, FL.

Feldon, D. F. (2007). Cognitive load and classroom teaching: The double-edged sword of automaticity. *Educational Psychologist, 42*, 123-137.

Graesser, A. C., D'Mello, S., & Cade, W. L. (2011). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction*. New York: Routledge.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*, 408-420.

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling* [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013a). *Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment.* Manuscript submitted for publication.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013b). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*, 242-260.

Hmelo-Silver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-based Learning, 1*, 21-39.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509-539.

King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology, 90*, 134-152.

Klauer, K. J. (1988). Teaching for learning-to-learn: A critical appraisal with some proposals. *Instructional Science, 17*, 351-367.

Leutner, D., Leopold, C., & Den Elzen-Rump, V. (2007). Self-regulated learning with a text-highlighting strategy: A training experiment. *Zeitschrift für Psychologie/ Journal of Psychology, 215*, 174-182.

Mandl, H., & Friedrich, H. F. (Eds.). (1992). *Lern- und Denkstrategien. Analyse und Intervention* [Learning and thinking strategies. Analysis and intervention]. Göttingen: Hogrefe.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction, 7*, 197-224.

Michael, J. A., Wenderoth, M. P., Modell, H. I., Cliff, W., Horwitz, B., McHale, P., … Whitescarver, S. (2002). Undergraduates' understanding of cardiovascular phenomena. *Advances in Physiology Education, 26*, 72-84.

Miller, P. H. (1990). The development of strategies of selective attention. In D. F. Bjorklund (Ed.), *Children's strategies: Contemporary views of cognitive development.* Hillsdale: Erlbaum.

Miller, P. H., & Seier, W. L. (1994). Strategy utilization deficiencies in children: When, where, and why. *Advances in Child Development and Behavior, 25*, 107-156.

Pelaez, N. J., Boyd, D. D., Rojas, J. B., & Hoover, M. A. (2005). Prevalence of blood circulation misconceptions among prospective elementary teachers. *Advances in Physiology Education, 29*, 172-181.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*, 5-13.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16*, 93-115.

Renkl, A. (2005). The worked-out examples principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning.* New York: Cambridge University Press.

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*, 534-574.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*, 422-445.

Sungur, S., & Tekkaya, C. (2003). Students' achievement in human circulatory system unit: The effect of reasoning ability and gender. *Journal of Science Education and Technology, 12*, 59-64.

General Discussion

One-on-one tutoring has been identified as a highly effective form of instruction (e.g., Bloom, 1984; Ritter et al., 2009; VanLehn, 2011). Moreover, assessments and assessment accuracy of an instructor are deemed crucial for the effectiveness of instruction in general (e.g. Klug et al., 2013; Schrader, 2010; Vogt & Rogalla, 2009). In assessing a tutee's understanding, tutors should particularly benefit from the one-on-one situation (Chi et al., 2004; Putnam, 1987; see also Snow & Swanson, 1992). Better assessments as a consequence of the one-on-one situation might contribute to the notable effectiveness of tutoring (Chi et al., 2004). This doctoral thesis aims at providing insight into tutors' assessments of a tutee's understanding in one-on-one tutoring. To do so, two studies were conducted. In the first study, the assessment accuracy of tutors with varying levels of experience was examined (see *Chapter 1*, articles 1 and 2). In this vein, the relationship between a tutor's formative assessments and a tutor's summative assessments was analyzed. Moreover, the significance of a tutor's formative assessments for the effectiveness of tutoring was examined. In the second study, it was experimentally tested whether tutors' assessments can be enhanced by a rather short training method. The aim of this training method was to render the tutors' style of tutoring more interactive (see *Chapter 2*, article 3).

Summary of Results

Findings Presented in Article 1

Article 1 focuses on a tutor's summative assessments after tutoring. The results reported in this article point out that the tutors generally had difficulty in summatively assessing a tutee's understanding correctly. This was true for a tutee's understanding at the level of single concepts and at the level of more complex mental models. Moreover, this was true for a tutor's relative assessment accuracy and a tutor's absolute assessment accuracy. In particular, the tutors largely overestimated their tutee's absolute levels of learning outcomes.

Nevertheless, experienced tutors (*teacher tutors*) differed from inexperienced tutors (*student tutors*) in the accuracy of their assessments. At the complex level of mental models, these differences were only subtle. That is, the teacher tutors were only more accurate than the student tutors in assessing a tutee's relative understanding after half of the tutoring session. Their relative assessments were not more accurate after tutoring any

longer. Neither were the teacher tutors' absolute assessments more accurate than the student tutors' absolute assessments. However, whereas the teacher tutors became slightly more accurate in assessing a tutee's understanding, the student tutors became less accurate in assessing a tutee's understanding.

At the level of single concepts, the teacher tutors generally outperformed the student tutors in accurately assessing a tutee's understanding. Finally, the teacher tutors could self-rate more accurately than the student tutors whether they were able to accurately assess their tutee's understanding at the level of mental models.

Findings Presented in Article 2

Article 2 concentrates on a tutor's formative assessments in the course of a tutoring session. The findings presented in this article reveal that all tutors engaged in strategies to formatively assess a tutee's expressed knowledge deficits during tutoring. They did so more often in response to tutor-initiated expressed knowledge deficits than in response to tutee-initiated expressed knowledge deficits. Larger amounts of formative assessment generally enhanced a tutee's learning. This was true, irrespective of the type of expressed knowledge deficit that was formatively assessed. Larger amounts of formative assessment also generally yielded more accurate summative assessments of the tutee's understanding at the level of single concepts. Likewise, this was true for both types of expressed knowledge deficits.

Teacher tutors differed from student tutors on several measures. These measures were, first, the interactivity of the tutors' tutoring style. More precisely, teacher tutors had their tutees more often express knowledge deficits than had the student tutors. Second, teacher tutors differed from student tutors in the amount of formative assessment. That is, the teacher tutors responded to a tutee's expressed knowledge deficits more often with strategies of formative assessment than did the student tutors. Third, teacher tutors differed from student tutors in their effectiveness. The teacher tutors produced more learning in their tutees than did the student tutors. This difference, however, was not explained by the difference in the amount of formative assessment between the teacher tutors and the student tutors.

On the contrary, the amount of formative assessment accounted for differences in the accuracy of summative assessments at the level of concepts (see article 1; *Chapter 1*) between teacher tutors and student tutors. Teacher tutors where more accurate in

summatively assessing a tutee's understanding than student tutors because they made use of strategies of formative assessment more often than student tutors.

Findings Presented in Article 3

The second study (see article 3; Chapter 2) showed, first, that the trained tutors elicited more expressions of knowledge deficits from their tutees (i.e., tutor-initiated expressed knowledge deficits) than did the untrained tutors. Second, however, the trained tutors and the untrained tutors did not significantly differ in the accuracy of their summative assessments after tutoring at the level of concepts. Thus, the overall effect of the training method on the accuracy of the tutors' assessments was not significant. Nevertheless, a mediation analysis, third, revealed a significant indirect effect of the training method on the accuracy of the tutors' summative assessments via the style of tutoring. Contrary to the hypothesis, the trained tutors were not more but less accurate than the untrained tutors in summatively assessing a tutee's understanding after tutoring. This was explained by a more interactive style of tutoring.

Tutors' Assessments and Assessment Difficulties

For the assessments and assessment difficulties of tutors, the results of the reported studies suggest at least two lines of implications. The first line of implications focuses on the tutors' formative assessments. The second line of implications is concerned with the tutors' summative assessments.

Formative Assessments – Tutors' Strengths

Formative Assessment and Learning. This doctoral thesis picks up on two area of research to examine a tutor's instructional strategies and to relate these strategies to a tutee's learning. The first area is tutoring research that studies a tutor's interactive instructional strategies. The second area is research on assessments in classroom situations that studies an instructor's formative assessments. Both tutoring research and research on classroom assessments indicate that instructional strategies such as scaffolding and giving feedback foster learning. In both areas, however, evidence is vague (Bennett, 2011; Kingston & Nash, 2011; VanLehn, 2011). In tutoring research, this is partially because scientists seldom measured learning outcomes when they studied tutoring processes (cf. Graesser et al., 2011; for exceptions see Chi et al., 2001, 2008). Moreover, in research on classroom assessments, the definition of formative assessment lacks clarity (cf. the section *Forms of*

Assessment). This issue makes it difficult to judge under which conditions formative assessment enhances learning (Bennett, 2011; Kingston & Nash, 2011; see also Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012; Kingston & Nash, 2012). In the studies presented by this doctoral thesis, the concept of formative assessment was used to examine a tutor's instructional strategies such as scaffolding and giving feedback during tutoring. In this vein the positive effect of formative assessment on a tutee's learning was confirmed. Consequently, the results of this doctoral thesis add to the evidence that instructional strategies such as scaffolding and giving feedback foster learning in tutoring research and in research on assessments. In particular, these results corroborate the assumption that these strategies are part of the mechanisms that make tutoring effective (e.g., Chi, 2009). Thereby, the doctoral thesis helps to disentangle the mechanisms at work in tutoring.

Formative Assessment and Summative Assessment. The formative assessment view on instructional strategies also has suggested that a tutor's scaffolding and feedback may be connected to a tutor's summative assessment after tutoring. The relationship between formative and summative assessments has been theoretically discussed in research on educational assessment (Bennett, 2011; Harlen & James, 1997; Perie et al., 2009). Yet, empirical evidence for such a relationship is lacking. In particular, no study has been conducted in the context of tutoring. The first study of this doctoral thesis (cf. article 2, *Chapter 1*), however, highlighted that more formative assessment in response to a tutee's expressed knowledge deficits yielded more accurate summative assessments after tutoring. This outcome, accordingly, promotes the theoretical understanding of educational assessments. It might also be possible to generalize this evidence to other forms of instruction. This is because the study has drawn on concepts of formative and summative assessment that are common, for example, in the study of classroom instruction, as well (e.g., Shepard, 2001).

Furthermore, discovering that formative assessments reliably inform summative assessment is encouraging with regard to the quality of tutoring in practical settings. Tutoring is most often implemented as an informal form of education such as private tutoring or parent tutoring (Haag, 2010; Mol et al., 2008). It can be assumed that tutors in these contexts seldom make use of formal forms of assessment such as tests and exams to document a tutee's learning outcomes. Ongoing formative assessments that do not inform or that even misinform summative assessments of the tutee's understanding would, thus, be a threat to the effectiveness of tutoring on the long run (see also Harlen & James, 1997).

Tutors Do Formatively Assess a Tutee's Understanding. Nevertheless, the positive relationship between formative assessments and summative assessments can only be assumed encouraging provided that tutors are able to formatively assess their tutee's understanding. In this regard, the tutors in the first study (article 2; see *Chapter 1*) at least responded to more than 90% of all expressed knowledge deficits with either scaffolding, feedback, or the correct answer. This finding does not necessarily imply that the tutors formatively assessed most of the expressed knowledge deficits. Yet, they neither ignored them (for a deviating result, see Chi et al., 2004). The tutors, overall, were rather weak in formatively assessing tutee-initiated expressed knowledge deficits. However, they showed some skills in formatively assessing the tutor-initiated expressed knowledge deficits. Even student tutors, on average, responded to almost half of the tutor-initiated expressed knowledge deficits with strategies of formative assessment. This result is encouraging because tutor-initiated expressed knowledge deficits made up the vast majority of all expressed knowledge deficits. Certainly, this formative assessment still leaves room for improvement. Nevertheless, tutors obviously do formatively assess a tutee's understanding to a reasonable extent.

Summative Assessments – Tutors' Limitations

The tutors' performance with regard to their summative assessments has to be interpreted as being more problematic (article 1; see *Chapter 1*). Although more formative assessments added to more accurate summative assessments, the tutors were not very accurate in summatively assessing a tutee's understanding after tutoring. The tutors seemed to struggle particularly with the summative assessments of a tutee's mental model. Yet, at this point, it cannot be determined as to which extent these problems were due to the complexity of the knowledge to be assessed and as to which extent these problems were due to the unfamiliarity with the method of assessing mental models (see article 1, *Chapter 1*). Assessment problems that are caused by unfamiliarity with a certain method of assessing understanding could be solved by introducing the method before it is used (cf. Yin et al., 2008).

Cognitive Biases. With regard to difficulties that are caused by the task of assessing a tutee's understanding itself, it has been assumed that tutors might too heavily draw on their own normatively correct understanding when they assess a tutee's understanding (cf. Chi et al., 2004; Nickerson, 1999). In his theory on how people construct a model of another

person's knowledge, for example, Nickerson (1999) states that people, first, use their own knowledge as an anchor to construct a default model of another person's knowledge. This default model is, second, adjusted into a more person-specific model, based on categorical information about the person. Such information can include the community to which the person belongs, for example. Third, the individual model is constantly adjusted in accordance with information obtained when interacting with the person. The tutors in the first study possessed a rather correct understanding of the human circulatory system. If they used their own understanding as an anchor to build a model of their tutee's understanding and did not adjust this model sufficiently, overestimation of the tutee's understanding should result.

Beyond this interpretation, other factors can affect a tutor's assessment accuracy. Just as the anchor-and-adjustment heuristic (Tversky & Kahneman, 1974) that Nickerson (1999) uses in his theory on how people construct a model of another person's knowledge, these factors can be categorized as (social) cognitive biases (Hesse & Latzko, 2011; Schrader, 2010). That is, the assessment process is influenced by certain systematic deviations from accurate judgments that can be relevant to other processes of perception and judgment as well. Many of these biases are interrelated and interact with each other in affecting processes of perception and judgment (Fisseni, 1997; Hesse & Latzko, 2011). Generally, a tutor's assessments may be influenced by this tutor's implicit and subjective theories of personality (Hofer, 1986) and by certain characteristics the tutor perceives in the tutee (e.g., the halo effect; Fisseni, 1997). Such factors can cause the tutor's absolute and relative assessment to deviate from the tutee's actual learning outcome.

Besides the tendency to draw too heavily on one's own knowledge, several biases can particularly account for a tutor's overestimation of a tutee's understanding. A self-concept of being an effective instructor, for example, might lead a tutor to expect large learning outcomes from the tutee. Subsequently, this tutor might tend to perceive and assess a high learning outcome in the tutee (cf. self-fulfilling prophecies, e.g., Fisseni, 1997; Hesse & Latzko, 2011). This tendency could be more pronounced for tutors who perceive small learning outcomes as threats to the own self-concept and related self-esteem (Rheinberg, 2002). Partially in line with Nickerson's (1999) reasoning (see also Keysar, Barr, & Horton, 1998), Graesser, D'Mello, and Person (2009, p. 376) have identified five illusions of tutors (and tutees) about communication processes and cognitions. Accordingly, tutors believe that they and the tutee have a common understanding about what is being discussed (*illusion of grounding*). Tutors believe that a tutee's feedback

about this tutee's understanding is correct, although often it is not (*illusion of feedback accuracy*). Tutors believe that the tutee generally understands the intentions of a tutor's contributions to the tutoring dialogue (*illusion of discourse alignment*). Tutors believe that the tutee has already understood much more in the course of tutoring than it is the case (*illusion of student mastery*). Finally, tutors believe that the tutee understands and learns whatever a tutor says (*illusion of knowledge transfer*). It is plausible to assume that these illusions also apply to the tutors studied in this doctoral thesis.

In the face of all these threats to assessment accuracy, Weinert and Schrader (1986) argue that informal educational assessments cannot and need not be perfectly accurate. They assume that assessments that are moderately biased in favor of the learner (i.e., overestimate learning outcomes) should have positive educational consequences for future learning. Nevertheless, they caution instructors to steadily monitor and correct their assessments as to avoid severe deviations in assessment accuracy. Similarly, Graesser et al. (2011) argue that tutors should be skeptical about their perceptions of a tutee's understanding. Future research could help to disentangle which factors most severely affect a tutor's assessments. Subsequently, measures could be developed or selected (e.g., Nickerson, 1999) that help tutors and other instructors to monitor and partially correct their assessments.

Implications for Group-Based Instruction. Thus, the effectiveness of tutoring could be enhanced by improving a tutor's assessments. However, it seems unrealistic to expect perfectly accurate assessments from tutors (cf. also Graesser, Conley, & Olney, 2012). Finding that even tutors in one-on-one tutoring have difficulty in assessing a tutee's understanding has implications for other instructional settings. The maximal accuracy with which a tutor can assess a tutee's absolute understanding (under optimal conditions) might form a threshold for the accuracy that can be attained by instructors in group-based instruction, such as classroom instruction. Dealing with a group of learners obviously increases the challenges an instructor has to face (Feldon, 2007). Consequently, there is less opportunity and less capacity to assess the understanding of single learners. It can, thus, be assumed that the absolute assessment accuracy of tutors sets a limit to the absolute assessment accuracy that can be expected for instructors in group-based instruction.

This doctoral thesis yields further evidence for tutors' difficulty in assessing a tutee's understanding (cf., Chi et al., 2004; Graesser et al., 1995). Yet, uniquely it also emphasizes that the tutors showed some skills in formatively assessing a tutee's

understanding. As has been suggested for research on personality judgment (Funder, 2012), it may be time to shift from detecting tutors' deficits and limitations in educational assessments. Instead, effort could be spent on finding their strengths, such as existing skills in formative assessment. These strengths could, in turn, be supported to improve assessments (cf. Dünnebier et al., 2009). This doctoral thesis made a step in this direction by analyzing a tutor's formative assessments and their relationships to learning outcomes and summative assessment. It made further steps in this direction by studying the assessment accuracy of tutors with varying levels of experience and by training tutors' formative assessments.

Influences on Assessments: Differences Between Experienced and Inexperienced Tutors

Previous research has suggested that tutors with teaching experience are more accurate in assessing a tutee's understanding than tutors without teaching experience (Dünnebier et al., 2009; Graesser et al., 2011; Mulholland & Berliner, 1992; cf. the section *Tutors' Assessments*). Therefore, this doctoral thesis compares teacher tutors' summative assessments with student tutors' summative assessments (article 1, cf. also article 2; *Chapter 1*). Of course, teaching experience is a rather broad category to study differential influences on assessment accuracy. Consequently, a goal of the doctoral thesis also is to uncover factors that hide behind the concept of teaching experience.

To pursue this goal the first approach (i.e., studying influences of general characteristics on tutoring) and the second approach (i.e., studying the structures and processes of tutoring) to the study of tutoring sensu Graesser et al. (2011) were integrated into the design of the first study (see *Chapter 1*). This was done to obtain more information about the assessments of tutors than by employing only one approach. In this vein, first, teacher tutors' and student tutors' formative assessments of a tutee's expressed knowledge deficits were studied as facets of a tutor's actual behavior (article 2; see *Chapter 1*). Second, the doctoral thesis revealed results that point towards differences in cognitive processing between teacher tutors and student tutors.

Teacher Tutors' Versus Student Tutors' Formative Assessments

Tutees of teacher tutors uttered more knowledge deficits than tutees of student tutors. In line with prior research (e.g., Chae et al., 2005; Chi et al., 2008), teacher tutors and student tutors also differed in the amount of scaffolding and feedback as compared with correct

answers that they provided following a tutee's expressed knowledge deficit. This difference, in turn, significantly accounted for the teacher tutors' more accurate assessments in comparison with the student tutors' assessments. Thus, the abstract concept of teaching experience is, indeed, reflected in certain observable behaviors that are relevant in assessing a tutee's understanding.

Teacher Tutors' Versus Student Tutors' Cognitive Processes

Furthermore, prior research has implied that cognitive processes of tutors might vary as a consequence of their level of teaching experience. For example, Graesser et al. (2009; 2011) have conjectured that experienced tutors likely are less prone to the illusions about communication processes and cognition they identified (cf. the section *Summative Assessments – Tutors' Limitations*). That is, experienced tutors might be better at meta-cognitively monitoring their behavior. Research on expertise in other domains than tutoring corroborates this conjecture. This research has found that experts in a particular domain are better able than non-experts to self-monitor their behavior in their domain of expertise (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Zimmerman, 2006). To be able to monitor their behavior, tutors, however, need free cognitive capacity at their disposal. In this regard, studies on expertise in teaching have established that experienced teachers, as compared with novice teachers, have automated basic teaching processes. Accordingly, experienced teachers can spend more cognitive resources on complex activities such as assessing learners' understanding while monitoring their own behavior and thoughts than can novice teachers (Feldon, 2007; see also Dünnebier et al., 2009).

Findings of this doctoral thesis suggest similar differences between cognitive processes of teacher tutors and of student tutors. More precisely, it can be conjectured that student tutors had less cognitive resources at their disposal than teacher tutors to assess a tutee's understanding and to monitor their own behavior and thoughts. Instead, student tutors' assessments were probably more impaired by social cognitive biases. These interpretations are based on the following results.

First, student tutors were not only less accurate in assessing a tutee's understanding at the level of concepts but they also increasingly overestimated a tutee's mental model in the course of a tutoring session. Teacher tutors' assessments of a tutee's mental model, on the contrary, became slightly more accurate in the course of a tutoring session (article 1; see also article 2; *Chapter 1*). Thus, the student tutors might have been less able than the teacher tutors to adjust their hypothetical model of the tutee's knowledge towards the

tutee's actual understanding (cf. Nickerson, 1999). Moreover, the student tutors might specifically have fallen prey to illusions of cognition and communication (Graesser et al., 2009) such as the illusion of student mastery and the illusion of knowledge transfer.

Second, student tutors were significantly worse at self-rating their assessment accuracy than were teacher tutors (article 1; see *Chapter 1*). Hence, they were possibly not able to spend as many resources on monitoring their assessments as teacher tutors. Of course, these conclusions await further empirical investigation. Future research could test for differences in cognitive strain between teacher tutors and student tutors, for example, by measuring cognitive load during tutoring (e.g., DeLeeuw & Mayer, 2008). Furthermore, tutors' meta-cognitive monitoring could be examined with the help of think-aloud protocols (e.g., Klingner, 2004).

Practical Implications of Differences Between Teacher Tutors and Student Tutors

These interpretations also have practical implications. As highlighted above, it is assumed that teacher tutors have more cognitive capacity at their disposal during tutoring than student tutors. This difference between teacher tutors and student tutors is attributed to the teacher tutors' expertise in teaching (cf. Dünnebier et al., 2009; Feldon, 2007; Zimmerman, 2006). To become an expert in a certain domain, however, it takes years of intensive practice (e.g., Ericsson, Krampe, & Tesch-Römer, 1993). Accordingly, the cognitive capacity of student tutors probably cannot be expanded easily to a reasonable extent. It would, thus, be very uneconomical to train student tutors' overall teaching skills as to free up cognitive capacity and consequently enhance the student tutors' assessments for common practical applications of tutoring (cf. Belzer, 2006).

To improve assessments for the purpose of tutoring, it might, however, not be necessary to train overall teaching skills. This is because tutors in private tutoring or parent tutoring often have relatively limited teaching tasks (Song, Park, & Sang, 2013; Whitehurst et al., 1988). These tasks can be assumed to be less complex than tasks of teachers in classroom instruction. Furthermore, tutors complete these tasks in a social situation that is also less complex than the situation in classroom instruction (cf., Feldon, 2007; Leutner, 2010).

For the purpose of tutoring, it might, consequently, be adequate to coach tutors in defined observable tutoring activities that have been found to foster assessments (cf. Graesser et al., 2012). Such a procedure would be much more economic for practical applications of tutoring. To do so, one option could be teaching tutors observable strategies

of formative assessment. These strategies were found to be central in assessing a tutee's understanding (article 2; *see Chapter 1*). Moreover, as discussed above, teacher tutors were more accurate in summatively assessing a tutee's understanding because they made use of these strategies to a larger amount. An approach to teaching tutors the application of strategies of formative assessment is discussed in the next section.

Training Inexperienced Tutors' Assessments and Assessment Accuracy

In the second study (see article 3; *Chapter 2*) of this doctoral thesis, student tutors were trained to implement an interactive style of tutoring that comprised strategies of formative assessment such as scaffolding and giving feedback. The training method aimed at improving the tutors' assessment accuracy. By experimentally manipulating activities of tutors, the study drew on the third approach to studying tutoring (i.e., manipulating tutoring activities) sensu Graesser et al. (2011).

Strengths and Limitations of the Training Method

Despite its brevity, the training method, indeed, succeeded in changing the tutoring style of the trained tutors. This was the case, although research has repeatedly reported that it is difficult to have tutors apply new instructional strategies (Bell, 2001; Belzer, 2006; Graesser et al., 2011). In this regard, the short computer-based training method was not only effective but also efficient.

The trained tutors, however, were not more accurate than the untrained tutors in summatively assessing the tutee's understanding after tutoring. Instead, the trained tutors' summative assessments were less accurate than the untrained tutors' summative assessments, as a consequence of a more interactive style of tutoring. In the first study, on the contrary, *teacher tutors'* summative assessments were more accurate than *student tutor's* summative assessments because the teacher tutors implemented a more interactive style of tutoring that comprised formative assessment (see article 2; *Chapter 1*). Remember that in the second study (*Chapter 2*) both trained tutors and untrained tutors were student tutors. This contrast in results between the two studies of this doctoral thesis, thus, supports the interpretation that student tutors differ from teacher tutors with respect to cognitive processing (cf. the section *Teacher Tutors' Versus Student Tutors' Cognitive Processes*). As the trained tutors in the second study likely did not have teaching experience they presumably were cognitively overstrained when they implemented the newly learned strategies. Hence, the trained tutors might have lacked cognitive capacity to process (most)

information they gained from the tutee via formative assessment. Consequently, their summative assessments of the tutee's understanding were even worse than the untrained tutors' assessments (cf. article 3, *Chapter 2*).

Implications for the Design of a Training Method

What do these findings mean for the design of a training method that aims at fostering tutors' assessment of a tutee's understanding? As in the laboratory study by Chi et al. (2001), the short training method implemented in the second study was successful in changing actual tutoring behavior. Yet, in both cases, the desired changes in effect measures did not result. Based on the conclusions about differences in cognitive processes between teacher tutors and student tutors, it can be assumed that the trained student tutors lacked disposable cognitive capacity. Hence, they might have been unable to process the information gained by using the strategies of formative assessment (see previous section).

Accordingly, tutors could intensively practice the implementation of interactive strategies during the training phase until strategy use becomes an automatic process. Implementation of automatic processes requires minimal cognitive resources (Feldon, 2007). Thus, after practice, trained tutors might have enough cognitive capacity at their disposal to process the information gained from the tutee. As a consequence, trained tutors' assessments might be enhanced (cf. Klauer, 1988; Friedrich & Mandl, 1992; Zimmerman, 2006). Such a training method, however, would be effortful in time and cost and, therefore, probably not feasible in the context of tutoring (Baker et al., 2000; Belzer, 2006; Graesser et al., 2011).

Nevertheless, field studies demonstrated promising results of training methods that were similarly short and economic as the training method applied in the second study of this doctoral thesis. The training methods in these field studies successfully enhanced the effectiveness of tutors in terms of a tutee's learning (Arnold et al., 1994; Blom-Hoffman et al., 2006; Chow & McBride-Chang, 2003). These field studies differed from the procedure applied in the second study of this doctoral thesis and from the procedure applied by Chi et al. (2001) with regard to the time of measurement. Effects in the investigations by Arnold et al. (1994), Blom-Hoffman et al. (2006), and Chow and McBride-Chang (2003) were measured after several tutoring sessions. Thus, it might not be necessary to intensify the training itself. Instead, practicing an interactive style of tutoring during in vivo tutoring sessions could be sufficient to attain a different quality in the use of trained strategies. Consequently, trained tutors might excel untrained tutors in summatively assessing a

tutee's understanding if measurement of summative assessment accuracy was carried out after more than one tutoring session.

Conclusion

This doctoral thesis examines tutors' assessments of a tutee's understanding from several angles using multiple approaches (cf. Graesser et al., 2011). It has revealed important results. It has found that tutors are, on average, at best moderately accurate in assessing a tutee's understanding. This was particularly true for the tutors' summative assessments. Thus, tutors obviously need not to be maximally skilled in assessing a tutee's understanding for tutoring to be effective. Nevertheless, the tutors showed some skills in formatively assessing a tutee's knowledge deficits. More formative assessment yielded more learning. Enhancing tutors' assessment skills could, therefore, make tutoring even more effective. Teacher tutors were more accurate in summatively assessing a tutee's understanding than student tutors on some measures. Hence, teaching experience, as a broad category, influences tutors' assessment accuracy. More precisely, the difference in assessment accuracy between teacher tutors and student tutors was explained by teacher tutors' prominent use of interactive instructional strategies to formatively assess a tutee's understanding. Thus, the doctoral thesis has found an observable indicator of teaching experience that is relevant to a tutor's assessments. In addition, it was possible to train student tutors to implement these interactive instructional strategies. However, the trained tutors' assessments did not become more accurate than the assessments of untrained student tutors. This result is interpreted with respect to possible differences in cognitive processing between teacher tutors and student tutors. Less efficient cognitive processing by student tutors might have barred the interactive style of tutoring from entailing more accurate assessments. Consequently, the design of the training method and the timing of measurement should be examined in more detail.

In this doctoral thesis, tutors' assessments were studied under laboratory conditions. To safeguard the external validity of the obtained findings, results should be validated under field conditions (Bortz & Schuster, 2010). To test the external validity of the findings, tutors' assessments should also be investigated using other content domains. By having tutors teach the human circulatory system, in this doctoral thesis a conceptual content domain has been used. Hence, the results should particularly be validated with a procedural content domain such as problem solving in sub-domains of physics or mathematics (Chi et al., 2004).

Overall, this doctoral thesis contributes to understanding the mechanisms that are at work in tutoring. Results of this doctoral thesis can be conducive to making tutoring even more effective and efficient. Moreover, implications for other forms of instruction, such as classroom instruction, can be derived from the obtained results. This doctoral thesis, thus, provides important insight into the field of effective instruction. In doing so, it contributes to good education.

Summary

The goal of this doctoral thesis is to study instruction that is effective in terms of supporting school-aged students' learning. This instruction, thus, contributes to good education. Instruction is of major interest because it can be modified to improve education. To do so, it is necessary to know as to what forms of instruction are effective. Moreover, it is necessary to know as to which mechanisms underlie these forms of instruction. One prominent form of effective instruction is one-on-one human tutoring. In this context, assessments and assessment accuracy of tutors are deemed central mechanisms with regard to the effectiveness of tutoring. However, these mechanisms have not been intensively studied yet. Therefore, this doctoral thesis investigates the assessments and the assessment accuracy of tutors. More precisely, two types of assessments are examined, namely, assessments that a tutor conducts continuously in the course of a tutoring session (i.e., formative assessment) and assessment that a tutor conducts after the completion of a tutoring session (i.e., summative assessment).

In this doctoral thesis, two empirical studies are reported. In both studies more knowledgeable tutors tutored school-aged tutees. It was assumed that tutors with teaching experience are more proficient in assessing a tutee's understanding than tutors without teaching experience. In the first study, the influence of teaching experience on a tutor's assessments was investigated empirically. That is, the assessment accuracy of tutors with teaching experience (i.e., teacher tutors) and the assessment accuracy of tutors without teaching experience (i.e., student tutors) were examined and compared with each other. In doing so, the relationship between a tutor's formative assessments and a tutor's summative assessments was analyzed. Moreover, the benefits of a tutor's formative assessments for a tutee's learning were investigated (see *Chapter 1*, articles 1 and 2). In the second study, it was experimentally tested whether the accuracy of student tutors' assessments can be enhanced by a short training method that aimed at fostering an interactive style of tutoring. The idea for the design of the training method was that tutees more likely express their own understanding when tutors implement an interactive style of tutoring. Based on the additional information about their tutee's understanding, tutors with an interactive style of tutoring should be better able to summatively assess the tutee's understanding than tutors with a style of tutoring that is less interactive (see *Chapter 2*, article 3).

In the first study (cf. *Chapter 1*), it was found that tutors were, on average, at best moderately accurate in summatively assessing a tutee's understanding. However, teacher

tutors were more accurate in summatively assessing their tutee's understanding than were student tutors (cf. article 1). Furthermore, the first study showed that all tutors engaged in interactive instructional strategies to formatively assess their tutee's understanding. More formative assessment, in turn, enhanced a tutee's learning. Similarly, larger amounts of formative assessment yielded more accurate summative assessments of a tutee's understanding. As in the case of summative assessments, teacher tutors and student tutors differed with regard to formative assessments. More precisely, teacher tutors more often engaged in strategies to formatively assess a tutee's understanding than student tutors. This difference in the amount of formative assessments, moreover, accounted for the difference in summative assessment accuracy between teacher tutors and student tutors (cf. article 2).

The results of the first study indicate that tutors, generally, are not very proficient at summatively assessing a tutee's understanding. Nevertheless, tutors do formatively assess a tutee's understanding to some extent. As more formative assessments entailed more learning, it can be assumed that formative assessments, indeed, belong to the mechanisms that make tutoring effective. Enhancing formative assessments, therefore, might make tutoring even more effective. Moreover, the observation that formative assessments led to better summative assessments suggests that fostering a tutor's formative assessments might yield more accurate summative assessments. Finally, teaching experience accounted for better assessments. More precisely, teacher tutors more often employed strategies of formative assessment than did student tutors. This difference also explained why teacher tutors produced more accurate summative assessments than student tutors. Obviously, a more intensive use of strategies to formatively assess a tutee's understanding is an observable indicator of teaching experience.

The second study (cf. *Chapter 2*) showed that tutors, indeed, implemented a more interactive style of tutoring when they were trained in using the interactive instructional strategies of formative assessment that were observed in the first study. However, trained tutors' summative assessments did not become more accurate than the summative assessments of untrained tutors. Instead, the trained tutors were less accurate than the untrained tutors in summatively assessing their tutee's understanding. This unexpected result was explained by the fact that trained tutors more intensively engaged in an interactive style of tutoring than the untrained tutors.

The results of the second study are interpreted with respect to possible deficiencies in cognitive processing on the part of the tutors. The tutors in this study were not experienced in teaching. Thus, implementing the newly learned strategies might have put a

high burden on a tutor's cognitive capacity. As a result, although the trained tutors elicited more information from their tutee by engaging in more interactive tutoring, they might not have been able to process this information appropriately. This interpretation explains why a more interactive style of tutoring failed to result in more accurate summative assessments. Consequently, the design of the training method could be modified. That is, the tutors' processing of information that is gained from the tutee might be enhanced by practicing the strategies of formative assessment more intensively during the training phase. Additionally, the design of the study that had been conducted to analyze the effects of the training method could be changed. Specifically, the timing of summatively assessing a tutee's understanding should be reconsidered. That is, the tutors' processing of information might also be enhanced when the summative assessment is delayed and the tutors get the opportunity to practice the strategies of formative assessment during several sessions of in-vivo tutoring. As a consequence, trained tutors might, indeed, excel untrained tutors in accurately summatively assessing a tutee's understanding.

By analyzing a tutor's assessments, this doctoral thesis contributes to understanding the mechanisms that underlie tutoring. The obtained results can, moreover, be conducive to making tutoring even more effective. This doctoral thesis, thus, provides an important insight into the field of effective instruction.

Zusammenfassung

Das Ziel dieser Dissertation ist es, zur Erforschung von Instruktion beizutragen, welche effektiv das Lernen von Schülerinnen und Schülern unterstützt. Derartige Instruktion leistet ihren Beitrag zu guter Bildung. Da Instruktion verändert werden kann, um Bildung zu verbessern, ist sie als Variable von großem Interesse. Um Instruktion im Sinne guter Bildung verändern zu können, ist es notwendig zu wissen, welche Instruktionsformen effektiv darin sind, das Lernen zu unterstützen. Weiterhin ist es notwendig zu wissen, welche Mechanismen diesen Instruktionsformen zugrunde liegen. Eine bedeutende Form effektiver Instruktion ist das Eins-zu-Eins-Tutoring. Mechanismen, die für die Effektivität des Tutorings als zentral gelten, sind dabei die Diagnosen und die Diagnosegenauigkeit von Tutorinnen und Tutoren. Diese Mechanismen sind bisher jedoch nicht intensiv untersucht worden. Aus diesem Grund werden in der Dissertation die Diagnosen und die Diagnosegenauigkeit von Tutorinnen und Tutoren näher betrachtet. Im Speziellen werden zwei Arten von Diagnosen untersucht. Dies sind erstens Diagnosen, die Tutorinnen und Tutoren fortlaufend während einer Tutoringsitzung durchführen (d. h. formative Diagnosen). Zweitens werden Diagnosen analysiert, die Tutorinnen und Tutoren nach dem Ende einer Tutoringsitzung erstellen (d. h. summative Diagnosen).

Im Zusammenhang mit dieser Dissertation wurden zwei empirische Studien durchgeführt. In beiden Studien wurden Tutandinnen und Tutanden im Schulalter von Tutorinnen und Tutoren unterrichtet, die über mehr Wissen verfügten, als ihre Lernenden. Es wurde angenommen, dass Tutorinnen und Tutoren mit Lehrerfahrung besser darin sind das Verständnis von Tutandinnen und Tutanden zu diagnostizieren als Tutorinnen und Tutoren ohne Lehrerfahrung. In der ersten Studie wurde der Einfluss von Lehrerfahrung auf die Diagnosen von Tutorinnen und Tutoren empirisch überprüft. Zu diesem Zweck wurden die Diagnosegenauigkeit von Tutorinnen und Tutoren mit Lehrerfahrung (d. h. Lehrkräfte) und die Diagnosegenauigkeit von Tutorinnen und Tutoren ohne Lehrerfahrung (d. h. Studierende) untersucht und miteinander verglichen. In diesem Zusammenhang wurde auch die Beziehung zwischen den formativen Diagnosen einer Tutorin oder eines Tutors und den summativen Diagnosen einer Tutorin oder eines Tutors analysiert. Weiterhin wurde der Nutzen der formativen Diagnosen einer Tutorin oder eines Tutors für das Lernen der Tutandin bzw. des Tutanden erforscht (siehe *Kapitel 1*, Artikel 1 und 2). In der zweiten Studie wurde experimentell geprüft, ob die Diagnosen von Studierenden als Tutorinnen und Tutoren durch ein kurzes Training verbessert werden können. Das Training

zielte dabei auf die Förderung eines interaktiven Tutoringstils. Die Idee für das Design des Trainings beruhte darauf, dass Tutandinnen und Tutanden wahrscheinlicher ihr eigenes Verständnis äußern, wenn Tutorinnen und Tutoren einen interaktiven Tutoringstil realisieren. Basierend auf den zusätzlichen Informationen über das Verständnis ihrer Tutandin bzw. ihres Tutanden sollten Tutorinnen und Tutoren mit einem interaktiven Tutoringstil besser in der Lage sein, summativ das Verständnis der Tutandin bzw. des Tutanden zu diagnostizieren als Tutorinnen und Tutoren mit einem weniger interaktiven Tutoringstil (siehe *Kapitel 2*, Artikel 3).

Wie die erste Studie (vgl. *Kapitel* 1) belegt, diagnostizieren Tutorinnen und Tutoren im Durchschnitt das Verständnis ihrer Tutandin oder ihres Tutanden bestenfalls mäßig genau. Allerdings waren Lehrkräfte genauer darin, summativ das Verständnis ihrer Tutandin bzw. ihres Tutanden zu diagnostizieren als Studierende (vgl. Artikel 1). Darüber hinaus zeigte die erste Studie, dass alle Tutorinnen und Tutoren interaktive Instruktionsstrategien einsetzten, um formativ das Verständnis ihrer Tutandin bzw. ihres Tutanden zu diagnostizieren. Mehr formative Diagnosen führten in diesem Zusammenhang zu mehr Lernen. In vergleichbarer Weise zogen mehr formative Diagnosen auch genauere summative Verständnisdiagnosen nach sich. Auch in Bezug auf die formativen Diagnosen unterschieden sich Lehrkräfte von Studierenden. Konkret heißt dies, dass Lehrkräfte häufiger Instruktionsstrategien zum formativen Diagnostizieren des Verständnisses ihrer Tudandin bzw. ihres Tutanden einsetzten als Studierende. Dieser Unterschied im Ausmaß formativer Diagnosen bedingte auch den Unterschied zwischen Lehrkräften und Studierenden bezüglich der summativen Diagnosegenauigkeit (vgl. Artikel 2).

Die Ergebnisse der ersten Studie weisen darauf hin, dass Tutorinnen und Tutoren im Allgemeinen nicht sehr gut darin sind, summativ das Verständnis von Tutandinnen und Tutanden zu diagnostizieren. Dessen ungeachtet diagnostizieren Tutorinnen und Tutoren formativ das Verständnis einer Tutandin bzw. eines Tutanden zumindest in gewissem Maß. Da zudem mehr formative Diagnosen zu mehr Lernen führen, kann angenommen werden, dass diese formativen Diagnosen tatsächlich zu den Mechanismen zählen, die Tutoring effektiv machen. Tutoring könnte somit noch effektiver werden, wenn man das formative Diagnostizieren förderte. Da außerdem beobachtet wurde, dass formative Diagnosen bessere summativen Diagnosen nach sich zogen, kann weiterhin angenommen werden, dass verbessertes formatives Diagnostizieren genauere summative Diagnosen mit sich brächte. Schließlich war auch Lehrerfahrung relevant für bessere Diagnosen. Im Speziellen verwendeten Lehrkräfte häufiger Strategien formativer Diagnose als Studierende. Dieser

Unterschied erklärte auch, warum Lehrkräfte genauer summativ diagnostizierten als Studierende. Offensichtlich kann somit ein intensiverer Gebrauch von Strategien zur formativen Verständisdiagnose als beobachtbarer Indikator für Lehrerfahrung angesehen werden.

Tutorinnen und Tutoren waren durchaus in der Lage, einen interaktiven Tutoringstil zu realisieren, wenn sie darin trainiert wurden, die interaktiven Instruktionsstrategien formativen Diagnostizierens zu verwenden, die in der ersten Studie beobachtet worden waren. Dies wurde in der zweiten Studie festgestellt (vgl. *Kapitel 2*). Allerdings wurden dadurch die summativen Diagnosen der trainierten Tutorinnen und Tutoren nicht genauer als die summativen Diagnosen der untrainierten Tutorinnen und Tutoren. Stattdessen waren die trainierten Tutorinnen und Tutoren weniger genau darin als die untrainierten, summativ das Verständnis ihrer Tutandin bzw. ihres Tutanden zu diagnostizieren. Dieses unerwartete Ergebnis wurde dadurch erklärt, dass die trainierten Tutorinnen und Tutoren stärker als die untrainierten einen interaktiven Tutoringstil realisierten.

Die Ergebnisse der zweiten Studie werden in Bezug auf mögliche Unzulänglichkeiten in der kognitiven Informationsverarbeitung auf Seiten der Tutorinnen und Tutoren interpretiert. Diese Unzulänglichkeiten hängen möglicherweise damit zusammen, dass die Tutorinnen und Tutoren in dieser Studie keine Lehrerfahrung besaßen. Die gerade gelernten Strategien umzusetzen mag daher die kognitive Kapazität der Tutoren stark beansprucht haben. Obwohl die trainierten Tutorinnen und Tutoren mehr Informationen von ihrer Tutandin bzw. ihrem Tutanden gewannen, waren sie in der Folge möglicherweise nicht in der Lage, diese Informationen angemessen zu verarbeiten. Auf diese Weise kann erklärt werden, warum ein interaktiverer Tutoringstil nicht zu genaueren summativen Diagnosen führte. Eine Konsequenz, die aus dieser Interpretation gezogen werden kann, wäre es, das Design des Trainings abzuwandeln. Genauer gesagt, könnten die Tutorinnen und Tutoren die von ihrer Tutandin bzw. ihrem Tutanden gewonnenen Informationen möglicherweise besser verarbeiten, wenn sie die Strategien formativen Diagnostizierens intensiver während der Trainingsphase übten. Zusätzlich könnte das Design der Studie verändert werden, die durchgeführt worden war, um die Trainingseffekte zu analysieren. Spezifischer, würde die Informationsverarbeitung der Tutorinnen und Tutoren möglicherweise auch verbessert, wenn die summative Diagnose später erfolgte und wenn die Tutorinnen und Tutoren so die Gelegenheit erhielten, die Strategien formativen Diagnostizierens während mehrerer realer Tutoringsitzungen zu üben. Im Ergebnis überträfen die trainierten Tutorinnen und Tutoren die untrainierten

eventuell tatsächlich darin, das Verständnis ihrer Tutandin bzw. ihres Tutanden summativ genau zu diagnostizieren.

Durch die Analyse der Diagnosen von Tutorinnen und Tutoren trägt diese Dissertation dazu bei, die dem Tutoring zugrunde liegenden Mechanismen zu verstehen. Die Ergebnisse der Dissertation können weiterhin dazu dienen, Tutoring noch effektiver zu machen. Die Dissertation bietet somit einen wesentlichen Einblick in das Feld effektiver Instruktion.

References

Arnold, D. H., Lonigan, C. J., Whitehurst, G. J., & Epstein, J. N. (1994). Accelerating language development through picture book reading: Replication and extension to a videotape training format. *Journal of Educational Psychology, 86*, 235-243. doi: 10.1037/0022-0663.86.2.235

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology, 29*, 344-370. doi: 10.1016/j.cedpsych.2003.09.002

Baker, S., Gersten, R., & Keating, T. (2000). When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training. *Reading Research Quarterly, 35*, 494-519. doi: 10.1598/rrq.35.4.3

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften [Keyword: Professional competencies of teachers]. *Zeitschrift für Erziehungswissenschaft, 9*, 469-520. Retrieved from: http://link.springer.com/journal/11618

Bell, J. (2001). Tutor training and reflection on practice. *The Writing Center Journal, 21*(2), 78-98. Retrieved from: http://casebuilder.rhet.ualr.edu/wcrp/wcjournal/bibliography.cfm

Belzer, A. (2006). Less may be more: Rethinking adult literacy volunteer tutor training. *Journal of Literacy Research, 38*, 111-140. doi: 10.1207/s15548430jlr3802_1

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18,* 5-25. doi: 10.1080/0969594x.2010.513678

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., … Nickmans, G. (2006). A learning Integrated Assessment System. *Educational Research Review*, 1, 61–67. doi: 10.1016/j.edurev.2006.01.001

Black, P. (1993). Formative and summative assessment by teachers. *Studies in Science Education, 21*, 49-97. doi: 10.1080/03057269308560014

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7-68. doi: 10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability, 21*, 5-31. doi: 10.1007/s11092-008-9068-5

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16. doi: 10.3102/0013189X013006004

Blom-Hoffman, J., O'Neil-Pirozzi, T., Volpe, R., Cutting, J., & Bissinger, E. (2006). Instructing parents to use dialogic reading strategies with preschool children: Impact of a video-based training program on caregiver reading behaviors and children's related verbalizations. *Journal of Applied School Psychology, 23*, 117-131. doi: 10.1300/J370v23n01_06

Borko, H., & Putnam, R. T. (1996). Learning to teach. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 673-707). New York, NY: Macmillan.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*, 259-278. doi: 10.1016/s0742-051x(96)00024-8

Bortz, J., & Schuster, C. (2010). Statistik für Human- und Sozialwissenschaftler [Statistics for scientists in human sciences and social sciences] (7th rev. ed.). Heidelberg, Germany: Springer.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice, 31*, 13-17. doi: 10.1111/j.1745-3992.2012.00251.x

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*, 3-12. doi: 10.1111/j.1745-3992.2010.00195.x

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher, 10*(2), 14-21. doi: 10.3102/0013189X010002014

Cade, W., Copeland, J., Person, N., & D'Mello, S. (2008). Dialogue modes in expert tutoring. In B. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent*

tutoring systems (Vol. 5091, pp. 470-479): Heidelberg, Germany: Springer. Retrieved from http://link.springer.com/

Casey, L. B., & Williamson, R. (2011). Training parents as effective literacy tutors: Increasing the procedural integrity of tutoring. *Mentoring & Tutoring: Partnership in Learning, 19*(3), 257-276. doi: 10.1080/13611267.2011.597118

Chae, H. M., Kim, J. H., & Glass, M. (2005). Effective behaviors in a comparison between novice and expert algebra tutors. In S. Hettiarachchi & R. Finkbine (Eds.), *Proceedings of the Sixteenth Midwest Artificial Intelligence and Cognitive Science Conference.* Retrieved from: http://www.mglass.org/papers/papers.html

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 21-30). Cambridge, MA: Cambridge University Press.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73-105. doi: 10.1111/j.1756-8765.2008.01005.x

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477. doi: 10.1207/s15516709cog1803_3

Chi, M. T. H., & Roy, M. (2010). How adaptive is an expert human tutor? In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems* (Vol. 6094, pp. 401-412) Heidelberg, Germany: Springer. doi: 10.1007/978-3-642-13388-6_44

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32,* 301-341. doi: 10.1080/03640210701863396

Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22*, 363-387. doi: 10.1207/s1532690xci2203_4

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533. doi: 10.1207/s15516709cog2504_1

Chow, B. W.-Y., & McBride-Chang, C. (2003). Promoting language and literacy development through parent-child reading in Hong Kong preschoolers. *Early Education and Development, 14*, 233-248. doi: 10.1207/s15566935eed1402_6

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Education outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237-248. doi: 10.2307/1162567

Corno, L., & Snow, R. E. (1986). Adapting teaching to individual differences among learners. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 605-628). New York, NY: Macmillan.

Cromley, J. G., & Azevedo, R. (2005). What do reading tutors do? A naturalistic study of more and less experienced tutors in reading. *Discourse Processes, 40,* 83-113. doi: 10.1207/s15326950dp4002_1

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100,* 223-234. doi: 10.1037/0022-0663.100.1.223

Dignath, C., Buettner, G., & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review, 3,* 101-129. doi: 10.1016/j.edurev.2008.02.003

Drechsel, B., Prenzel, M., & Seidel, T. (2009). Nationale und internationale Schulleistungsstudien [National and international student assessment studies]. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (pp. 353-382). Heidelberg, Germany: Springer.

Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung. Eine experimentelle Studie zu Ankereffekten [Biases in teachers' assessments of student performance: An experimental study of anchoring effects]. *Zeitschrift für Pädagogische Psychologie*, 23, 187-195. doi: 10.1024/1010-0652.23.34.187

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83-87. doi: 10.1111/1467-8721.01235

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading

failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*, 605-619. doi: 10.1037/0022-0663.92.4.605

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406. doi: 10.1037/0033-295x.100.3.363

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research, 102*, 453-462. doi: 10.3200/joer.102.6.453-462

Feldon, D. F. (2007). Cognitive load and classroom teaching: The double-edged sword of automaticity. Educational Psychologist, 42, 123-137. doi: 10.1080/00461520701416173

Fisseni, H.-J. (1997). Lehrbuch der psychologischen Diagnostik mit Hinweisen zur Intervention [Textbook of psychological diagnostics with references to intervention] (2nd rev ed.). Göttingen, Germany: Hogrefe.

Friedrich, H. F., & Mandl, H. (1992). Lern- und Denkstrategien - ein Problemaufriß [Learning and thinking strategies – outline of the problem]. In H. Mandl & H. F. Friedrich (Eds.), *Lern- und Denkstrategien. Analyse und Intervention* (pp. 3-54). Göttingen, Germany: Hogrefe.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*, 177-182. doi: 10.1177/0963721412445309

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*, 360-389. doi: 10.1080/08957340802347852

Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal, 89*, 253-300. doi: 10.1086/461577

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In S. Graham & K. Harris (Eds.), *APA educational psychology handbook* (Vol. 3, pp. 451-473). Washington, DC: American Psychological Association.

Graesser, A. C., D'Mello, S., & Cade, W. L. (2011). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 408-426). London, United Kingdom: Routledge.

Graesser, A. C., D'Mello, S., & Person, N. K. (2009). Meta-knowledge in tutoring. In A. C. Graesser, D. J. Hacker, & J. Dunlosky (Eds.), *Handbook of metacognition in education* (pp. 362-382). London, United Kingdom: Routledge.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 495-522. doi: 10.1002/acp.2350090604

Haag, L. (2010). Nachhilfeunterricht [Private tutoring]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (4th rev. ed., pp. 591-599). Weinheim, Germany: Beltz.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*, 365-379. doi: 10.1080/0969594970040304

Hattie, J. (2009). Visible learning. A synthesis of over 800 meta-analyses relating to achievement. London, United Kingdom: Routledge.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013a). *Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment.* Manuscript submitted for publication.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013b). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*, 242-260. doi: 10.1080/00220973.2012.699900

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (in press). Benefits for processes cause decrements in outcomes: Training improves tutors' interactivity at the expense of assessment accuracy. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Hesse, I., & Latzko, B. (2011). Diagnostik für Lehrkräfte [Diagnostics for teachers] (2nd rev. ed.). Opladen, Germany: Barbara Budrich.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied, 5*, 205-221. doi: 10.1037/1076-898X.5.2.205

Hmelo-Silver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-based Learning*, *1*(1), 21-39. doi: 10.7771/1541-5015.1004

Hofer, M. (1986). Sozialpsychologie erzieherischen Handelns. Wie das Denken und Verhalten von Lehrern Organisiert ist [Social psychology of educational actions. How teachers‘ thinking and behavior is organized]. Göttingen, Germany: Hogrefe.

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

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539. doi: 10.1007/s10648-007-9054-3

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science, 7,* 46-50. doi: 10.1111/1467-8721.ep13175613

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*, 28-37. doi: 10.1111/j.1745-3992.2011.00220.x

Kingston, N., & Nash, B. (2012). How many formative assessment angels can dance on the head of a meta-analytic pin: .2. *Educational Measurement: Issues and Practice*, *31*, 18-19. doi: 10.1111/j.1745-3992.2012.00254.x

Klauer, K. J. (1988). Teaching for learning-to-learn: A critical appraisal with some proposals. *Instructional Science, 17,* 351-367. doi: 10.1007/bf00056221

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for Effective Intervention, 29*(4), 59-67. doi: 10.1177/073724770402900408

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38-46. doi: 10.1016/j.tate.2012.10.004

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess. Der Lehrer als “flexibler Denker” [Goal-directed processing of students’ attributes: The teacher as “flexible thinker”]. *Zeitschrift für Pädagogische Psychologie, 23*, 175-186. doi: 10.1024/1010-0652.23.34.175

Lehman, B., D’Mello, S., Cade, W., & Person, N. (2012). How do they do it? Investigating dialogue moves within dialogue modes in expert human tutoring. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol.

7315, pp. 557-562). Heidelberg, Germany: Springer. doi: 10.1007/978-3-642-30950-2_72

Leutner, D. (2010). Instruktionspsychologie [Instructional psychology]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (4th rev. ed., pp. 289-298). Weinheim, Germany: Beltz.

Leutner, D., Leopold, C., & Den Elzen-Rump, V. (2007). Self-regulated learning with a text-highlighting strategy: A training experiment. *Zeitschrift für Psychologie, 215*, 174-182. doi: 10.1027/0044-3409.215.3.174

Lipowsky, F. (2009). Unterricht [Instruction]. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (pp. 73-102). Heidelberg, Germany: Springer.

Lonigan, C. J., & Whitehurst, G. J. (1998). Relative efficacy of parent and teacher involvement in a shared-reading intervention for preschool children from low-income backgrounds. *Early Childhood Research Quarterly, 13*, 263-290. doi: 10.1016/s0885-2006(99)80038-6

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N.G. Lederman (Eds.), *Examining pedagogical content knowledge: The construct and its implications for science education* (pp. 95–132). Boston, MA: Kluwer.

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*, 78-102. doi: 10.1080/10627190903039429

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction, 7,* 197-224. doi: 10.1207/s1532690xci0703_2

Mehan, H. (1979). *Learning lessons: social organization in the classroom*. Cambridge, MA: Harvard University Press.

Michael, J. A., Wenderoth, M. P., Modell, H. I., Cliff, W., Horwitz, B., McHale, P., … Whitescarver, S. (2002). Undergraduates' understanding of cardiovascular phenomena. *Advances in Physiology Education, 26*, 72–84. doi: 10.1152/advan.00002.2002

Mulholland, L. A., & Berliner, D. C. (1992). *Teacher experience and the estimation of student achievement.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. Retrieved from: http://www.eric.ed.gov/PDFS/ED348350.pdf

Mol, S. E., Bus, A. G., de Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent-child book readings: A meta-analysis. *Early Education and Development, 19*, 7-26. doi: 10.1080/10409280701838603

Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal, 40*, 905-928. doi: 10.3102/00028312040004905

Nickerson, R. S. (1999). How we know – and sometimes misjudge – what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*, 737-759. doi: 10.1037/0033-2909.125.6.737

Organisation for Economic Co-operation and Development (OECD) (2013, May). OECD Programme for International Student Assessment (PISA) [Homepage]. Retrieved from http://www.oecd.org/pisa

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*, 5-13. doi: 10.1111/j.1745-3992.2009.00149.x

Pressley, M., Goodchild, F., Fleet, J., Zajchowski, R., & Evans, E. D. (1989). The challenges of classroom strategy instruction. *The Elementary School Journal, 89*, 301-342. doi: 10.1086/461578

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*, 13-48. doi: 10.3102/00028312024001013

Randell, T., Hall, M., Bizo, L., & Remington, B. (2007). DTkid: Interactive simulation software for training tutors of children with autism. *Journal of Autism and Developmental Disorders, 37*, 637-647. doi: 10.1007/s10803-006-0193-z

Renkl, A. (2005). The worked-out examples principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 229-245). Cambridge, MA: Cambridge University Press.

Rheinberg, F. (2002). *Motivation* [Motivation] (4th rev. ed.). Stuttgart, Germany: Kohlhammer.

Riemeier, T., Jankowski, M., Kersten, B., Pach, S., Rabe, I., Sundermeier, S., & Gropengießer, H. (2010). Wo das Blut fließt. Schülervorstellungen zu Blut, Herz und Kreislauf beim Menschen [Where our blood flows. Students' conceptions about

blood, heart and circulation]. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 77-93. Retrieved from http://gandalf.ipn.uni-kiel.de/zfdn/

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research, 79*, 3-38. doi: 10.3102/0034654308325690

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*, 57-84. doi: 10.1002/tea.20163

Schmidt, H. G., & Moust, J. H. C. (1995). What makes a tutor effective? A structural-equations modeling approach to learning in problem-based curricula. *Academic Medicine, 70*, 708-714. Retrieved from http://journals.lww.com/academicmedicine/pages/default.aspx

Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern [Diagnostic competencies of parents and teachers]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (4th rev. ed., pp. 102-108). Weinheim, Germany: Beltz.

Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology, 99*, 285-296. doi: 10.1037/0022-0663.99.2.285

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). *Standards für die Lehrerbildung: Bildungswissenschaften* [Standards for teacher education: Educational siences] (Beschluss der Kultusministerkonferenz vom 16.12.2004). Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

Shavelson, R. J. (2006). *On the integration of formative assessment in teaching and learning with implications for teacher education*. Paper prepared for the Stanford Education Assessment Laboratory and the University of Hawaii Curriculum Research and Development Group. Retrieved from: http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/Paper.htm

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., … Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment

developers. *Applied Measurement in Education, 21*, 295-314. doi: 10.1080/08957340802347647

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *The handbook of research on teaching* (4th ed., pp. 1066–1101). Washington, D.C.: American Educational Research Association.

Smith, P. L., & Ragan, T., J. (2005). *Instructional design. Hoboken*, NJ: Wiley.

Snow, R. E., & Swanson, J. (1992). Instructional psychology: Aptitude, adaptation, and assessment. *Annual Review of Psychology, 43*, 583-626. doi: 10.1146/annurev.ps.43.020192.003055

Song, K.-O., Park, H.-J., & Sang, K.-A. (2013). A cross-national analysis of the student- and school-level factors affecting the demand for private tutoring. *Asia Pacific Education Review*. Advance online publication. doi: 10.1007/s12564-012-9236-7

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. Journal of Educational Psychology, 104, 743-762. doi: 10.1037/a0027627

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der simulierte Klassenraum. Eine experimentelle Untersuchung zur diagnostischen Kompetenz [The simulated classroom: An experimental study on diagnostic competence]. *Zeitschrift für Pädagogische Psychologie, 22*, 261-276. doi: 10.1024/1010-0652.22.34.261

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education, 32*, 321-345. doi: 10.1007/BF00138870

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25*, 631-645. doi: 10.1080/01443410500345172

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. doi: 10.1126/science.185.4157.1124

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

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197-221. doi: 10.1080/00461520.2011.611369

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*, 1-60. doi: 10.1080/03640210709336984

VanLehn, K., Siler, S. A., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*, 209-249. doi: 10.1207/s1532690xci2103_01

Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teaching and Teacher Education, 25*, 1051-1060. doi: 10.1016/j.tate.2009.04.002

Vosniadou, S. (1999). Conceptual change research: State of the art and future directions. In W. Schnotz, S. Vosniadou, & M. Carretero (Eds.), *New perspectives on conceptual change* (pp. 3-13). Amsterdam, The Netherlands: Pergamon.

Vosniadou, S., Vamvakoussi, X., & Skopeliti, I. (2008). The framework theory approach to the problem of conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 3-34). London, United Kingdom: Taylor and Francis.

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring - a review of 5 programs. *Reading Research Quarterly, 28*, 179-200. Retrieved from http://www.jstor.org/stable/747888

Weinert, F. E., Helmke, A., & Schrader, F.-W. (1992). Research on the model teacher and the teaching model. In F. K. Oser, A. Dick, & J.-L. Patry (Eds.), *Effective and responsible teaching: The new synthesis* (pp. 249-260). San Francisco, CA: Jossey-Bass.

Weinert, F.E. & Schrader, F.-W. (1986). Diagnose des Lehrers als Diagnostiker [Diagnosis of the teacher as diagnostician]. In H. Petillon, J., Wagner, & B. Wolf (Eds.), *Schülergerechte Diagnose. Theoretische und empirische Beiträge zur Pädagogischen Diagnostik* (pp. 11–29). Weinheim, Germany: Beltz.

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology, 24*, 552-559. doi: 10.1037/0012-1649.24.4.552

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

Wittwer, J., Nückles, M., Landmann, N., & Renkl, A. (2010). Can tutors be supported in giving effective explanations? *Journal of Educational Psychology, 102*, 74-89. doi: 10.1037/a0016727

Wittwer, J., Nückles, M., & Renkl, A. (2010). Using a diagnosis-based approach to individualize instructional explanations in computer-mediated communication. *Educational Psychology Review, 22*, 9-23. doi: 10.1007/s10648-010-9118-7

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

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., … Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*, 335-359. doi: 10.1080/08957340802347845

Zimmerman, B. J. (2006). Development and adaptation of expertise: The role of self-regulatory processes and beliefs. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 705-722). Cambridge, MA: Cambridge University Press.

Acknowledgements

First of all, I would like to express my sincere gratitude to Prof. Dr. Jörg Wittwer for giving me the opportunity to work on this inspiring research project. Without his enduring support, patience and enthusiasm throughout the last years this doctoral thesis would not have been possible. I also owe many thanks to Prof. Dr. Susanne Bögeholz for her guidance and advice and for the fruitful discussions about my work across interdisciplinary boundaries.

Furthermore, I would like to thank Prof. Dr. Hannes Rakoczy, who is a member of the examination board, for his support during the time of my doctoral project. I am, moreover, grateful to Prof. Dr. Matthias Nückles and Prof. Dr. Alexander Renkl for sharing knowledge and ideas at many stages of the project. I would also like to thank both of them for polishing the articles this doctoral thesis is based on and Prof. Dr. Nückles for becoming a member of the examination board. In addition, I want to extend my thanks to Prof. Dr. Roland H. Grabner and Prof. Dr. Michael R. Waldmann for becoming members of the examination board. I thank Prof. Dr. Nele McElvany who drew my interest to the study of instruction and learning in the first place.

It is a pleasure to thank my colleagues in Kiel, Göttingen, Freiburg and elsewhere for stimulating discussions, interesting exchanges of ideas, for proofreading, and for the inspiring working atmosphere. Special thanks go to my colleague Natalie Ihme for her constant support and input and the great time of being fellow doctoral students.

I further would like to thank the student assistants who participated in the research performed in the context of this doctoral thesis for helping me with many practical aspects of the project. Thank you, Hannah, Victoria, Julian, Angela, Imme, Sophia, Rico, Amelie, Annette, Karina, Bosse, Anne-Kristin, Tatjana, Eva, Lisa, and Raoul. I have greatly enjoyed working with all of you.

Of course, I also owe many thanks to everyone who participated as tutor or as tutee in the studies conducted for the project. Furthermore, the German Research Foundation (DFG: WI 3348/2-1) kindly provided financial support.

Last but not least, I would like to thank my friends and, in particular, my family for their continuous support, encouragement, and motivation.

Curriculum Vitae

Personal Information

Name:	Stephanie Herppich
Date of birth:	November 24, 1982
Place of birth:	Bayreuth
Nationality:	German

Education and Research Experience

Since 03/2011	Doctoral Student and Research Associate at the University of Göttingen, Germany, Educational Institute, Empirical Educational Research on Instruction and Learning, in the project: *Conditions and promotion of the diagnosis of misconceptions in tutoring* (DFG), and subsequently in the project: *Diagnostic expertise of instructors: Development and validation of a test measuring knowledge about pedagogical diagnostics* (*PAEDDI*; PRO*Niedersachsen)
05/2009 – 02/2011	Doctoral Student and Research Associate at the Leibniz Institute for Science and Mathematics Education (IPN) at the University of Kiel, Germany, Department of Educational Science, in the project: *Conditions and promotion of the diagnosis of misconceptions in tutoring* (DFG: WI 3348/2-1)
04/2009	Diplom in Psychology, University of Potsdam, Germany, Thesis: *Mütterliches Steuerungsverhalten in der Essenssituation: Eine Beobachtungsstudie zum Vergleich zwischen über- und normalgewichtigen Müttern*
12/2006 – 03/2009	Student Assistant at the Max Planck Institute for Human Development, Berlin, Germany, Center for Educational Research
09/2005 – 02/2006	Semester abroad at the University of Leuven, Belgium
10/2004 – 02/2005	Student Assistant at University of Potsdam, Germany, Research Methods in Human Sciences
10/2002 – 03/2009	Studies of Psychology at the University of Potsdam, Germany
06/2002	Abitur, Ernst-Haeckel-Gymnasium, Werder/Havel, Germany

Teaching

WS 12/13	Proseminar: Lehr- (und Lern-)Methoden (Teaching (and Learning) Methods)
SS 12	Proseminar: Methoden des Lehrens und Lernens im Klassenzimmer (Methods of Teaching and Learning in the Classroom; 2 Courses)
WS 11/12	Proseminar: Verfahren und Methoden der Diagnostik von Schülerinnen und Schülern (Techniques and Methods of Diagnosing Students)

Grants and Awards

2012	Best Paper Award of the JURE (Junior Researchers of the European Association for Research on Learning and Instruction) 2012 Conference, Regensburg, Germany
2011	German Academic Exchange Service (DAAD) travel grant supporting attendance of the 33rd Annual Conference of the Cognitive Science Society, Boston, MA, USA

Publications

Peer Reviewed Publications

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (in press). Benefits for processes cause decrements in outcomes: Training improves tutors' interactivity at the expense of assessment accuracy. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*, 242-260. doi: 10.1080/00220973.2012.699900

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2011). Does teaching experience help? Differences in the assessment of tutees' understanding between teacher tutors and student tutors. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 78-83). Austin, TX: Cognitive Science Society. Retrieved from: http://csjarchive.cogsci.rpi.edu/Proceedings/2011/papers/0019/paper0019.pdf

Van Steensel, R., McElvany, N., Kurvers, J., & Herppich, S. (2011). How effective are family literacy programs: Results of a meta-analysis. *Review of Educational Research, 81*, 69-96. doi:10.3102/0034654310388819

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2010). Do tutors' content knowledge and beliefs about learning influence their assessment of tutees' understanding? In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 314-319). New York, NY: Erlbaum. Retrieved from: http://mindmodeling.org/cogsci2010/papers/0052/paper0052.pdf

McElvany, N., Herppich, S., Van Steensel, R. & Kurvers, J. (2010). Wirksamkeit familiärer Frühförderungsprogramme im Bereich Literacy [On the effectiveness of family programs of preschool education in the field of literacy - results of a meta-analysis]. *Zeitschrift für Pädagogik, 56*, 178-192. Retrieved from: http://www.fachportal-paedagogik.de/start.html

Book Chapters

McElvany, N., Van Steensel, R., Guill, K., Van Tuijl, C., & Herppich, S. (2012). Family literacy programs in the Netherlands and in Germany: Policies, current programs, and evaluation studies. In B. Wasik & B. Van Horn (Eds.), *Handbook on family literacy* (2nd ed., pp. 339-353). London, United Kingdom: Routledge.

Van Steensel, R., Herppich, S., McElvany, N., & Kurvers, J. (2012). How effective are family literacy programs for children's literacy skills? A review of the meta-analytic evidence. In B. Wasik & B. Van Horn (Eds.), *Handbook on family literacy* (2nd ed., pp. 135-148). London, United Kingdom: Routledge.

Overview of Articles

The following articles are part of this doctoral thesis:

Article 1 (cf. *Chapter 1*):

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013b). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*, 242-260. doi: 10.1080/00220973.2012.699900

This article has been accepted by *The Journal of Experimental Education.* It has been published on January 1, 2013. Copyright © 2013 by Taylor & Francis Ltd. Reproduced with permission. The official citation that should be used in referencing this material is Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. *The Journal of Experimental Education, 81*, 242-260. doi:10.1080/00220973.2012.699900. The article is available online at http://www.tandfonline.com/10.1080/00220973.2012.699900

Article 2 (cf. *Chapter 1*):

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013a). *Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment.* Manuscript submitted for publication.

This article has been submitted to the *Journal of Educational Psychology*.

Note. At the time this doctoral thesis was published, a revised version of article 2 had been accepted and published by the *Journal of Educational Psychology*. Copyright © 2014 by the American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2014, March 10). Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment. *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/a0036076. This article may not exactly replicate the final version published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written

permission from the American Psychological Association. The article is available online at http://dx.doi.org/10.1037/a0036076

Article 3 (cf. *Chapter 2*):

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (in press). Benefits for processes cause decrements in outcomes: Training improves tutors' interactivity at the expense of assessment accuracy. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

This article has been accepted for presentation as poster at the 35th Annual Conference of the Cognitive Science Society and for publication in the *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Acceptance was based on a strict review process.

Note. At the time this doctoral thesis was published, article 3 had been published in the *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. The official citation that should be used in referencing this material is Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Benefits for Processes Cause Decrements in Outcomes: Training Improves Tutors' Interactivity at the Expense of Assessment Accuracy. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2530-2535). Austin, TX: Cognitive Science Society. The version of the article printed in this doctoral thesis might not exactly replicate the final version published in the proceedings. The article is available online at http://mindmodeling.org/cogsci2013/papers/0458/paper0458.pdf

Statement of Originality and Description of Own Contributions to the Publications

(Erklärung über eigene Leistungen und Darstellung der geleisteten Eigenanteile an den Publikationen)

I hereby declare that this doctoral thesis is my own work and that I have strictly complied with all conditions stated in the *Promovierenden-Erklärung der Georg-August-Universität Göttingen (Anlage 1 zu § 4 Abs. 7 der RerNat-O vom 04.07.2012).*

This doctoral thesis is based on two studies (cf. *Chapter 1* and *Chapter 2*) conducted in the context of a research project that was funded by the German Research Foundation (DFG: WI 3348/2-1). The DFG-project was granted to Prof. Dr. Jörg Wittwer together with Prof. Dr. Matthias Nückles and Prof. Dr. Alexander Renkl. The DFG-project was supervised by Prof. Dr. Wittwer. Prof. Dr. Wittwer also supervised all work for my doctoral project leading to this doctoral thesis.

In the context of my doctoral project, I added conceptual aspects to the general research questions of the DFG-project. I substantiated the research questions covered by the two studies, I added to the variables measured, and I independently analyzed and interpreted the obtained results. More specifically, the extensions that I made to the general aims of the DFG-project refer to (1) the investigation of the relationships between formative assessments and summative assessments of tutors and of the relationships between formative assessments of tutors and the tutees' learning, (2) the design and the content of the training method to enhance a tutor's assessment accuracy, and (3) the conceptual and statistical analyses of the indirect relationships between tutors, process measures, and effect measures.

I independently developed the material for both studies. In doing so, I was assisted by student assistants of the Leibniz Institute for Science and Mathematics Education at the University of Kiel (IPN) and of the Georg-August University Göttingen. In particular, Raoul Zimmermann programmed the IT environment for training tutors in the second study. Furthermore, I independently conducted both studies. Again, I was assisted by student assistants. I wrote all three articles that this doctoral thesis is based on as first author (cf. *Chapter 1* and *Chapter 2*). All co-authors contributed to finalizing all three articles.

Göttingen, May 31, 2013 Stephanie Herppich