Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 9 (2017), No. 3, 82–101

Anne B. Wiesbeck, Johannes Bauer, Martin Gartmeier,
Claudia Kiessling, Grit E. Möller, Gudrun Karsten,
Martin R. Fischer & Manfred Prenzel

Simulated conversations for assessing professional conversation competence in teacher-parent and physician-patient conversations

Abstract

Simulated conversations (SC) with trained actors are a performance-oriented method for assessing communicative competences in authentic task situations. This study evaluated the psychometric properties of parallel designed SC in a cross-professional setting: In teacher-parent and physician-patient conversations. Specifically, we addressed three research questions regarding the reliability and construct validity of the SC: (1) whether trained observers reach a satisfactory

Dr. Anne B. Wiesbeck, TUM School of Education, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany
e-mail: anne.wiesbeck@tum.de

Prof. Dr. Johannes Bauer (corresponding author), Chair for Educational Research and Methods, University of Erfurt, Nordhäuser Str. 63, 99089 Erfurt, Germany
e-mail: johannes.bauer@uni-erfurt.de

Dr. Martin Gartmeier, TUM Medical Education Center, Technical University of Munich, Nigerstr. 3, 81675 Munich, Germany
e-mail: martin.gartmeier@tum.de

PD Dr. Claudia Kiessling, Coordinator, Assessment and Examination Organization, Brandenburg Medical School, Fehrbelliner Straße 38, 16816 Neuruppin, Germany
e-mail: claudia.kiessling@mhb-fontane.de

Grit E. Möller, Faculty of Medicine, CAU Kiel, Christian-Albrechts-Platz 4, 24098 Kiel, Germany (at time of study)
e-mail: moeller@communication-modell.de

Dr. Gudrun Karsten, Faculty of Medicine, CAU Kiel, Arnold-Heller-Straße 3, 24105 Kiel, Germany
e-mail: karsten.dekanat@med.uni-kiel.de

Prof. Dr. Martin R. Fischer, Chair for Medical Education, LMU Munich, Ziemsenstr. 1, 80336 Munich, Germany
e-mail: martin.fischer@med.uni-muenchen.de

Prof. Dr. Manfred Prenzel, TUM School of Education, TU Munich, Arcisstr. 21, 80333 Munich, Germany
e-mail: manfred.prenzel@tum.de

interrater reliability in rating examinees performance; (2) whether correlations among three types of ratings (external observers', SC partners', and students' self-ratings) correspond to expectations; and (3) whether hypothesized correlations with external criteria (prior communication training, semester of study, high school grade point average) could be found. To answer these questions, n = 72 undergraduate medical students and n = 96 pre-service teachers conducted SC. Results showed sufficient interrater reliability (ICC = 0.71). Moreover, the pattern of correlations among the observer ratings, the other two types of ratings, and external criteria emerged as expected. These results provide evidence for the reliability and validity of the developed SC assessment.

Keywords
Simulated conversations; Assessment; Teacher education; Medical education; Communication competence

Simulierte Gespräche als Instrument zur Messung professioneller Gesprächsführungskompetenz in Lehrer-Eltern- und Arzt-Patienten-Gesprächen

Zusammenfassung
Simulierte Gespräche mit trainierten Schauspielern sind eine performanzorientierte Methode zur Erfassung kommunikativer Kompetenzen. In der vorliegenden Studie analysierten wir die Reliabilität und Konstruktvalidität eines parallel entwickelten Assessments mit simulierten Gesprächen in einem domänenübergreifenden Setting: In Lehrer-Eltern- und Arzt-Patienten-Gesprächen. Dabei untersuchten wir drei Fragestellungen: (1) ob trainierte Beobachter eine hinreichende Interrater-Reliabilität bei der Einschätzung der Performanz der Teilnehmenden erzielen; (2) ob die Korrelationen zwischen den Ratings der Beobachter, der eingesetzten Schauspieler und der Selbsteinschätzung der Probanden ein erwartetes Muster aufweisen; und (3) ob sich vermutete Korrelationen zu externen Variablen (vorausgehende Kommunikationstrainings, Abiturnote, Studiensemester) nachweisen lassen. Hierfür nahmen n = 72 Medizinstudierende und n = 96 Lehramtsstudierende an je zwei simulierten Gesprächen teil. Die Ergebnisse zeigten eine insgesamt zufriedenstellende Interrater-Reliabilität (ICC = 0.71). Zudem fielen die Korrelationen zwischen den verschiedenen Ratings und den Außenkriterien erwartungskonform aus. Diese Ergebnisse liefern Hinweise für die Reliabilität und Konstruktvalidität des entwickelten Assessments.

Schlagworte
Simulierte Gespräche; Assessment; Lehrerbildung; Medizinerausbildung; Kommunikationskompetenz

1. Introduction

Conducting professional conversations is an important task in many professional domains (Hargie, 2011; Association of Standardized Patient Educators, 2017). Consequentially, current frameworks for qualifications in higher education highlight the development of communicative competences (e.g., KMK, 2005). Specifically, professional conversation competence (PCC) and respective training programs have grown in importance for many fields of professional education. This is particularly true for medical education, as indicated by an increasing body of research (for an overview see e.g., Association of Standardized Patient Educators, 2017). In other fields, like teacher education, such research is only emerging, despite common emphasis on the importance of preparing students for professional conversations, particularly with parents (e.g., Aich, 2011; Dotger, 2013; Gartmeier et al., 2015; Hertel, 2009; Wiesbeck, 2015).

Assessing the outcomes of such training and providing feedback to learners requires reliable and valid measures of PCC. For this purpose, a specific challenge is to create performance-oriented assessment methods that measure participants' communicative behavior in authentic task situations (Braun, Athanassiou, & Pollerhof, 2016; cf. Blömeke, Gustafson & Shavelson, 2015; Shavelson, 2013). Simulated conversations (SC), as established in medical education, are a promising method for this purpose (e.g., Lane & Rollnick, 2007; for more information on simulated patients, see Association of Standardized Patient Educators, 2017; Barrows & Abrahamson, 1964). In SC, examinees lead a simulated professional conversation about a pre-defined authentic case scenario with actors trained to portray a standardized role. For example, a patient seeking medical advice on different treatment options. Beyond medical training, SC have started to expand into other professional domains, such as teacher education (Dotger, Harris, & Hansel, 2008; Gerich & Schmitz, 2016). However, while medical research maintains substantial evidence for the psychometric quality of SC (e.g., Cleland, Abe, & Rethans, 2009; Newble, 2004), such research is lacking in teacher education.

The present study contributes to closing this gap by investigating the interrater reliability and aspects of construct validity of SC for assessing conversation competence in a cross-professional setting: In teacher-parent conversations in education (cf. Chrispeels & Coleman, 1996; Jeynes, 2011; Kreider, Caspe, Kennedy, & Weiss, 2007) and physician-patient conversations in medicine (cf. Bennett, Fuertes, Keitel, & Phillips, 2011; Street, Makoul, Arora, & Epstein, 2009). The study is part of a larger project targeting the development of parallel-designed training programs to foster initial competence of shared decision-making conversations in medical and teacher education (Gartmeier, Bauer, Fischer, Karsten, & Prenzel, 2011; Gartmeier et al., 2015). We chose this interdisciplinary perspective, not only to build on existing research within the medical domain, but also to investigate domain-general questions. This cross-domain perspective may seem unusual at first glance, given that teachers and physicians have rather distinct occupations. This notwithstand-

ing, the review by Berkhof, van Rijssen, Schellart, Anema, and van der Beek (2011) suggests that cross-domain studies have an added value for the generalization of results and for answering questions about the domain-generality vs. specificity of communication skills, their training, and assessment (cf. Braun et al., 2016). Despite the considerable differences between teaching and medicine, a closer look reveals that *structurally* equivalent conversation types – i.e. with comparable communicative goals and procedures – exist in both professions. Several typical tasks shared by both physicians and teachers include counseling clients[1] (e.g., patients, parents or students for making a well-informed decision on a problem), delivering bad news (e.g., about a medical diagnosis or a student's achievement problems), or resolving conflicts (e.g., about compliance with a medical treatment or disruptive student behavior).

The present study focuses on one of these particular situations that frequently occur within both domains, namely shared decision-making conversations (Medicine: Makoul & Clayman, 2006; Teaching: Aich, 2011; Hanafin & Lynch, 2002; Staples & Diliberto, 2010). In such conversations, the goal is to reach a mutual decision regarding a problem for which several viable solutions must be weighed (Charles, Gafni, & Whelan, 1999; Loh, Simon, Kriston, & Härter, 2007). Shared decision-making conversations are special cases of expert-layperson[2] communication (Bromme, Nueckles, & Rambow, 1999) and counseling (Bamberger, 2015; Burks & Stefflre, 1979; Rogers, 1951). To lead such conversations competently, a core of generic communication skills seem to apply across many professional domains and types of conversations. Such generic skills include: Creating common ground to advance a problem solution, establishing a positive interpersonal relationship, or structuring the conversation proactively and solution-oriented through meta-communication (Aich, 2011; Beck & Daughtridge, 2002; Bruder, 2011; Gartmeier et al., 2015; Hargie, 2011; Hertel, 2009; Clark & Brennan, 1991; Lawrence-Lightfoot, 2004; Rogers, 1951; Weisbach & Sonne-Neubacher, 2005; Wiesbeck, 2015). Developing reliable and valid measures for such generic aspects of PCC can provide a basis for further research and training practices, as well as measuring communicative competences in higher education.

Below, we first elaborate on the theoretical background of the PCC model, which provided the foundation for constructing the SC. Subsequently, we outline the main findings on SC as an assessment method from existing research.

1.1 Professional conversation competence

PCC is typically conceived as a sub-aspect of the broader concept of individual communication competence (e.g., Berger, 2008; Traut-Mattausch & Frey, 2006).

1 We use the term "clients" broadly here, though the term is not frequently used in the school context.

2 In this context, experts are defined as persons who – unlike laypersons – can rely on professional education and related job experience in their domain (Bromme et al., 2003).

PCC can be defined as personal dispositions (skills, knowledge, attitudes, etc.) that allow professionals to attain communicative goals in a given professional conversation (Weisbach & Sonne-Neubacher, 2005). As mentioned, the underlying model of PCC in this study is based on theoretical approaches of expert-layperson communication (Bromme et al., 1999; Bromme, Jucks, & Rambow, 2003) and counselling (Bamberger, 2015; Burks & Stefflre, 1979; Rogers, 1951). Both approaches describe key-challenges to problem-oriented conversations between professionals and clients. Specifically, expert-layperson communication involves the effort of sufficiently aligning two individual cognitive frames of reference in order to establish adequate common ground for reaching the specific goal of a conversation (Bromme, Jucks, & Rambow., 2004; Clark & Brennan, 1991). Below, we describe three crucial communicative skills the professional requires to lead such conversations effectively (cf. Gartmeier et al., 2011, 2015).

Advancing a problem solution refers to communicative techniques that increase the likelihood of meeting the client's concern(s) and developing a promising solution. In this respect, one major task for the professional is to establish common ground with the conversational partner (Bromme et al., 1999, 2003; Clark & Brennan, 1991; Horton & Keysar, 1996). That is, the professional takes charge of negotiating a shared understanding of the problem, its symptoms, and its putative causes. On this basis, the conversation partners can identify possible solutions and weigh them together. Finally, concrete agreements should be made on the ways in which to proceed (Weisbach & Sonne-Neubacher, 2005). One critical aspect here is that while professionals must ensure that the conversation is goal-oriented, they typically cannot solve the problem on the client's behalf. Instead, professionals should foster their clients' understanding of the problem, weighing the pros and cons of different options, and come to a joint solution, which is in line with the clients' preferences and current situation (cf. Bamberger, 2015).

Building a supportive interpersonal relationship implies the establishment of a climate of mutual respect and trust as a prerequisite for mutually working on the problem that constitutes the topic of communication (cf. Rogers, 1951). This aspect is crucial because communication on the content-level of a conversation is inevitably intertwined with the quality of processes on the interpersonal-level (e.g., Traut Mattausch & Frey, 2006; Schulz von Thun, 1998; Watzlawick, Beavin, & Jackson, 1967). Evidence suggests the equivalent importance of targeting this aspect in both medicine and teaching (Beck & Daughtridge, 2002; Lawrence-Lightfoot, 2004). The professional can foster a positive interpersonal relationship by empathizing with the conversational partner (e.g., through reflecting facial expressions and gestures), by providing unconditional positive regard and by acting authentically (i.e., through bringing his/her own thoughts and feelings into the conversation; Rogers, 1951).

Structuring the conversation proactively and solution-oriented refers to the organization of the conversation on the meta-level (Berger, 2008; Bieber, Loh, Ringel, Eich, & Härter, 2007; Weisbach & Sonne-Neubacher, 2005). This implies that the professional takes charge in organizing the conversation in a step-

wise manner (*conversation phases*). Meta-communication – such as negotiating an agenda, summarizing essential points, or moderating transitions between conversation phases – makes this organization transparent to the client (Maclure & Walker, 2000). Following a conversational script can considerably facilitate the task of structuring a conversation – an approach frequently applied to medical conversation trainings (e.g., Baile, 2000; Charles et al., 1999). Such scripts provide step-by-step models for structuring different types of professional conversations in sequential phases. For the present study, a shared decision making script was adapted to both professional domains and applied to the training and SC assessment (cf. Bieber et al., 2007). It comprised the following phases: (1) welcoming the patient/parent; (2) clarifying patients'/parents' concern and problem to be discussed; (3) offering to conduct a shared-decision-conversation; (4) naming and explaining options; (5) checking patients'/parents' understanding; (6) exploring patients'/parents' concerns, wishes, and expectations regarding the options; (7) eliciting patients'/parents' preferences; (8) negotiating options; (9) eliciting patients'/parents' decision; (10) making concrete agreements; (11) eliciting patients'/parents' satisfaction; (12) ending the conversation.

For the purpose of the present study, we conceptualize PCC as a hierarchical construct that involves the three communicative aspects elaborated above. These aspects of PCC are important from the perspective of expert-layperson communication and counseling theory, and are reasonably applicable across domains. Moreover, there is corroborating evidence for this structure from confirmatory factor analyses (Gartmeier et al., 2015; Wiesbeck, 2015). This is not to claim, however, that the present conceptualization of PCC covers all potentially relevant aspects comprehensively or applies to any type of conversation. Presumably, there are other important communicative skills in professional conversations that are more domain-specific. However, based on the discussion above, we believe that the three described aspects of competence are core skills for many professional conversations in diverse settings. Beyond shared decision-making conversations, Gartmeier et al. (2011) discussed how this model applies to other types of professional conversations; e.g., to conversations breaking bad news. Wiesbeck (2015) elaborated on how these aspects tie in with existing models of teacher-parent conversations (Aich, 2011) and counseling (Bruder, 2011; Hertel, 2009).

1.2 Assessing professional conversation competence through simulated conversations

The most straightforward way to measure PCC might be to focus on concrete, naturally occurring conversations that require minimal inferences regarding the targeted competence (Shernoff & Kratochwill, 2004). However, using real teacher-parent conversations for assessment purposes would be impractical under most circumstances, and would offer little standardization of the assessment situation. Hence, a plausible alternative is SC with trained actors. This method involves creating stan-

dardized situations that are authentic in a sense that the scenario and its task-demand-structure closely resemble real-world applications (Shavelson, 2013). SC with simulated patients were first introduced in the 1960s in order to train and assess undergraduate medical students (Barrows & Abrahamson, 1964). Today, they are an established method for measuring clinical skills and communication competence (Association of Standardized Patient Educators, 2017; Cleland et al., 2009; Ortwein, Fröhmel, & Burger, 2006; United States Medical Licensing Examination, 2016). Prior to the assessment, the actors are trained to display certain behaviors or symptoms comparable to real patients in a standard, unchanging manner (Dotger et al., 2008). The participants' performance in the SC can then be rated by following an established coding manual (e.g., Kurtz, Silverman, Benson, & Draper, 2003; Makoul, 2001; tEACH Assessment subgroup, 2012).

There is a large body of research on SC in medicine, e.g., demonstrating that they are generally well accepted and perceived as authentic (Rees, Sheard, & McPherson, 2004). Typically, physicians cannot distinguish simulated patients from real patients (Beullens, Rethans, Goedhuys, & Buntinx, 1997; Rethans, Drop, Sturmans, & van der Vleuten, 1991). Moreover, research syntheses indicate that SC, if constructed and conducted properly, are highly objective, reliable, and valid (Barman, 2005; Cleland et al., 2009; Newble, 2004). Though the implementation of SC has spread to teacher education in the form of simulated students, parents, colleagues, and school leaders (Dotger et al., 2008; Gerich & Schmitz, 2016), applications have been restricted to training of (pre-service) teachers (Dotger, 2013). So far, there is little evidence on the use of SC as an assessment method in teacher education or its psychometric quality (Dotger, Dotger, & Maher, 2010; Wiesbeck, 2015).

Because SC are a form of rater-mediated assessment, particular challenges apply from a psychometric perspective (Engelhard, 2002; Furr & Bacharach, 2013; Shavelson & Webb, 1991). Specifically, measurement error introduced by the raters should be minimized, e.g., by rater training, and analyzed, e.g., in form of interrater reliability (Chesser, Cameron, Evans, Cleland, Boursicot, & Mires, 2009; Gwet, 2014; Lurie, Mooney, Nofziger, Meldrum, & Epstein, 2008). Interrater reliability is a key concern because it is the very basis of building scores from the assessment. Moreover, construct validity is a general issue and should be evaluated whenever adapting instruments to new contexts (e.g., Furr & Bacharach, 2013).

1.3 The present study

To address the stated gap in research, the present study aimed at analyzing aspects of reliability and construct validity of the developed SC assessment (AERA, APA, & NCME, 2014; Furr & Bacharach, 2013). First, we investigated whether trained raters (*external observers*) reach a satisfactory interrater reliability when rating the SC according to a theory-based coding rubric (Gwet, 2014; Uebersax, 2016).

Research Question 1: Do trained raters reach a satisfactory interrater reliability in rating examinees' performance in the SC assessment?

Next, as a source of construct validity, we investigated expected correlation patterns regarding the different ratings of the SC, as well as their correlation patterns with other variables. In general, there is evidence for construct validity when associations of an instrument with other relevant variables meet expectations based on theory or prior research (AERA, APA, & NCME, 2014; Furr & Bacharach, 2013; Raykov & Marcoulides, 2011). For investigating this, we focused on two sets of correlations. First, we used a multimethod approach to convergent/discriminant validity (Eid & Diener, 2006; Furr & Bacharach, 2013). In multimethod measurements, different instruments are combined to measure one construct in order to provide convergent and /or discriminant validity evidence (Eid & Diener, 2006). For this purpose, we analyzed correlations among ratings from different sources: The external observers, the SC partners' (i.e. the actors), and the students' self-rating. Given that all ratings tap into the same underlying construct, we hypothesized overall positive correlations among them, albeit of differential size. Specifically, we expected stronger correlations between the two external ratings (external observers and SC partners) than among the external ratings and the students' self-rating. This expectation was based on prior research indicating that students' self-assessed communication competence often has low correlations with external assessments that target the same competences (Aich, 2011; Hertel, 2009).

Research Question 2: Do correlations between a) external observers' ratings, b) SC partners' ratings and c) students' self-ratings, correspond to the expected pattern?

As a second aspect of construct validity, we tested SC score associations with external variables, specifically students' prior communication training experiences, average grade in high school leaving examination (*Abitur*), and semester of study. We expected that students with prior communication training, having a better Abitur grade and in a more advanced semester would score better in the SC assessment. Concerning the first assumption, prior communication training constitutes a rough indicator of prior knowledge and, thus, should facilitate performance in the SC. Djakovic and Hertel (2013) illustrated that training teachers in communication or cooperation with parents improves teachers' self-assessed competence, knowledge, and professional belief. Communication training has also been identified in the improvement of doctor-patient communication (Ha & Longnecker, 2010). The second assumption, Abitur grades, can be seen as an indicator for students' ability to communicate subject related knowledge, which is an important part of expert-layperson communication. Moreover, Abitur grades are a proxy for intellectual ability and performance (Deary, Strand, Smith, & Fernandes 2007) and, thus, should facilitate performance in assessment situations. Finally, we expected that students in more advanced semesters would have had more relevant learning opportunities during their studies (e.g., in internships or courses) and, consequently, would score better in the SC. Communication competence training has been a compulsory part of German medical licensing since 2012 and 95 % of German

medical schools prepare their students – mostly via SC (Görlitz et al., 2014, p. 1). Analyses of teacher education curricula and teacher surveys reveal that competencies needed for conversations with parents have yet to be systematically integrated into German teacher education (Hertel, Bruder, Jude, & Steinert, 2013). However, teachers who entered the profession in recent years reported significantly higher amounts of preparation for parent-teacher conversations than practicing teachers (Bruder, 2011). This indicates that longer study time can be expected to coincide with increased learning opportunities and related competence development.

Regarding the size of the expected correlations, we assume that the two external ratings (observers, SC partners) will have larger correlations with the external variables compared to students' self-ratings because of the aforementioned validity problems associated with self-ratings.

Research Question 3: Do correlations among the three types of ratings and (a) prior communication training, (b) Abitur grade, and (c) semester of study, correspond to the expected pattern?

2. Method

2.1 Sample and design

The sample consisted of N = 168 students (n = 72 undergraduate medical students: 72 % female, M = 6.6 semesters;[3] n = 96 pre-service teachers: 65 % female, M = 4.6 semesters). Participation was voluntary. To mitigate potential motivational selection biases, the participants received a certificate of participation and a 25 Euro voucher as incentives. The data for this study were collected in the course of a randomized experimental study investigating the effectiveness of the *ProfKom*-training. Because this experiment is of less concern to the purpose of the present article, we shall not elaborate on it further (for details see Gartmeier et al., 2015).

2.2 Simulated conversation procedure and materials

Procedure. Each participant conducted two shared decision making conversations with two different simulated parents/patients (one male, one female; random assignment of participants to actors). Two SC per participant were implemented, as prior research indicates the advantages to validity with the use of several cases (Barman, 2005; Iramaneerat, Yudkowsky, Myford, & Downing, 2008). Prior to the conversations, the participants received a standardized introduction and two case vignettes and had then 20 minutes to prepare the conversations based on the case vignettes (see *Instruments*). After the preparation, each study participant was as-

3 The regular duration of study for physicians is around twelve semesters/6 years, for teachers it is around nine semesters/4.5 years.

signed a conference room in which the two SC took place successively. Each of the SC lasted around 10 minutes, and were video-taped. During the complete procedure, participants were separated from their peers until all had completed the SC.

Cases. In order to develop authentic case scenarios, we applied the following procedures. In teaching, cases were created on basis of a Delphi-study (Gartmeier, Bauer, Noll, & Prenzel, 2012) concerned with challenging parent-teacher conversations. In medicine, medical education experts were asked to construct cases with corresponding difficulty levels and prevalence in their field. All cases were optimized through expert consultation and a consensual validation procedure. Below is a list of the applied case topics (cf. Wiesbeck, 2015):

- Teaching, case 1: Teacher and parent discuss different options with regard to helping a student who recently received bad grades. They weigh the pros and cons of the alternatives, aiming toward the student's improved performance.
- Teaching, case 2: Teacher and parent discuss different options regarding the choice between the linguistic or scientific branch for a student's secondary education academic track, weighing the related pros and cons.
- Medicine, case 1: Physician and patient discuss different treatment options for a patient's broken arm and weigh their pros and cons.
- Medicine, case 2: Physician and patient discuss different treatment options regarding a patient's alcohol abuse problems and weigh their pros and cons.

Actors. Six professional actors played the simulated patients/teachers. Each of them took the role of one parent and one patient to avoid bias. In order to standardize their performance, the actors completed a training that included rehearsal of the respective roles and joint analysis of their videotaped portrayal, along with expert feedback.

2.3 Instruments

External observer ratings. The videotaped SC were rated by two observers using a coding rubric on the three aspects of PCC discussed above. In the construction, we followed methodological guidelines for video based coding (Seidel, Prenzel, & Kobarg, 2005; Seidel & Prenzel, 2010). Moreover, the coding rubric was partly inspired by existing instruments from medicine (e.g., EPSCALE by Edgcumbe, Silverman, & Benson, 2012). We could not completely rely on an existing instrument, however, due to the cross-domain approach and specific competence aspects that were the focus of our study.[4] The coding rubric contains a set of 43 behavior-anchored (low inference) items and 9 high inference items pertaining to the three competence aspects discussed above. We used this combination of high inference and low inference items because the two item types have different advantages and

4 Wiesbeck (2015) investigated relations of the coding rubric developed for this study with six established instruments from medicine and found an average correlation of r = .55 (range: .38 to .70).

disadvantages (e.g., Seidel & Prenzel, 2010). Low inference items refer to directly observable behavior and, thus, are relatively straightforward to rate and tend to have high interrater reliability. In contrast, high inference items demand qualitative judgements that may be more difficult for raters, making it more challenging to attain high interrater reliability (Seidel et al., 2005). However, there seem to be some advantages to high inference items in terms of validity (Newble, 2004; Regehr, MacRae, Reznick, & Szalay, 1998; Seidel et al., 2005; Seidel & Prenzel, 2010). Often, they have a more direct relationship to the theoretical construct being measured and, thus, better content validity (Seidel & Prenzel, 2010). Moreover, Regehr and colleagues (1998) hinted at advantages for predictive validity.

In the development of the coding rubric, a pilot study (N = 49, raters = 2) indicated good interrater reliability for the high inference items (ICC = .82). For the present study, we added low inference items to enrich the interpretation of the assessment data. Moreover, we wanted to give the participants feedback on their performance after the study, and low inference items are advantageous for this purpose (Altmann, 2014). Below are examples of high and low inference items for the three conversational competence aspects, respectively (cf. Wiesbeck, 2015). (1) *Advancing a problem solution*: "By the end of the conversation, the student comes to a concrete agreement with the conversational partner about how to further proceed" (high inference); "The student explains advantages and disadvantages of the options" (low inference). (2) *Building a positive relationship*: "The student shows unconditional positive regard and respect to the conversational partner" (high inference); "The student reflects the facial expression, gestures and tone of voice of the conversational partner" (low inference). (3) Structuring the conversation: "The fundamental phases of a shared decision making conversation are clearly visible" (high inference); "The student gives an advanced organizer of the different options" (low inference). All items were rated on a 5-point Likert scale with lower values indicating better performance.

The raters received extensive training on the use of the coding rubric (a one-day communication training and a two-day rater training). After the latter training, they rated 10 videos in a trial run. We compared their ratings to expert ratings (two experts from the educational domain and one expert from the medical domain) and calculated interrater reliability. Interrater reliability was $ICC \geq .60$ for all possible combinations of raters and experts. Subsequently, we trained the raters in a second workshop, based on empirical analysis of the results of the trial runs (aiming at equalizing their leniency/strictness; Langer & Schulz von Thun, 2007; Seidel et al., 2005; Wirtz & Caspar, 2002).

Actors' ratings and students' self-ratings. The actors rated the quality of each conversation immediately afterwards on three global ratings for the three competence facets (α = .82). As the actors conducted up to eight SC in a row, they rated only 3 items per participant to keep their workload within a reasonable range. The participants self-rated their performance immediately after the SC, on a questionnaire with 10 items pertaining to the three competence facets described above (α = .84).

Student Questionnaire. Students' prior communication training experience was measured with 5 items (e.g., "I have already participated in trainings, seminars or courses on conducting conversations"; yes/no answer format). Abitur grade and semester of study were asked in open response format.

2.4 Analyses

Regarding research question 1, we used intraclass correlations (*ICC* [C, 2]) as measures of interrater reliability across the 336 videos resulting from the SC (McGraw & Wong, 1996; Uebersax, 2016). According to Cicchetti (1994) interrater reliability below .40 is poor, between .40 and .59 fair, between .60 and .74 good, and above .75 excellent. We set an *ICC* ≥ .60 as threshold for acceptable reliability.

Concerning research questions 2 and 3, we conducted standard correlation analyses. For this purpose, composite PCC scores were built per participant by aggregating over the raters and the two SC. To compare the three types of ratings we used average correlations per type of rating with the three external criteria as summative validity coefficients (Raykov & Marcoulides, 2011). Correlations were Fisher z-transformed before averaging. Judging effect sizes was based on Cohen (1988) guidelines for correlations (r = .1 small, r = .3 medium, r = .5 large).

3. Results

3.1 Research question 1: Interrater reliability

Intraclass correlations indicated a satisfactory overall interrater reliability across all videos and all high and low inference items (*ICC* = .71). We also evaluated interrater reliability for the high and low inference items separately. For the high inference items, interrater reliability was *ICC* = .54 and below the pre-set threshold. For the behavior-anchored items, interrater reliability was *ICC* = .73.

3.2 Research questions 2 and 3: Relations to other variables

The left side of table 1 shows the correlations between the observer ratings, the ratings of the SC partners, and the students' self-ratings. The right side of table 1 displays the correlations between the observer ratings of the performance of the students in the SC and external variables.

Table 1: Correlations among (1) trained external observers' ratings of the simulated conversations, (2) simulated conversation partners' ratings (actors), (3) students' self-ratings, and external variables.

Ratings	(2)	(3)	PCT	Semester	Abitur grade
(1) Observers	.35**	.16*	.26**	.27**	.38**
(2) Actors		.23**	.14*	.27**	.10
(3) Students			.17*	.03	.00

Note. PCT = Prior communication training; * $p \leq .05$; ** $p \leq .01$ (one-tailed).

Regarding research question 2, the pattern of correlations among the three types of ratings occurred completely as expected. There was a significant medium-sized correlation between the observers' and the SC partners' ratings. The correlations of the two external ratings (observers, SC partners) with the students' self-rating were also significant, but smaller than the correlation between the two external ratings.

Regarding research question 3, the ratings correlated in the expected directions with the external criteria, with one exception: Unexpectedly, students' self-ratings had no correlations with Abitur grade and semester of study. Moreover, the data corroborated the assumption that the two external ratings would have higher correlations with the external criteria. The external observers' ratings had medium sized correlations with the three external criteria and the largest average correlation with them (r = .30). The actors' ratings correlated lower than the observers' ratings with prior communication training and Abitur grade (n.s.), but had a similarly sized correlation with semester of study. Consequently, as expected the average correlation with the external criteria was lower for the actors' than for the observers' ratings (r = .17), but still larger than the average correlation of the students' self-rating (r = .07). The latter was only weakly correlated to prior communication training and uncorrelated to other criteria.

4. Discussion

Communicative competences, and specifically the ability to communicate with laypersons, are central elements of professional higher education (KMK, 2005). Therefore, there is a growing general interest in the development of reliable and valid instruments for assessing them (e.g., Braun, 2016). SC assessments are particularly promising in this regard because they provide performance based data in authentic task situations rather than self-report or paper-pencil test items (cf. Aich, 2011; Blömeke et al., 2015; Hertel, 2009; Shavelson, 2013). In this study, we aimed at gathering evidence on the reliability and validity of a parallel designed SC assessment in medical and teacher education. Whereas SC are a standard tool for training and assessment in medical education, evidence on their application and psychometric quality in teacher education or as a cross-professional instrument is

still scarce. Our study adds to the existing literature in this regard by assessing interrater reliability and providing construct validity evidence for the developed SC procedure and coding rubric.

Concerning research question 1, we found good interrater reliability for the complete coding rubric as well as for the low inference items. For the high inference ratings, unlike in our pilot study, interrater reliability was below the pre-set threshold, but still "fair" according to Cicchetti's (1994) guidelines (cf. LeBreton & Senter, 2008). This finding is in line with existing evidence indicating that it is more challenging to reach a high interrater reliability with high inference items (Newble, 2004; Seidel et al., 2005). One possible explanation is that additional low inference items might have interfered with the raters' judgments of the high inference items. For example, raters might have considered high inference items as mere summaries rather than as independent ratings. Such interferences should be investigated in continuing research using cognitive interviewing techniques (Wilson, 2005).

Regarding research questions 2 and 3, the results of the correlation analyses provided preliminary evidence for the construct validity of the developed SC assessment for measuring the focused aspects of PCC in both domains. Overall, the observed patterns of correlations corresponded to those expected regarding the directions and sizes of the correlations among the different types of ratings and external criteria. Particularly, the external observers' ratings had the largest validity coefficient with the external criteria. In contrast, the finding that students' self-rating had the lowest overall correlation with the external criteria adds to prior evidence of validity problems with self-rated PCC (Aich, 2011; Hertel, 2009).

Regarding the possibility of a domain-general assessment of PCC with SC, the results are consistent with the assumption that some components of PCC generalize across domains. Across the two domains, observers were able to reach a satisfactory interrater reliability and the pattern of correlations between several types of ratings and external variables corresponded to expectations. Hence, the study extends existing research targeting only single domains by providing tentative evidence for domain-general components of PCC and by illuminating the possibility to assess PCC across domains. Nevertheless, we emphasize that we deliberately focused on aspects of PCC that are assumedly domain-general and excluded other, potentially more domain-sensitive aspects. The next steps should be to target the cross-domain transferability more explicitly by systematically comparing the assessment of PCC at the domain level and, subsequently, at the scenario level. This will allow singling out corresponding domain/scenario-specific aspects of PCC and investigating how domain-general components of PCC interact with domain-specific ones and with domain-related knowledge. For the future development of SC the proportion of domain-general, domain-specific and scenario-specific components should be carefully weighed with respect to the aim of the assessment. Domain-general assessments hold the potential to be more efficient due to their wider scope of applicability. However, for some purposes, assessments targeting domain/scenario-specific aspects of PCC might be better suited. In a similar vein, our study

was not designed to investigate how domain differences affect the implementation of SC and its psychometric properties. Studies on cross-domain measurement invariance (Millsap, 2011) of SC assessments would be required to answer that question. Such research would also be helpful to clarify whether PCC related constructs have an equivalent meaning and interpretation across domains and, thus, could contribute evidence on questions of domain specificity. Invariance analyses require large sample sizes, however, that were beyond the scope of the present study.

Regarding further limitations of our study, we acknowledge that we could only include a limited number of external criteria in the validity analyses given the context of our project. Though these criteria proved useful for the present purpose, future research might considerably expand the selection of relevant variables. Particularly, convergent and incremental validity with other instruments for measuring PCC could be investigated more closely. Finally, our study was restricted to shared decision-making conversations. Hence, it is unclear to what degree the findings generalize to other types of conversations discussed above.

In sum, we are of the opinion that SC has great potential as an instrument for assessing PCC within and across domains. Since they are also highly motivating for students and create effective learning possibilities, educators beyond the medical sector might want to establish SC as standard instruments in their curricula in order to establish closer connections between learning environments at university and challenging situations in later practice (Dotger, 2013).

Acknowledgment

The project ProfKom – Professionalization of future physicians and teachers in the area of communication competence was funded by the German Federal Ministry of Education and Research (Grant Code 01PH08015).

References

AERA, APA, & NCME – American Educational Research Association, American Psychological Assocation, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aich, G. (2011). *Professionalisierung von Lehrenden im Lehrer-Eltern-Gespräch* [Teachers‘ professional development for teacher-parent conversations]. Hohengehren, Germany: Schneider.

Altmann, R. (2014). *Entwicklung und Evaluierung von zwei verschiedenen Feedback-Varianten zur Gesprächsführungskompetenz angehender Lehrpersonen in simulierten Lehrer-Eltern-Gesprächen* [Development and evaluation of two types of feedback on simulated teacher parent conversations] (Unpublished Bachelor’s thesis). München, Germany: Technische Universität München.

Association of Standardized Patient Educators (2017). *The global network for human simulation education*. Retrieved from http://www.aspeducators.org/

Baile, W. F. (2000). SPIKES – a six-step protocol for delivering bad news: Application to the patient with cancer. *The Oncologist, 5*(4), 302–311.

Bamberger, G. G. (2015). *Lösungsorientierte Beratung: Praxishandbuch* [Solution oriented counseling: Practice guide] (5th ed.). Weinheim, Germany: Beltz.

Barman, A. (2005). Critiques on the objective structured clinical examination. *Annals Academy of Medicine Singapore*, *34*(8), 478–482.

Barrows, H. S., & Abrahamson, S. (1964). The programmed patient: A technique for appraising student performance in clinical neurology. *Journal of Medical Education*, *39*(8), 802–805.

Beck, R., & Daughtridge, R. (2002). Physician-patient communication in the primary care office: A systematic review. *Journal of the American Board of Family Medicine, 15*(1), 25–38.

Bennett, J. K., Fuertes, J. N., Keitel, M., & Phillips, R. (2011). The role of patient attachment and working alliance on patient adherence, satisfaction, and health-related quality of life in lupus treatment. *Patient Education and Counseling*, *85*(1), 53–59.

Berger, C. R. (2008). Interpersonal communication. In W. Donsbach (Ed.), *The international encyclopedia of communication* (pp. 2473–2486). Malden, MA: Blackwell Pub.

Berkhof, M., van Rijssen, J., Schellart, A., Anema, J., & van der Beek, A. (2011). Effective training strategies for teaching communication skills to physicians: An overview of systematic reviews. *Patient Education and Counseling, 84*(2), 152–162.

Beullens, J., Rethans, J.-J., Goedhuys, J., & Buntinx, F. (1997). The use of standardized patients in research in general practice. *Family Practice*, *14*(1), 58–62.

Bieber, C., Loh, A., Ringel, N., Eich, W., & Härter, M. (2007). *Patient als Partner. Patientenbeteiligung bei medizinischen Entscheidungen*. Heidelberg: Universitätsklinikum.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R.J. (2015). Approaches to competence measurement in higher education. *Zeitschrift für Psychologie, 223*(1), 1–2.

Braun, E., Athanassiou, G., & Pollerhof, K. (2016). *Entwicklung eines Messverfahrens zur Erfassung kommunikativer Fähigkeiten* [Development of an instrument for measuring communicative competences]. Paper presented at the 4th meeting of the Gesellschaft für Empirische Bildungsforschung (GEBF), Berlin. Retrieved from http://www.gebf2016.de/aktuelles/Symposien_Abstracts_240216.pdf

Bromme, R., Jucks, R., & Rambow, R. (2003). Wissenskommunikation über Fächergrenzen: Ein Trainingsprogramm [Communicating knowledge across domain borders: A training program]. *Wirtschaftspsychologie*, *5*(3), 94–102.

Bromme, R., Jucks, R., & Rambow, R. (2004). Experten-Laien-Kommunikation im Wissensmanagement. In G. Reinmann & H. Mandl (Eds.), *Der Mensch im Wissensmanagement: Psychologische Konzepte zum besseren Verständnis und Umgang mit Wissen* (pp. 176–188). Göttingen: Hogrefe.

Bromme, R., Nueckles M., & Rambow R. (1999). Adaptivity and anticipation in expert-laypeople communication. In S. E. Brennan, A. Giboin, & A. Traum (Eds.), *Psychological models of communication in collaborative systems* (pp. 17–24). Menlo Park, CA: AAAI.

Bruder, S. (2011). *Lernberatung in der Schule* [Counselling learners in school] (Doctoral dissertation, Technische Universität Darmstadt, Germany). Retrieved from http://tuprints.ulb.tu-darmstadt.de/2432/

Burks, H. M., & Stefflre, B. (Eds.). (1979). *Theories of counseling*. New York, NY: McGraw-Hill.

Charles, C., Gafni, A., & Whelan, T. (1999). Decision-making in the physician-patient encounter: Revisiting the shared treatment decision-making model. *Social Science & Medicine, 49(5)*, 651–661.

Chesser, A., Cameron, H., Evans, P., Cleland, J., Boursicot, K., & Mires, G. (2009). Sources of variation in performance on a shared OSCE station across four UK medical schools. *Medical Education*, *43*(6), 526–532.

Chrispeels, J., & Coleman, P. (1996). Improving schools through better home-school partnerships. *School Effectiveness and School Improvement*, *7*(4), 291–296.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 453–494). Washington, DC: APA.

Cleland, J. A., Abe, K., & Rethans, J.-J. (2009). The use of simulated patients in medical education: AMEE guide no 421. *Medical Teacher*, *31*(6), 477–486.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: LEA.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13–21.

Djakovic, S.-K., & Hertel, S. (2013). *Wie gut können Lehrkräfte im Rahmen von Fortbildungen erworbene Beratungskompetenzen im Schulalltag umsetzen?* 1. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Kiel.

Dotger, B. (2013). *Clinical simulations for teacher development: A companion manual for teachers*. Charlotte, NC: Information Age Publishing.

Dotger, B. H., Dotger, S. C., & Maher, M. J. (2010). From medicine to teaching: The evolution of the simulated interaction model. *Innovative Higher Education*, *35*(3), 129–141.

Dotger, B. H., Harris, S., & Hansel, A. (2008). Emerging authenticity: The crafting of simulated parent-teacher candidate conferences. *Teaching Education*, *19*(4), 337–349.

Edgcumbe, D.P., Silverman, J., & Benson, J. (2012). An examination of the validity of EPSCALE using factor analysis. *Patient Education and Counseling, 87*(1), 120–124.

Eid, M., & Diener, E. (Eds.). (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: APA.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: LEA.

Furr, R. M., & V. R. Bacharach (2013). *Psychometrics*. Los Angeles, CA, Sage.

Gartmeier, M., Bauer, J., Fischer, M. R., Hoppe-Seyler, T., Karsten, G., Kiessling, C., Möller, G., Wiesbeck, A. B., & Prenzel, M. (2015). Fostering professional communication skills of future physicians and teachers: Differential effects of e-learning and role-play. *Instructional Science, 43*(4), 443–462.

Gartmeier, M., Bauer, J., Fischer, M. R., Karsten, G., & Prenzel, M. (2011). Modellierung und Assessment professioneller Gesprächsführungskompetenz von Lehrpersonen im Lehrer-Elterngespräch. In O. Zlatkin-Troitschanskaia (Ed.), *Stationen Empirischer Bildungsforschung. Traditionslinien und Perspektiven* (pp. 412–424). Wiesbaden: Springer VS.

Gartmeier, M., Bauer, J., Noll, A., & Prenzel, M. (2012). Welchen Problemen begegnen Lehrkräfte beim Führen von Elterngesprächen? *Die Deutsche Schule, 104*(4), 374–382.

Gerich, M., & Schmitz, B. (2016). Using simulated parent-teacher talks to assess and improve prospective teachers' counseling competence. *Journal of Education and Learning, 5*(2), 285–301.

Görlitz, A., Bachmann, C., Blum, K., Höfer, S., Peters, T., Preusche, I., Raski, B., Rüttermann, S., Wagner Menghin, M., & Kiessling, C. (2014, September). *Lehren und*

Prüfen kommunikativer Kompetenzen im Medizinstudium – Ergebnisse einer Umfrage im deutschsprachigen Raum. Vortrag auf der Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA), Hamburg.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics.

Ha, J. F., & Longnecker, N. (2010): Doctor-patient communication: A review. *The Ochsner Journal 10*(1), 38–43.

Hanafin, J., & Lynch, A. (2002). Peripheral voices: Parental involvement, social class, and educational disadvantage. *British Journal of Sociology of Education, 23*(1), 35–49.

Hargie, O. (2011). *Skilled interpersonal communication*. London, GB: Routledge.

Hertel, S. (2009). *Beratungskompetenz von Lehrern: Kompetenzdiagnostik, Kompetenzförderung, Kompetenzmodellierung* [Teachers' conselling competence: Diagnostics, training, and modeling]. Münster, Germany: Waxmann.

Hertel, S., Bruder, S., Jude, N., & Steinert, B. (2013). Elternberatung an Schulen im Sekundarbereich. Schulische Rahmenbedingungen, Beratungsangebote der Lehrkräfte und Nutzung von Beratung durch die Eltern. In N. Jude, & E. Klieme (Eds.), *PISA 2009 – Impulse für die Schul- und Unterrichtsforschung* (59th supplement of the Zeitschrift für Pädagogik, pp. 40–62). Weinheim u.a.: Beltz.

Horton, W.S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*(1), 91–117.

Iramaneerat, C., Yudkowsky, R., Myford, C.M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, *13*(4), 479–493.

Jeynes, W. (2011). *Parental involvement and academic success*. New York, NY: Routledge.

KMK – Kultusministerkonferenz [Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany]. (2005). *Qualifikationsrahmen für Deutsche Hochschulabschlüsse* [Qualification framework for higher education in Germany]. Bonn, Germany: KMK.

Kreider, H., Caspe, M., Kennedy, S., & Weiss, H. (2007). *Family involvement in middle and high school students' education*. Retrieved from http://www.hfrp.org/publications-resources/publications-series/family-involvement-makes-a-difference/family-involvement-in-middle-and-high-school-students-education

Kurtz, S., Silverman, J., Benson, J., & Draper, J. (2003). Marrying content and process in clinical method teaching: Enhancing the calgary-cambridge guides. *Academic Medicine, 78*(8), 802–809.

Lane, C., & Rollnick, S. (2007). The use of simulated patients and role-play in communication skills training: A review of the literature to August 2005. *Patient Education and Counseling, 67*(1-2), 13–20.

Langer, I., & Schulz von Thun, F. (2007). *Messung komplexer Merkmale in Psychologie und Pädagogik: Ratingverfahren* [Measuring complex traits in psychology and education: Rating procedures]. Münster, Germany: Waxmann.

Lawrence-Lightfoot, S. (2004). *The essential communication: What parents and teachers can learn from each other*. New York, NY: Random House.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852.

Loh, A., Simon, D., Kriston, L., & Härter, M. (2007). Patientenbeteiligung bei medizinischen Entscheidungen [Patient participation in medical decisions]. *Deutsches Ärzteblatt, 104*(21), A1483–A1488.

Lurie, S. J., Mooney, C. J., Nofziger, A. C., Meldrum, S. C., & Epstein, R. M. (2008). Further challenges in measuring communication skills: Accounting for actor effects in standardised patient assessments. *Medical Education*, *42*(7), 662–668.

Maclure, M., & Walker, B. M. (2000). Disenchanted evenings: The social organization of talk in parent-teacher consultations in UK secondary schools. *British Journal of Sociology of Education, 21*(1), 5–25.

Makoul, G. (2001). Essential elements of communication in medical encounters: The kalamazoo consensus statement. *Academic Medicine*, *76*(4), 390–393.

Makoul, G., & Clayman, M. L. (2006). An Integrative Model of Shared Decision Making in Medical Encounters. *Patient Education and Counseling*, *60*, 301–312.

McGraw, K. O., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Psychology Press.

Newble, D. (2004). Techniques for measuring clinical competence: Objective structured clinical examinations. *Medical Education*, *38*(2), 199–203.

Ortwein, H., Fröhmel, A., & Burger, W. (2006). Einsatz von Simulationspatienten als Lehr-, Lern- und Prüfungsform. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, *56*(1), 23–29.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Rees, C., Sheard, C., & McPherson, A. (2004). Medical students' views and experiences of methods of teaching and learning communication skills. *Patient Education and Counseling*, *54*(1), 119–121.

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, *73*(9), 993–997.

Rethans, J.-J., Drop, R., Sturmans, F., & van der Vleuten, C. (1991). A method for introducing standardized (simulated) patients into general practice consultations. *British Journal of General Practice*, *41*(344), 94–96.

Rogers, C. R. (1951). *Client-centered therapy*. Boston, MA: Houghton Mifflin.

Schulz von Thun, F. (1998). *Miteinander Reden 1: Störungen und Klärungen. Allgemeine Psychologie der Kommunikation* [Talking with each other 1: Disturbances and clarifications. Psychology of communication]. Reinbek, Germany: Rowohlt.

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist, 48*(2), 73–86.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Seidel, T., & Prenzel, M. (2010). Beobachtungsverfahren: Vom Datenmaterial zur Datenanalyse [Observtion studies: From data to analysis]. In H. Holling, & B. Schmitz (Eds.), *Handbuch Statistik, Methoden und Evaluation* (pp. 139–152). Göttingen, Germany: Hogrefe.

Seidel, T., Prenzel, M., & Kobarg, M. (Eds.) (2005). *How to run a video study*. Münster, Germany: Waxmann.

Shernoff, E.S., & Kratochwill, T. R. (2004). The Application of behavioral assessment methodologies in educational settings. In S. N. Haynes & E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment* (Vol. 3. Behavioral assessment, pp. 365–385). Hoboken, NJ: Wiley.

Staples, K. E., & Diliberto, J. A. (2010). Guidelines for successful parent involvement. *Teaching Exceptional Children, 42*(6), 58–63.

Street, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Theories in Health Communication Research*, *74*(3), 295–301.

tEACH Assessment subgroup (2012). *Assessment Tools for Communication Skills*. Utrecht, The Netherlands: EACH.

Traut-Mattausch, E., & Frey, D. (2006). Kommunikationsmodelle [Communication models]. In H. W. Bierhoff & D. Frey (Eds.), *Handbuch der Sozialpsychologie und Kommunikationspsychologie* (pp. 536–544). Göttingen, Germany: Hogrefe.

Uebersax, J. (2016). *Statistical methods for rater and diagnostic reliability*. Retrieved from http://www.john-uebersax.com/stat/agree.htm

United States Medical Licensing Examination. (2016). *Step 2 CS*. Retrieved from http://www.usmle.org/step-2-cs/

Watzlawick, P., Beavin, J., & Jackson, D. D. (1967). *Pragmatics of human communication*. New York, NY: Norton.

Weisbach, C.-R., & Sonne-Neubacher, P. (2005). *Leadership in professional conversation*. Munich, Germany: DTV.

Wiesbeck, A. B. (2015). *An evaluation of simulated conversations as an assessment of pre-service teachers' communication competence in parent-teacher conversations*. Technical University of Munich, Munich. Retrieved from http://mediatum.ub.tum.de?id=1271404

Wilson, M. (2005). *Constructing measures*. Mahwah, NJ: Erlbaum.

Wirtz, M. A., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen, Germany: Hogrefe.