J Canc Educ (2019) 34:73–81
DOI 10.1007/s13187-017-1269-5

Assessing Communication Skills in Real Medical Encounters in Oncology: Development and Validation of the ComOn-Coaching Rating Scales

Marcelo Niglio de Figueiredo[1,2] · **Lorena Krippeit**[1] · **Johanna Freund**[1] · **Gabriele Ihorst**[3] · **Andreas Joos**[1] · **Juergen Bengel**[4] · **Alexander Wuensch**[1,5]

Published online: 16 August 2017

Abstract One of the challenges in research on teaching physician-patient communication is how to assess communication, necessary for evaluating training, the learning process, and for feedback. Few instruments have been validated for real physician-patient consultations. Real consultations involve unique contexts, different persons, and topics, and are difficult to compare. The aim of this study is to develop and validate a rating scale for assessment of such consultations. For the evaluation study of a communication skills training for physicians in oncology, real consultations were recorded in three assessment points. Based on earlier work and on current studies, a new instrument was developed for assessment of these consultations. Two psychologists were trained in using the instrument and assessed 42 consultations. For inter-rater reliability, interclass correlation (ICC) was calculated. The final version of the rating scales consists of 13 items evaluated on a 5-point scale. The items are grouped in seven areas: "Start of conversation," "assessment of the patient's perspective," "structure of conversation," "emotional issues," "end of conversation," "general communication skills," and "overall evaluation." ICC coefficients for the domains ranged from .44 to .77. An overall coefficient of all items resulted in an ICC of .66. The ComOn-Coaching Rating Scales are a short, reliable, and applicable instrument for the assessment of real physician-patient consultations in oncology. If adapted, they could be used in other areas. They were developed for

Electronic supplementary material The online version of this article (doi:10.1007/s13187-017-1269-5) contains supplementary material, which is available to authorized users.

✉ Marcelo Niglio de Figueiredo
marcelo.de.figueiredo@uniklinik-freiburg.de

Lorena Krippeit
lorena.krippeit@uniklinik-freiburg.de

Johanna Freund
johanna.freund@gmail.com

Gabriele Ihorst
gabriele.ihorst@uniklinik-freiburg.de

Andreas Joos
andreas.joos@uniklinik-freiburg.de

Juergen Bengel
bengel@psychologie.uni-freiburg.de

Alexander Wuensch
alexander.wuensch@uniklinik-freiburg.de

[1] Department of Psychosomatic Medicine and Psychotherapy, Medical Center—University of Freiburg, Faculty of Medicine, University of Freiburg, Hauptstr. 8, Freiburg 79104, Germany

[2] Department of Dermatology and Venereology, Medical Center—University of Freiburg, Faculty of Medicine, University of Freiburg, Hauptstr. 7, Freiburg 79104, Germany

[3] Clinical Trials Unit, Medical Center—University of Freiburg Faculty of Medicine, University of Freiburg, Freiburg, Germany

[4] Rehabilitation Psychology and Psychotherapy, Department of Psychology, University of Freiburg, Freiburg, Germany

[5] Psychosomatic Medicine and Psychotherapy, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

research and teaching purposes and meet the required methodological criteria. Rater training should be considered more deeply by further research.

Keywords Clinical competence · Communication · Communication assessment · Data accuracy · Medical oncology/education · Physician-patient relations

Introduction

Both patients and experts agree in emphasizing the importance of good communication in health care [1–3]. This consensus is based on sound research in the last decades [4] demonstrating positive effects of good communication for providers (e.g., lower burn-out levels, higher work satisfaction [5]), for patients (e.g., higher levels of trust, compliance, and quality of life [5, 6]), and, more indirectly, also for the health system (e.g., better use of financial resources [5, 7]). Two meta-analyses about efficiency of communication skills training (CST) showed that communication skills can be successfully taught [8, 9]. However, the effective sizes of the evaluated communication skills trainings are medium to small and seem to be short-lived [8, 9]. CST trainers and researchers in this field are thus confronted with a complex and multifaceted task: developing teaching methods in order to assure a good outcome of the training with long-lasting results. This task involves the development of assessment tools that allow an objective measure of the training results.

Previous research on the ComOn group of the Medical Center—University of Freiburg, Germany, was dedicated to the evaluation of two content-specific CSTs for oncological consultations. One focused on the transition from curative to palliative care (COM-ON-p [10]) and the other focused on informing patients about clinical trials in oncology (COM-ON-rct [11]). For these projects, we developed a specific assessment tool, which matched very closely the taught skills according to recommendations of Cegala and Lenzmeier Broz [12]: the COM-ON-Checklist [13]. The theoretical framework was the SPIKES Model [14], which was adapted to the specific content of the two trainings. The instrument was divided into two parts: a generic one, consisting of items assessing general communication skills (e.g., "Did the doctor show empathy to the patient?"); and a specific one, with items assessing specific skills for the transition from curative to palliative care, (e.g., "Did the doctor adequately answer the question about prognosis?"); and for the information about clinical trials, (e.g., "Did the doctor provide an appropriate explanation of randomization?"). This checklist showed moderate to good reliability in both its forms (ICC between 0.5 and 0.8 [13]).

However, these two studies were limited in their external validity as the evaluation was assessed through standardized patients, so that further investigation about their effect was needed. In a subsequent project, we have expanded the original CST to a broader spectrum of oncological consultations and settings and wanted to prove the efficacy of this new CST concept with real physician-patient encounters. Detailed information about this bicentric study evaluated on different levels in a randomized controlled trial can be found in the study protocol of *ComOn-Coaching* [15].

The expansion process raised the central question of physician-patient communication research again: "What is good communication?" A patient-centered approach to this question was followed by Bensing et al. [2], who asked laypeople in different European countries what they assume to be good communication. They analyzed interviews of focus groups after providing a video of a medical consultation to stimulate discussion. The quintessence of different focus groups in different countries and settings were 14 "tips" for physicians during a consultation. These tips show astonishing similarity with expert recommendations such as the ones from the SPIKES Model. Langewitz and colleagues complement these recommendations stressing the importance for the physician to actively structure the consultation and propose the book metaphor as a specific structuring technique [16]. The abovementioned sources were integrated in our training content and in our assessment tool, so that the content of the training could match the assessment tool [12]. The skills taught are summarized in 13 items organized in seven areas (see also [15]):

A1 Start of the conversation

- A1 Does the physician initiate the conversation appropriately?

A2 Patient's Perspective

- A2 Does the physician manage to get an idea of the patient's perspective at the beginning of or during the consultation?

B Structure of conversation

- B1 Does the physician actively give structure to the conversation (set an agenda of central topics and follow it)?
- B2 Does the physician set sub-sections in the course of the conversation (in detail)?

C Emotional issues

- C1 Does the physician recognize the patient's emotions and does he/she name them; evaluation based on NURSE by Back (2008)
- C2 Does the physician offer emotional support?

D End of conversation

D1 Does the physician summarize the content of the consultation and does he/she close the conversation appropriately?

E General communication skills

E1 Does the physician use clear and appropriate words during the conversation?
E2 Does the physician use appropriate non-verbal communication during the consultation?
E3 Does the physician adjust his pace during the consultation and does he make appropriate pauses?
E4 Does the physician offer the patient the chance to ask questions during the consultation?
E5 Does the physician check whether the patient has understood the consultation?

F Overall Evaluation

F1 How do you assess the communication skills of the physician in this conversation?

Difficulty

How difficult do you assess this consultation for the physician?

A first version of the assessment tool was published as the *ComOn-Check Rating Scales*, which was developed for a parallel project to evaluate a new program in undergraduate medical education [17]. This assessment tool showed to be an applicable instrument in short consultations with actor patients. As it will be described below, the process of transposing the *ComOn-Check Rating Scales* into the project *ComOn-Coaching*, in which longer, real-life consultations should be rated, brought new difficulties to light so that this instrument also had to be revised. The aim of this article is to describe and to discuss the development and validation of the instrument developed in this revision process, the *ComOn-Coaching Rating Scales*.

Methods

For an evaluation study of a CST for physicians working with cancer patients consisting of a workshop and individual coaching, real physician-patient consultations were video-recorded. A rating instrument was developed and validated in order to allow an external assessment of these consultations. The results of the main study are being prepared for publication.

Research Framework and Population

ComOn-Coaching [15] is a bicentric evaluation study of a CST for oncological consultations consisting of a 1–1/2 day workshop followed by an individual coaching. This RCT has two main aims: firstly, to evaluate a CST with real consultations; and secondly, to prove if a more intensive coaching can enhance the results obtained by only one coaching session. The training was based on the 13 skills mentioned above; in the form of a "memory card," they were used for feedback during the workshop and the coaching. More details on training content and study design can be found in the study protocol [15]. Seventy-three physicians (64% female; 30% hemato-oncology, 22% gynecology, 16% other internal medicine specialties, 12% radiology, 19% other specialties; age: mean = 34 years, SD = 8) from two medical centers in Germany participated in the CST and in the study. For evaluation, two real consultations per physician were video-recorded at three assessment points: before the workshop, after the workshop, and after the last coaching session. This design with three assessment points allowed us to better accompany the learning process of the physicians: the videos of the second assessment points were used in the manualized coaching sessions. Furthermore, this design permits the comparison of the effect of the two coaching models. Altogether, 431 consultations were recorded (431 different patients, 53% female; age: mean = 59 years, SD = 16; treatment status: 49% curative, 33% palliative, 18% unclear; distress: mean = 48 (SD = 28, range 0–100)). The study was open for all clinicians working with cancer patients; for the patients, there were no restrictions. The patients were chosen by the physicians among the ones they were presently treating; the consultations had such different themes as breaking bad news, providing information, discussing palliative treatment, counseling about cryo-conserve of sperm before chemotherapy, etc.

This recruitment by the physicians could conduct to a selection bias, only the simplest consultations being recorded. This problem represents an important limitation of the representativeness of our sample of consultations. As the assessment of the patients' distress shows, however, even very distressed patients were asked and agreed to participate (data is being prepared for publication), so that the selection bias seems to have been limited. Both physicians and patients had to give their written consent to participate in the study. The two raters who assessed the consultations of the main study took part in the validating work of the instrument.

Development of the Rating Scale and Raters' Training

Two psychologists, both in post-graduate education for psychotherapy and working in the psychosomatic liaison service (psychosocial care of physically ill inpatients) of our department, were trained in using the *ComOn-Check Rating Scales* [17] on videos of the abovementioned projects COM-ON-p and COM-ON-rct, in which physicians

communicated with standardized patients. This training consisted of theoretical information about physician-patient communication and on the risks of bias in the rating procedure. As soon as a good agreement on the ratings of the videos COM-ON-p and COM-ON-rct was achieved, raters were invited to rate 40 recordings randomly taken from the project *ComOn-Coaching*. The ratings showed weak ICCs for some items (< .30, see Table 3). We realized that the real-life consultations needed more detailed items, as many different kinds of consultations should be compared. So, we defined more precise guidelines for the ratings and laid them down in a manual.

The original *COM-ON-Checklist* used a 5-point scale (0–4) for each item and offered definitions for 3 anchor points: the two extremes (0 and 4) and the medium point (2). The *ComOn-Check Rating Scales* offered definitions for all 5-scale points in some items. In a first step, we revised and complemented the definitions of all scale points of all items (see Table 1). In this process, we tried to incorporate the experience of the raters, as they had to make decisions that were not included in the manual.

After one test phase, we noticed that some items are defined not only through one kind of behavior but also through complex behavior clusters. For example, the first item "start of conversation" includes greeting the patient, explaining the objectives of the consultation, asking the patient about his/her concerns, and setting a common agenda for the consultation. Each of these single behaviors can be performed better or worse, more or less appropriately. Moreover, each behavior

Table 1 Comparison of the three ComOn-Instruments–ComOn-Coaching rating by definition of all scale points

COM-ON-Checklist	ComOn-Check Rating Scales	ComOn-Coaching Rating Scales
	Context of validation study	
Standardized oncological consultations, transition from curative to palliative care; residents and specialists	Standardized consultations in general practice; undergraduate medical students	Real-life oncological consultations, thematic dependent on the actual work of the physician; residents and specialists
	Example question for "rating by definition of all scale points"	
Did the doctor explore the patient's perception of the situation?	Does the student manage to get an idea of the patient's perspective at the beginning of or during the consultation?	Does the physician manage to get an idea of the patient's perspective at the beginning of or during the consultation?
	Description in the manual	
Background information It is important that the physician captures the patient's understanding of his/her situation, so that the physician gets an idea of the patient's level of knowledge and can adapt the conversation correspondently. Positive example: "I would like to ensure that we both have the same level of knowledge. Could you tell me how you see your present situation? Could you tell me in your own words why this exam was necessary?" Rating: 0 = The physician doesn't show this behavior at all. (1 = The physician asks how the patient is doing, how he/she feels) 2 = The physician asks how the patient is feeling and if he already received information from other physicians (closed-ended question). Or: Physician himself makes a summary of past exams and treatments. 4 = The physician behaves as described above in the "background information" and the "positive example".	Positive example: "What brings you here?"; "Can you describe it more precisely?"; "What do you suppose is going on?"; "How have you dealt with these problems till now?" Rating: 0 = The physician asks about symptoms but not about past exams/treatments. 1 = The patient (or the physician, if he/she already knows the patient) reports about past exams/treatment. 2 = The physician asks the patient by means of closed-ended questions about past exams/treatments. 3 = The physician asks the patient by means of open-ended questions about past exams/treatments. 4 = The physician asks open questions about previous findings and treatments, inquires about the patient's perspective and makes an effort to get to know the patient's view.	Positive example: "What brings you here?"'; "Can you describe it more precisely?"; "What do you suppose is going on?"; "How have you dealt with these problems till now?" Rating: 0 = Patient's perspective is not included (e.g.: "We did an MRI and now we are discussing it.") 1 = The physician reports previous findings and treatments in a reassuring way (this can also be a closed question). 2 = The patient (or the physician if he/she already knows the treatment history) explains the previous findings/treatments and their relation to one another (rather than just listing them). 3 = The physician asks open questions about previous findings and treatments. Example: "Can you tell me again in your own words what was done until now?" "Can you describe it in more detail?" 4 = The physician asks open questions about previous findings and treatments, inquires about the patient's perspective and makes an effort to get to know the patient's view. E.g.: "What do you think of this?" "What was your experience of this like?" "What do you think about this?"

could be more difficult or easier to perform depending on the kind of consultation. We defined each of the behaviors as a "sub-item" and the raters would give each of them up to 1 point. The raters would subtract up to 1 point for undesirable communication behavior. The items A1, B1, B2, E1, and E2 were defined this way (see Table 2).

Pause management and pace adjustment (item E3) presented a particular challenge. Pause management and speech pace are two different categories but very difficult to separate, so they were kept together. This means that it was not possible to give only one definition of all 5-scale points for both categories. On the other hand, pauses can and should occur in many different situations: a first try to define "sub-items" revealed to be so complicated that it was not viable. After a test phase, we decided to provide the item with a longer explanation of when and how pauses are to be expected and are adequate as well as what adequate speech pace means and only to define the extreme points of the scale. Also, item F1 (overall evaluation) had only the extreme points defined, as this seemed more adequate for a global rating. In one extra item, the raters assessed the difficulty of the consultation.

The 13 items of the rating scales were summed up into seven areas: Start of Conversation, item A1; Assessing the Patient's Perspective, item A2; Structure of Conversation, items B1 and B2; Emotional Issues, items C1 and C2; End of Conversation, item D; General Communication Skills, items E1–E5; and Overall Evaluation, item F. In its first form [15], the items A1 (starting the conversation) and A2 (assessing the patient's perspective) were clustered together. During the development of the *ComOn-Check Rating Scales*, we decided to keep them separated, as the assessment of the patient's perspective can occur in a point of the consultation other than the beginning. The numbers of the items were kept in order to facilitate the comprehension of their development.

For the rating, the raters watched each video twice. The first time, they rated items of the A, B, C, and D groups and the second time, the item of the E group (general communication skills) and the overall evaluation. This system was kept,

Table 2 Comparison of the three ComOn-Instruments–ComOn-Coaching rating by sub-items

COM-ON-Checklist	ComOn-Check Rating Scales	ComOn-Coaching Rating Scales
	Context of validation study	
Standardized oncological consultations, transition from curative to palliative care; residents and specialists	Standardized consultations in general practice; undergraduate medical students	Real-life oncological consultations, thematic dependent on the actual work of the physician; residents and specialists
	Example question for "rating by sub-items"	
Did the physician open the consultation appropriately?	Does the student initiate the conversation appropriately?	Does the physician initiate the conversation appropriately?
Manual	Manual	Manual
Background information		
The physician offers an overview of the consultation. 'Appropriately' means in this case that the physician explains the objectives of the consultations, offers an overview of its topics and asks the patient and his or her relative about their concerns. Positive example 'Today I'd like to talk to you about the results of the examination. First I will tell you the results and then we can plan the next steps together. Do you have any questions or concerns?' Rating 0 = The physician offers no introduction 2 = The physician offers a short introduction 4 = The physician offers a detailed introduction and waits for a sign of assent from the patient.	In addition to introducing him/herself and greeting the patient, the student takes time for the conversation in which he/she asks personal questions. Rating 0 = The student introduces him/herself (name). 1 = The student introduces him/herself properly (name and shaking hands) but starts the conversation immediately. 2 = The student introduces him/herself properly and takes time for the beginning of the consultation (e.g. takes a seat before he starts to speak). 3 = The student introduces him/herself properly, takes time for the beginning of the consultation and asks about wellbeing. 4 = The student introduces him/herself properly and takes time for the beginning of the consultation (and asks about wellbeing), addresses the setting and/or defines with the patient a common purpose/goal for the conversation.	In addition to introducing him/herself and greeting the patient, the doctor takes time for the conversation in which he/she asks personal questions. For each criterion one point: Physician a) takes enough time; b) asks at the start of the conversation about overall wellbeing (rather than about patient's symptoms); c) addresses the setting (how much time etc. This may include addressing the situation of video recording) d) defines (together with patient) purpose/goal for the conversation Up to one point may be subtracted if the physician presents inadequate behavior (e.g. makes an improper joke). Rating 0 = The physician introduces him/herself briefly (0 points) 1 = 1 point 2 = 2 points 3 = 3 points 4 = 4 Points

even though most of the videos could be thoroughly rated the first time, in order to ensure that all videos were treated equally.

After the redefinition, the raters were trained again with randomly chosen real-life recordings of the present project ComOn-Coaching. Differences were discussed until an agreement was achieved. After the raters reached a good agreement, they rated ten recordings separately and ICCs were calculated. The procedure was repeated until at least "moderate" ICCs (ICC > 0.40; classification after Landis and Koch, 1977 [18]) was achieved in almost all items. Time and organizational constraints forced us to accept mere "fair" agreement in two items (E3 and F). Altogether, we had eight training sessions of 3–4 h.

Statistical Analysis

Ten percent of the 431 recordings from the ComOn-Coaching Study (42 recordings) were assessed by two raters, both blind to assessment point (t0/t1/t2) and study arm (intervention group (IG)/control group (CG)). The videos were randomly chosen among the ones that had not yet been seen, as some of them had been used as training material (see above). In order to assure enough variance, it is necessary that all the scale's range is used [19]. We achieved this by randomizing the videos in clusters: 14 videos from each assessment point were included, as a change through time was expected. The inter-rater reliability was calculated by means of intra-class correlation, as recommended in the literature (ICC absolute agreement, two-way random, single measure, cf. [20]). All statistical analyses were carried out using the Statistical Package for Social Sciences (SPSS, version 22).

Results

The ComOn-Coaching Rating Scales

The ComOn-Coaching Rating Scales consist of 13 items covering seven areas of medical consultations: Start, structure, and end of conversation; assessment of the patient's perspective; emotional issues; general communication skills; and overall evaluation. Each item is assessed by means of a 5-point Likert scale. The complete Manual of the Rating Scales is presented in Appendix.

Inter-rater Reliability

Table 3 shows the ICCs of the item clusters, of the single items, and of the average of all items.

Discussion

Our aim was to develop and validate an assessment instrument for physician-patient communication adequate for real consultation in a wide range of settings and topics. This instrument should be efficient (short, allow on-the-spot ratings to provide feedback) and detailed (allow a qualitative assessment of the physician's performance in different sets of skills). The *ComOn-Coaching Rating Scales* is a short instrument (13 items) and applicable in assessing real-life consultations in a number of different settings in oncology. Its flexibility makes use in other medical areas thinkable.

The work within the task of redefining the item formulation and rating rules revealed three main challenges:

1. Great heterogeneity of the raw data: While most instruments were validated with standardized consultations, our aim was to validate one instrument with real consultations—the ratings should make very different consultations comparable (e.g., consultations where the near death of a patient is communicated to relatives and the last consultation after a successful treatment by means of chemotherapy).

Table 3 Inter-rater reliability of all items and item clusters (shown in italic) with confidence intervals

	1st round		2nd round		
	N	ICC	N	ICC	95% CI
A1—Start of the conversation	*40*	*.36*	*39*	*.44*	*.16–.66*
A2—Patient's perspective	*40*	*.76*	*42*	*.49*	*.23–.69*
B—Structure of conversation		*.60*		*.64*	*.42–.79*
B1 Active structuring	38	.56	42	.67	.46–.81
B2 Setting sub-sections	38	.60	42	.46	.18–.66
C—Emotional issues		*.62*		*.54*	*.29–.72*
C1 Recognizing emotions	40	.59	42	.59	.36–.76
C2 Offering emotional support	39	.51	42	.43	.15–.65
D—End of conversation	*35*	*.44*	*39*	*.42*	*.12–.65*
E—General communication skills		*.42*		*.70*	*.51–.83*
E1 Clear and appropriate words	40	.27	42	.41	.14–.63
E2 Non-verbal communication	39	.51	42	.61	.38–.77
E3 Pacing and making pauses	40	< .01	42	.38	.09–.60
E4 Offering to ask questions	38	.90	42	.88	.79–.93
E5 Checking understanding	39	.36	42	.70	.14–.88
F—Overall evaluation	*40*	*.49*	*42*	*.35*	*.07–.58*
Difficulty	*40*	*.06*	*42*	*.51*	*.24–.71*
All items		*.70*		*.66*	*.46–.80*

Suggested interpretation: < 0.2 = poor, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = good, 0.81–1.00 = very good

We dealt with this problem by redefining the items. This redefinition brought on the one hand a better understanding of the items, but on the other hand made the rating process more complex with three kinds of rating systems (items with definitions for each Likert point, items with summative system, items with definition of two Likert points). While clearer definitions were generated, new freedom with unclearness (e.g., the deduction of points for undesirable communication behavior) made the rating work more challenging, as discussed in the following point:

2. Unwritten rating rules: Item definitions will never be complete enough to define all possibilities. They refer always to "model situations." That means that in the real rating process, raters have to complete the item definitions themselves. The rules that guide these decisions can be called "unwritten rating rules." This is one aspect of influence of subjectivity and is a challenge inherent to every rating process [21]. The more heterogeneous the material to be rated, the less the items will be able to provide detailed definitions for all possible situations and the more the raters are going to depend on the unwritten rating rules, thereby jeopardizing objectivity. Training of the raters can be thus understood as the work to make the unwritten rating rules of each rater explicit in order to allow them to converge.

We dealt with this problem in two ways: We defined the items as detailed as possible without overloading them, and we offered training as intensive as possible for the raters. Moreover, for organizational reasons, the number of raters had to be limited to two, what presumably made the "converging process" simpler. Both raters, furthermore, shared a broad background through their study and work experience, which should minimize the effect of professional education on the rating process. In spite of that, the ICCs show a wide range (.35–.88), revealing differences in the difficulty of item definition and assessment. Furthermore, while this process brought a greater agreement in some items (ICC of cluster E changed from .42 to .70), some remained unchanged (cluster B, from .60 to .64; cluster C, from .61 to .54; cluster D, from .44 to .42; and the mean of all items, from .70 to .66), and two changed downwards (cluster A2, from .76 to .50; and F, from .50 to .35). Surprisingly, items A2 and F were not affected by the redefinition work. This fact raises the question of intra-rater reliability ("Does the same rater assess the same consultation in the same way at different times?") that we were not able to consider in this study.

Peterson et al. [22] compared the use of the Gap-Kalamazoo Communication Skills Assessment Form by untrained faculty members as well as by peers on assessing communication skills in simulated conversations. In their study, the peer raters showed a higher inter-rater reliability than the faculty members, and the items of the instrument also showed a wide range of ICC mainly among the faculty members (.52–.80). The observation, that more experience (as by Peterson et al.) or, as in our study, more training does not necessarily bring a better inter-rater reliability, is surprising but not new and not yet clearly answered, as Kroboth et al. discuss [23]. Feldman et al. [24], in an article providing guidelines for rater training in high-stakes simulation-based assessments, further observe that experienced physicians seem to be more resistant to rater trainings "when the referent expert rating model differs from their own schemas of good and poor performance." Our concept of subjective, unwritten rating rules may offer a partial explanation: It drives the researcher's attention to the subjective part of the rating process and conceptualizes it as "rules" that can be brought into consciousness and, to some extent, changed. Our study suggests, however, that these rules may be relatively stable and require focused work to be changed. Future research should try to better understand the rating process as part of the instrument, and the training of the raters should receive more attention (cf. 28).

3. Variance challenge: in order to ensure enough variance, all degrees of the scale should be used [19]. In the COMON-Checklist, this was achieved by means of redefining the middle point of the items. While avoiding ceiling/bottom effects, this system has the disadvantage that scales get dependent on the sample.

In this study, as the item definitions were not as flexible as in the earlier instrument, we dealt with this problem by ensuring that the rated videos were equally distributed among the three assessment points. Our definitions were thus not so sample-dependent, but would rather allow comparison with other samples. The balance between sample dependence and comparability to other samples is always a sensible one and determines that no instrument can serve all goals.

The main challenge we faced was the question of the rater training as discussed above. The quality of the training is the guarantee of validity and reliability of the instrument [24, 25]. For instruments developed for evaluation studies, it is enough if the study raters alone show enough reliability. Instruments, however, which aim at a broader use such as, in an educational setting, to evaluate medical communicative performance or, in research, to compare trainings with each other, need a solid training which should ensure that people other than the original raters could also be trained and come to similar results. That means that the rater training itself should be evaluated. It is thus surprising that so few rater training concepts have been published or are available [25]. The development of a validated rater training would imply for example a manualization of the training and the availability of a corpus of rated videos offering an anchor for the learning process. These are future steps for our research group as well as further open questions such as concurrent validity and responsiveness.

The ComOn-Coaching Rating Scales represent one further step towards the development of an assessment instrument for physician-patient consultation in oncology. In spite of its limitations, we reached the goals we set:

- An existing checklist was adapted to real-life settings/real conversations;
- Real conversations were made evaluable and comparable;
- The physician's performance and its change can be evaluated;
- The instrument allows the provision of feedback during the training; and
- The assessment of the conversations together with the physician makes it possible to individualize the CST and to concentrate on the strengths and deficits of the particular physician.

With an intensified training of the raters who evaluate these conversations, we see a high potential for our checklist: it contributes to evaluating the physician's performance in different areas, so that training can be tailored to the physicians' and patients' reality, taking into account individual communicative needs and limitations.

Acknowledgements The ComOn-Coaching Project was made possible by the financial support of the German Cancer Aid. The article processing charge was funded by the German Research Foundation (DFG) and the University of Freiburg in the funding program Open Access Publishing. We thank Angela Vöhringer and Christopher Koppermann for the rating work and all physicians and patients for making this project possible.

Compliance with Ethical Standards The study was fully approved by the ethics committees of the Medical Center – University of Freiburg, Freiburg, Germany, and of the University Hospital Klinikum rechts der Isar, Munich, Germany, and is registered under DRKS00004385 in the DRKS (German Clinical Trials Register). The corresponding author (Marcelo Niglio de Figueiredo) has full control of all primary data and agrees to allow the journal to review the data if requested.

Conflict of Interest The authors have no conflict of interests to declare.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Hack TF, Degner LF, Parker PA (2005) The communication goals and needs of cancer patients: a review. Psychooncology 14:831–845. doi:10.1002/pon.949
2. Bensing JM, Deveugele M, Moretti F, Fletcher I, van Vliet L, Van Bogaert M, Rimondini M (2011) How to make the medical consultation more successful from a patient's perspective? Tips for doctors and patients from lay people in the United Kingdom, Italy, Belgium and the Netherlands. Patient Educ Couns 84:287–293. doi:10.1016/j.pec.2011.06.008
3. Baile WF, Aaron J (2005) Patient-physician communication in oncology: past, present, and future. Curr Opin Oncol 17:331–335
4. Pham AK, Bauer MT, Balan S (2014) Closing the patient–oncologist communication gap: a review of historic and current efforts. J Cancer Educ 29:106–113. doi:10.1007/s13187-013-0555-0
5. Epstein RM, Street RL (2007) Patient-centered communication in cancer care: promoting healing and reducing suffering. NIH Publication No. 07-6225. National Cancer Institute, Bethesda Online: https://healthcaredelivery.cancer.gov/pcc/pcc_monograph.pdf
6. Lundberg KL (2014) What are internal medicine residents missing? A communication needs assessment of outpatient clinical encounters. Patient Educ Couns 96:376–380. doi:10.1016/j.pec.2014.07.015
7. Thorne SE, Bultz BD, Baile WF (2005) Is there a cost to poor communication in cancer care? A critical review of the literature. Psychooncology 14:875–884 (discussion 885-886. doi:10.1002/pon.947
8. Barth J, Lannen P (2011) Efficacy of communication skills training courses in oncology: a systematic review and meta-analysis. Ann Oncol 22:1030–1040. doi:10.1093/annonc/mdq441
9. Moore PM, Mercado SR, Artigues MG, Lawrie TA (2013) Communication skills training for healthcare professionals working with people who have cancer. Cochrane Database Syst Rev 3: CD003751. doi:10.1002/14651858.CD003751.pub3
10. Goelz T, Wuensch A, Stubenrauch S, Ihorst G, de Figueiredo M, Bertz H, Wirsching M, Fritzsche K (2011) Specific training program improves oncologists' palliative care communication skills in a randomized controlled trial. J Clin Oncol 29:3402–3407. doi:10.1200/JCO.2010.31.6372
11. Wuensch A, Goelz T, Ihorst G, Terris DD, Bertz H, Juergen Bengel J, Wirsching M, Fritzsche K (2017) Effect of individualized communication skills training on physicians' discussion of clinical trials in oncology: results from a randomized controlled trial. BMC Cancer 17:264. doi:10.1186/s12885-017-3238-0
12. Cegala DJ, Broz SL (2002) Physician communication skills training: a review of theoretical backgrounds, objectives and skills. Med Educ 36:1004–1016
13. Stubenrauch S, Schneid E-M, Wünsch A, Helmes A, Bertz H, Fritzsche K, Wirsching M, Gölz T (2012) Development and evaluation of a checklist assessing communication skills of oncologists: the COM-ON-Checklist. J Eval Clin Pract 18:225–230. doi:10.1111/j.1365-2753.2010.01556.x
14. Baile WF, Buckmann R, Lenzi R, Glober G, Beale EA, Kudelka AP (2000) SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer. Oncologist 5:302–311
15. de Figueiredo N, Marcelo BR, Bylund CL, Goelz T, Heußner P, Sattel H, Fritzsche K, Wuensch A (2015) ComOn Coaching: study protocol of a randomized controlled trial to assess the effect of a varied number of coaching sessions on transfer into clinical practice following communication skills training. BMC Cancer 15:503. doi:10.1186/s12885-015-1454-z
16. Langewitz W, Ackermann S, Heierle A, Hertwig R, Ghanim L, Bingisser R (2015) Improving patient recall of information: harnessing the power of structure. Patient Educ Couns 98:716–721. doi:10.1016/j.pec.2015.02.003
17. Radziej, Katharina, Alexander Wuensch, Johanna Loechner, Cosima Engerer, Marcelo Niglio de Figueiredo, Johanna Freund, Heribert Sattel, Cadja Bachmann, Pascal O Berberat & Andreas Dinkel. In Review. How to assess communication skills? Development of the rating scale ComOn Check
18. Landis JR, G. G. Koch GG. (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
19. Wirtz M, Caspar F (2002) Beurteilungsübereinstimmungen und Beurteilerreliabilität. [Assessment Agreement and Rater Reliability] Hogrefe, Göttingen
20. Landers RN (2015) Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. Winnower 2: e14351881744. doi:10.15200/winn.143518.81744

21. Salmon P, Young B (2011) Creativity in clinical communication: from communication skills to skilled communication. Med Educ 45:217–226. doi:10.1111/j.1365-2923.2010.03801.x
22. Peterson EB, Calhoun AW, Rider EA (2014) The reliability of a modified Kalamazoo Consensus Statement Checklist for assessing the communication skills of multidisciplinary clinicians in the simulated environment. Patient Educ Couns 96:411–418. doi:10.1016/j.pec.2014.07.013
23. Kroboth FJ, Hanusa BH, Parker S, Coulehan JL, Kapoor W, Brown FH, Karpf M, Levey GS (1992) The inter-rater reliability and internal consistency of a clinical evaluation exercise. J Gen Intern Med 7:174–179. doi:10.1007/BF02598008
24. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D (2012) Rater training to support high-stakes simulation-based assessments. J Contin Educ Heal Prof 32:279–286. doi:10.1002/chp.21156
25. Eppich W, Nannicelli AP, Seivert NP, Sohn M-W, Rozenfeld R, Woods DM, Holl JL (2015) A rater training protocol to assess team performance. J Contin Educ Heal Prof 35:83–90. doi:10.1002/chp.21270