Patient Education and Counseling 105 (2022) 1283–1289

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pec

A video-based situational judgement test of medical students' communication competence in patient encounters: Development and first evaluation

Sabine Reiser[a,*], Laura Schacht[a], Eva Thomm[a], Christina Figalist[b], Laura Janssen[b], Kristina Schick[b], Eva Dörfler[c], Pascal O. Berberat[b], Martin Gartmeier[b], Johannes Bauer[a]

[a] University of Erfurt, Educational Research and Methodology, Erfurt, Germany
[b] Technical University of Munich, TUM School of Medicine, TUM Medical Education Center, Munich, Germany
[c] Technical University of Munich, ProLehre | Media and Didactics, Munich, Germany

ARTICLE INFO

Article history:
Received 23 November 2020
Received in revised form 22 June 2021
Accepted 20 August 2021

Keywords:
Medical education
Basic communication skills
Situational judgement test
Video-based assessment
Construct-driven development

ABSTRACT

Objective: We developed and evaluated the *Video-Based Assessment of Medical Communication Competence* (VA-MeCo), a construct-driven situational judgement test measuring medical students' communication competence in patient encounters.

Methods: In the construction phase, we conducted two expert studies ($n_{\text{panel1}} = 6$, $n_{\text{panel2}} = 13$) to ensure curricular and content validity and sufficient expert agreement on the answer key. In the evaluation phase, we conducted a cognitive pre-test ($n = 12$) and a pilot study ($n = 117$) with medical students to evaluate test usability and acceptance, item statistics and test reliability depending on the applied scoring method (raw consensus vs. pairwise comparison scoring).

Results: The results of the expert interviews indicated good curricular and content validity. Expert agreement on the answer key was high (*ICCs* > .86). The pilot study showed favourable usability and acceptance by students. Irrespective of the scoring method, reliability for the complete test (Cronbach's α > .93) and its subscales (α > .83) was high.

Conclusion: There is promising evidence that medical communication competence can be validly and reliably measured using a construct-driven and video-based situational judgement test.

Practice Implications: Video-based SJTs allow efficient online assessment of medical communication competence and are well accepted by students and educators.

1. Introduction

Patient encounters constitute a main aspect of physicians' work [1,2]. To manage this task effectively, physicians need significant medical communication competence (MCC; [3–6]). Therefore, MCC is increasingly an obligatory part of medical education in many countries [7–10]. Efficient and high-quality measurement instruments are essential to provide MCC training and support the learning process by giving feedback [11], to observe training effects [12] and to evaluate teaching methods [13]. Simulated patient encounters in objective structured clinical examinations (OSCEs; [14–16]) comprise a well-established method to evaluate MCC [8,17,18] and can attain high validity. However, they are also time- and resource-consuming [19,20]. Therefore, alternative measures, such as objective structured video examinations [21–24] or situational judgement tests (SJTs) [25–27], that can be easily administered online to large groups are important and their development is receiving growing interest. In general, the literature on video-based assessments of interpersonal skills shows that they can attain good reliability and validity and are well accepted by students [26,28,29].

This paper discusses the conceptual development and first empirical inspection of a video-based SJT to assess critical aspects of MCC in medical students, the *VA-MeCo*. SJTs present hypothetical situations to examinees in a standardized form and require them to evaluate the possible courses of action provided and to select a suitable one [30,31]. Prior research suggests that SJTs allow assessing the cognitive aspects related to MCC in a reliable, valid and efficient manner [32–38]. To ensure broad applicability in medical education (in Germany), we closely linked the VA-MeCo with established

[*] Correspondence to: Nordhäuser Str. 63, 99089 Erfurt, Germany.
E-mail address: sabine.reiser@uni-erfurt.de (S. Reiser).

https://doi.org/10.1016/j.pec.2021.08.020
0738-3991/© 2021 The Author(s). Published by Elsevier B.V.
CC_BY_4.0

standards of MCC curricula [10,39,40] and a theoretical model of professional conversational competence [41]. Based on this, we used a construct-driven development approach [42,43] that differs from traditional SJT development in its emphasis that test construction starts from theoretically defined personal traits (i.e. constructs) for which the test tasks are indicators. Methodological research on SJT development [25,26,43,44] shows that, next to conceptual advantages, construct-driven SJTs often have more favourable psychometric properties; they are also better grounded in psychometric theory [45]. To our knowledge, this construct-driven development approach distinguishes our test from other SJTs that measure competences related to medical communication [20,26,46].

1.1. Measuring MCC using SJTs

Researchers and lecturers in medical education are increasingly using SJTs to measure a variety of traits, such as social competencies [20], shared decision-making skills [25], empathy [26], and (knowledge about) non-cognitive skills like communication skills [21]. Originally, SJTs were devised for personnel selection research [31]. The core of the SJT method is to present examinees with a set of scenarios that represent typical and/or critical work-related situations in a standardized format. The presentation mode of SJTs in general is frequently text-based [30,47]; however, prior research underscores the advantages of using video stimuli to model situations and contexts [30,34]. Among other aspects, videos provide a richer contextualisation [48], resulting in better face validity [49], as well as higher motivation and acceptance on the part of participants [50,51]. Typically, the scenarios stop at a critical point, and the examinees have to evaluate different response options that indicate alternative ways to proceed within the scenario [31]. Different evaluation formats of the answer options are available, depending on the assessment purpose, such as choosing a single best option, ranking options or rating their efficiency vis-à-vis a goal [52]. To obtain individual test scores, the examinees' answers are compared with an answer key, often based on the judgements of subject matter experts (SMEs) or on theoretical reasoning [53,54]. This comparison yields a numeric score by applying an appropriate scoring method. Many such methods exist, from simple correct/incorrect scoring to elaborate methods quantifying the degree of deviation between examinees and experts. Because there is no single best scoring method [55,56], an appropriate method needs to be determined in the SJT development process. In brief, SJTs measure participants' abilities to interpret and evaluate action options in work-related situations [31]. According to research on competence assessment, such situation-specific skills are crucial because they relate individuals' dispositions (e.g. knowledge) to their real-world performance in task situations [57].

In line with a growing body of research on SJTs in general [30,33,34,37] there is evidence that, if constructed carefully, SJTs in medical education yield good psychometric properties [20,25–27,46]. Moreover, they are well accepted by participants [26,30,35] and easy to use on large groups as online tests. Video-based scenarios in online-tests seem particularly promising for measuring MCC [20,58], because they illustrate authentic acts of communication, including verbal, non-verbal and para-verbal cues. Even though the initial costs of developing video-based SJTs are high, they may eventually fall below those of simulated patient encounters, as educators and researchers can reuse SJTs without additional personnel and time [26,59].

1.2. Aims

We aimed to develop and evaluate a construct-driven, video-based SJT that measures medical students' MCC. Our research questions concerned the tests' curricular and content validity, usability and reliability. Specifically, regarding content, we investigated (1) whether subject matter experts judged the test materials (i.e. patient scenarios, video stimuli, communication goals, answer options) as correct, relevant and authentic. Moreover, we checked (2) whether an answer key with sufficient interrater agreement could be established. Regarding usability, we investigated (3) whether the test meets the prerequisites and needs of the test audience. Finally, concerning reliability, we tested (4) whether the complete test and its sub-scales have sufficient internal consistency, and to what degree this depends on the applied scoring method.

Table 1
Participant characteristics of the four conducted studies.

Study	Participants	Sample characteristics
Expert study 1	*n* = 6 SME	*Expertise criteria*: experienced active lecturers of medical communication skills *Professional background*: 3 physicians, 3 psychologists/psychotherapists/other *Gender distribution*: 3 females, 3 males
Expert study 2	*n* = 13 SME	*Expertise criteria*: same as Expert study 1 *Professional background*: 6 physicians, 7 psychologists/psychotherapists/other *Gender distribution*: 8 females, 5 males
Cognitive pre-test	*n* = 12 MS	*Semesters of study*: *M* = 9.08 (*SD* = 2.43) *Gender distribution:* 9 females, 3 males
Pilot study	*n* = 117 MS	*Semesters of study*: *M* = 8.74 (*SD* = 2.60) *Gender distribution:* 82 females, 34 males, 1 other

Note. SME = subject matter experts; MS = medical students.

2. Methods

Data for answering these research questions were collected in the multi-step construction phase (expert studies 1 and 2), the subsequent evaluation phase (cognitive pre-test and pilot study), or both. In the following, we elaborate on these points. Sample descriptions are available in Table 1.

2.1. Conceptual basis of the VA-MeCo

We followed established procedures for designing construct-driven SJTs to attain high curricular and content validity [42,43]. For this purpose, we drew upon (a) established standards of physician–patient communication (i.e. the Calgary-Cambridge Guide [40] and the CanMeds Framework [39]) and their implementation in the German National Competence-Based Catalogue of Learning Objectives [10], and (b) a model based on communication theory which specifically addresses professional communication and has been used in prior medical communication research (i.e. the Munich Model of Professional Conversation Competence [41,60]). This latter model describes professional communication competence as a hierarchical, multidimensional construct comprising three ability sub-dimensions: advancing a joint problem solution, structuring the conversation in a pro-active and transparent manner, and establishing a good working relationship. An extant study has empirically corroborated this proposed structure [41]. In the VA-MeCo's construction process, the Calgary-Cambridge Guide provided the basis for selecting and designing specific communicative tasks and quality standards for the test items. The Munich Model provided further theoretical grounding, as well as the test's target structure. Specifically, we designed the test to cover its three proposed ability dimensions:

i. Competent medical communication requires the ability to *advance the content level* of the conversation. In terms of the Calgary-Cambridge Guide, this includes gathering information (e.g. exploring the patient's problems) and explaining and planning (e.g. providing the correct amount and type of information [40]);

ii. Physicians need to be able to *structure the conversation* actively. Providing structure involves guiding the patient systematically through the conversation in an understandable and systematic way (e.g. making organisation overt, explicitly opening and closing the session [40]);

iii. Each conversational act inevitably contains a relational dimension. Physicians must be able to *establish a good working relationship* with the patient throughout the consultation, for example, through demonstrating the willingness to help and/or involving the patient in decision making [40].

2.2. Construction phase

2.2.1. Test design and task components

In the first step of the construction, we developed the general test design and a set of items. The draft test consisted of 14 tasks, each featuring the following components:

i. a background description of a patient (e.g. name, age, reason for the consultation, relevant medical history);

ii. a video clip of the physician–patient conversation (~1 min in length) stopping at a critical moment during the conversation;

iii. a communication goal the physician seeks to achieve next (e.g. gathering information about the patient's concern; summarizing and moving on to the next step; acknowledging the patient's feelings about the disease);

iv. five possible statements the physician could make to continue the conversation and achieve the communication goal (i.e. answer options).

The participants' task is to judge the effectiveness of each statement in reaching the communication goal regarding all three dimensions of MCC (i.e. advancing content, providing structure and building a relationship) on a 6-point rating scale (1 = 'very ineffective' to 6 = 'very effective'). Fig. 1 provides an example task with one answer option. Another example task is available in the supplementary material (Fig. S1). The answer options vary in their efficiency in terms of reaching the communication goal and are presented sequentially.

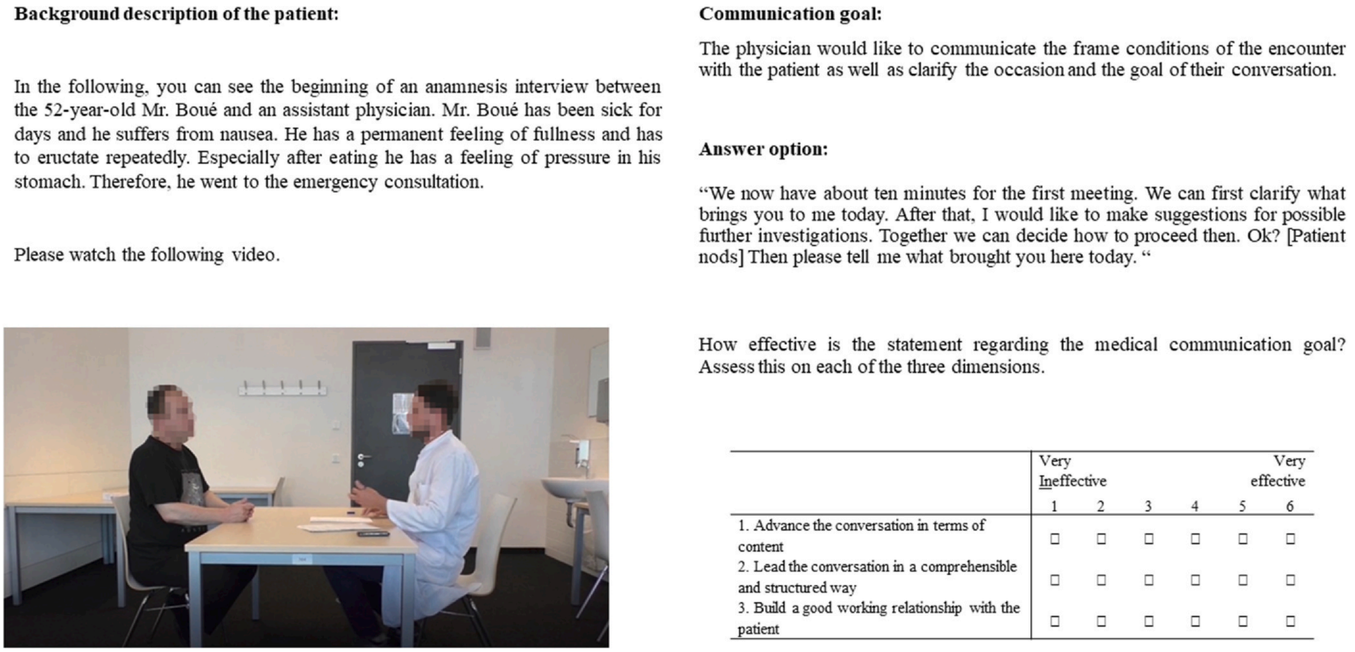
We created the video clips and answer options from material recorded during simulated patient encounters in a competence-based simulation of a resident's working day [61]. In this setting, the participating medical students were tasked with taking a medical history in an initial contact with a (simulated) patient. Drawing on the conceptual framework discussed above, we selected relevant video clips and created answer options based on statements made by the medical students in the simulation.

To assess the test's content validity, we conducted semi-structured expert interviews (*expert study 1*). To combine both medical and communication expertise, we invited experienced active lecturers of medical communication skills to participate. Moreover, we aimed at a rough gender balance to take different perspectives into account. In an iterative revision process, the experts commented on the content and format of the test (e.g. clarity of instruction and usability for the target group; variation in the quality of the answer options; medical correctness of the content). Moreover, they rated each task for its authenticity and relevance for medical communication on a 5-point rating scale, with higher values indicating higher authenticity and relevance, respectively. Answers to open-ended questions were transcribed and categorized. We revised both the content and the format of the test in light of the experts' comments.

2.2.2. Expert answer key and scoring method

To create the answer key, we recruited another panel of SMEs, applying the same criteria for expertise as above (*expert study 2*). Because the experts' task was less costly in time and effort than in the previous step, we could extend the size of the panel and thereby include experts from a broader range of faculties. The experts completed the test and for each task indicated their confidence in their judgement (5-point rating scale, 1 = not confident at all, 5 = very confident).

We analysed expert agreement in two steps. First, we checked whether a homogenous rater set could be identified using exploratory principal axis factor analysis (EFA) [62]. Having established this, we assessed agreement for the three dimensions and the



Background description of the patient:

In the following, you can see the beginning of an anamnesis interview between the 52-year-old Mr. Boué and an assistant physician. Mr. Boué has been sick for days and he suffers from nausea. He has a permanent feeling of fullness and has to eructate repeatedly. Especially after eating he has a feeling of pressure in his stomach. Therefore, he went to the emergency consultation.

Please watch the following video.

Communication goal:

The physician would like to communicate the frame conditions of the encounter with the patient as well as clarify the occasion and the goal of their conversation.

Answer option:

"We now have about ten minutes for the first meeting. We can first clarify what brings you to me today. After that, I would like to make suggestions for possible further investigations. Together we can decide how to proceed then. Ok? [Patient nods] Then please tell me what brought you here today. "

How effective is the statement regarding the medical communication goal? Assess this on each of the three dimensions.

	Very Ineffective					Very effective
	1	2	3	4	5	6
1. Advance the conversation in terms of content	□	□	□	□	□	□
2. Lead the conversation in a comprehensible and structured way	□	□	□	□	□	□
3. Build a good working relationship with the patient	□	□	□	□	□	□

Fig. 1. Example of a test task.

complete test using intra-class correlations (*ICC* 2, *k*). An *ICC* > .70 is evidence of strong agreement [63].

We selected two methods that are appropriate for our answer format and tested their impact on test reliability in the evaluation phase (see 2.3.2.). (i) In classical *raw consensus scoring* [56,64], test scores are computed per item as the squared difference between the participant's answer and the rounded mean of the experts' ratings. This method is often applied in SJTs and delivers item-level statistics that can inform test evaluation and refinement [55,56]. (ii) An alternative method receiving increasing attention is *pairwise comparison scoring* [65,66]. In this method, ratings of the answer options within tasks are compared pairwise per participant (e.g. in task 1, participant X rates answer option A higher than option B). The results are then compared to the experts' pairwise comparisons. The participant receives one point for each matching comparison. Thus, in contrast to raw consensus scoring, psychometric properties relate to pairs of items instead of individual items. In our scoring procedure, we included pairwise comparisons having 65% or higher agreement among the expert sample [65].

2.3. Evaluation phase

2.3.1. Usability and acceptance

We conducted *cognitive pre-test interviews* [67] with a panel of medical students using a mix of the think-aloud technique during test completion and after-assessment interviews with both open-ended questions and ratings. In the interviews, students commented on the usability of the test (i.e. overall design, instructions and test tasks), the effort they perceived during test completion and their interest in the theme of the test. Verbal protocols of students' comments were analysed, and test improvements were made accordingly. In the subsequent pilot study (see 2.3.2.), we re-assessed test usability and acceptance using ratings of the quality of the videos, the comprehensibility of the task elements, the handling of the test and students' interest in test processing and effort perceived during test processing (4-point rating scale, higher ratings = higher quality, etc.).

2.3.2. Reliability

In the *pilot study*, we aimed for a larger sample of n ~ 100 medical students to complete the test to analyse classical item statistics (i.e. means as an indicator of psychometric difficulty, standard deviations and discriminations) and internal consistency reliability using Cronbach's α [68]. Reliability ≥ .80 is desirable from a psychometric perspective.

In the analysis, we excluded items with problematic psychometric properties (i.e. discrimination < .10, difficulty < 10% or > 90%). Complete tasks were removed if fewer than three answer options remained after applying the exclusion criteria. To understand the impact of the scoring method, we conducted separate analyses for each method and compared the results. At best, a test's reliability does not significantly depend on a specific scoring method because this shows that results are robust to methodological decisions.

3. Results

3.1. Content validity

In *expert study 1*, participants rated the tasks as authentic (*M* = 4.36; *SD* = 0.50) and relevant for medical communication (*M* = 4.73; *SD* = 0.29). Overall, they agreed that the answer options in every task varied from high to low effectiveness in terms of the respective communication goal. Additionally, all experts considered the entire test to be well structured and very comprehensible for the target group. The usability of the test (e.g. clarity of instruction) was also assessed as positive. The task contents were mostly considered medically correct. Altogether, following the experts' suggestions, we made minor changes in the test content and format. To reduce the test length, we excluded three of the 14 tasks that the experts judged as highly similar in content to other tasks. Therefore, this shortening did not reduce content validity.

3.2. Expert solution

N = 13 SMEs participated in *expert study 2*, but missing data occurred for three experts who omitted several tasks. Therefore, we had to remove them from the analyses. As suggested by the scree plot (Fig. S2 in the supplementary material), the EFA indicated a clear one-factorial solution. It identified a homogenous rater set including nine out of the ten analysed experts who had high loadings on the primary factor and, thus, agreed in their ratings regarding the effectiveness of the answer options (Table S1 in the supplementary material). One expert loaded primarily on a second nuisance factor and thus was excluded from further analyses. The subsequent analyses with the homogenous rater set confirmed strong agreement for both the complete test (*ICC* = .88) and the individual dimensions of MCC (advancing content: *ICC* = .86; providing structure: *ICC* = .89; building a relationship: *ICC* = .90). Furthermore, the experts reported high confidence in their judgements (*M* = 3.90; *SD* = 0.56).

3.3. Usability

The verbal protocols from the *cognitive pre-test interviews* indicated that the participating students perceived the test as clearly structured (e.g. regarding the sequence of the presentation of its elements). Although the students considered the instructions extensive, they found all provided information essential. Concerning acceptance, the students reported that completing the test was moderately but not excessively demanding. Furthermore, they perceived the test as highly interesting. In summary, these findings indicated satisfactory usability and favourable acceptance of the test. Based on the students' suggestions, we made minor rearrangements and revisions.

The results from the *pilot study* quantitatively corroborated the findings of the cognitive pre-test interviews. The students judged the test's usability to be high regarding the quality of the videos (*M* = 3.60, *SD* = 0.55), the comprehensibility of task elements (*M* = 3.65, *SD* = 0.33) and the overall handling of the test (*M* = 3.52, *SD* = 0.65). Moreover, the medical students considered the test interesting (*M* = 2.97, *SD* = 0.58), and they were highly engaged while working on the test, as indicated by their reported effort (*M* = 3.54, *SD* = 0.47). Test completion took students 41 min on average (*SD* = 16).

3.4. Reliability

Table 2 lists the remaining number of items after applying the criteria for task selection and the estimated reliabilities separately for raw consensus scoring and pairwise comparison scoring. Note that the number of items is artificially higher in the latter method

Table 2
Internal consistencies for raw consensus and pairwise comparison scoring.

Dimension of MCC	Cronbach's α	
	Raw consensus scoring	Pairwise comparison scoring
Advancing content	.85 (*m* = 32)	.83 (*m* = 52)
Providing structure	.86 (*m* = 32)	.86 (*m* = 57)
Building relationship	.84 (*m* = 32)	.84 (*m* = 56)
Total	.94 (*m* = 96)	.93 (*m* = 165)

Note: m = number of items.

because it is based on the $k(k-1)/2 = 10$ pairwise comparisons per task, whereas raw consensus scoring uses the $k = 5$ answer options per task. Removing ill-fitting items marginally increased the reliability (max. difference = .07). The final internal consistencies were equally satisfactory for both scoring methods. Further descriptive statistics about the items and average student performance are available for each MCC dimension and scoring method in the supplementary material (Table S2 to S7).

4. Discussion and conclusions

4.1. Discussion

We aimed to develop and evaluate a construct-driven, video-based SJT that (i) reliably measures medical students' MCC; (ii) satisfies curricular and content validity regarding curricular standards [10,39,40] and communication theory [60]; and (iii) facilitates broad applicability in medical education with large audiences by being easy to use and well accepted. In developing and evaluating the test, we conducted a mixed-method study consisting of two studies with expert panels and two studies with medical students comprising a mix of quantitative and qualitative methods.

Concerning *curricular and content validity*, the expert interviews confirmed that the test content was medically correct and relevant for measuring medical students' MCC. The experts' feedback proved useful to refine and improve the material and to make a non-redundant selection of tasks from the initial set. Based on the gathered evidence and the conceptual rationale, we conclude that the VA-MeCo sufficiently satisfies content and curricular validity. Regarding the content, it should be noted that the scenarios focus exclusively on communication in medical history taking during a first-time patient encounter. Although this is arguably a major communicative task of physicians [1], the test does not directly cover other important types of medical communication, such as shared decision making [69] or breaking bad news [70]. The design of the VA-MeCo allows extending its content to such topics, which could be one promising direction of future development. Regarding curricular validity, it is worth mentioning that the implementation of MCC training standards varies across medical faculties [71]. Therefore, the test contents may fit local curricula differentially. Even though we included experts and students from a range of faculties in the development process, a broader investigation of the test's applicability across medical units is desirable.

Regarding the answer key, we found strong agreement among the experts, as supported by the high ICCs and successful identification of a homogenous rater set that covered nine of the ten experts who provided complete data. Please note that we had to exclude three experts who provided incomplete data. Therefore, the agreement among the experts was only calculated for the ten remaining ones. This limitation notwithstanding, the developed answer key rests on a more than two-thirds majority of the originally invited experts. These experts still covered the intended selection of lecturers of medical communication skills with diverse professional backgrounds (i.e. five physicians, five psychologists/psychotherapists/other). This high agreement on a 'correct' solution further strengthens the content validity. Conceptually, one might question whether defining a correct solution for reaching a communicative goal is appropriate because different responses may be equally effective. For this reason, we chose scoring methods that evaluate the students' judgement of each answer option (in terms of the degree of deviation from the average expert score or the consistency of pairwise comparisons) instead of assigning correct/incorrect scores if students picked the purportedly best answer. However, all these methods require a high level of agreement in terms of the experts' judgement of the effectiveness of the answer options.

Regarding *usability and acceptance*, we found no evidence of technical or motivational issues when administering the SJT to students. The experts judged the test to fit the target group well. In the cognitive pre-test interviews and the pilot study, medical students had no problems understanding and working with the test. Additionally, the students reported high levels of motivation and engagement. One practical limitation is the long testing time. We are currently working on a reduced version that includes the psychometrically best items and cuts testing time to about 30 min.

Finally, the results of the pilot study showed good *reliability* for all dimensions of MCC and both scoring methods. Although replication is warranted, the first inspection of the test suggests it meets common testing standards regarding reliability. In a replication study, we aim to check whether excluded items have truly problematic psychometric properties and to further advance this first version of the test. Moreover, we will continue to scrutinize potential effects of the scoring method on the test's psychometric properties (e.g. difficulty, factorial structure, validity coefficients with related measures) and characteristics, such as its sensitivity to instruction.

Beyond the issues already mentioned, we acknowledge some further limitations regarding the presented research. First, although the pilot study provided useful information, the evidence is only preliminary due to the small sample size. This also prevented us from conducting more elaborate psychometric investigations, such as confirming the three dimensions of MCC through factor analysis. Second, we only assessed aspects of validity relating to the test content (i.e. theoretical and curricular fit, expert agreement on correct answers) and, partly, participants' answer processes (i.e. students' understanding of the tasks and engagement with the test) [68]. In-depth investigations of convergent validity (e.g. correlates with alternative measures of MCC or related constructs, such as empathy), discriminant validity (i.e. distinguishability from unrelated constructs, such as personality or intelligence) and predictive validity for later communicative behaviour in the work environment are required. This is in line with current calls that, despite promising evidence for SJTs' good validity in general [72,73], there is still a need for further validation research on their use for assessment [31,72]. Finally, we focused on developing the SJT as an assessment instrument in the present study. To enhance its didactic use, we plan to implement a feedback option to foster and support students' learning processes [26].

4.2. Conclusions

This study showed that the VA-MeCo shows satisfying content validity, high usability and good reliability for measuring medical students' MCC. The video-based format proved beneficial for presenting medical communication scenarios in an authentic and information-rich way, thereby making the test appealing to medical educators and students [20,26]. The good acceptance by experts and students increases the chance that medical education faculties will adopt the test in their teaching. As a final note, we would like to emphasize that in its current form the VA-MeCo should be seen as a useful tool for enhancing the teaching of communication that complements (rather than replaces) established forms of assessing MCC. Despite the promising results of this study and other research on SJTs in medical education [20,25–27,46], we would currently caution against the sole use of SJTs for making high-stakes decisions, such as student selection. This would require a stronger evidence base and a better theoretical understanding of which aspects of MCC SJTs can measure validly.

4.3. Practice implications

- Video-based SJTs, such as the VA-MeCo, allow the reliable assessment of different aspects of students' MCC using authentic scenarios of medical communication.
- SJTs can be administered easily to large groups online and are well accepted by students and educators in medical education.
- To attain curricular and content validity, the conceptual basis of an SJT must align with established standards of MCC training and theoretical models of communication competence.
- Effectiveness ratings of different statements to yield a communication goal are a conceptually appropriate answer format for measuring MCC and attained high agreement among experts.
- The reliability of the developed VA-MeCo was high and did not hinge on a specific scoring method.

Ethics approval

The study was approved by the medical ethics commission of the Technical University of Munich [code 14/20S]. The students' participation in the studies was voluntary. They were informed in advance and explicitly provided informed consent.

Funding

This work was supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) [grant number 16DHB2134].

CRediT authorship contribution statement

Sabine Reiser: Conceptualization, Investigation, Validation, Formal analysis, Writing – original draft. **Laura Schacht:** Validation, Investigation, Formal analysis, Writing – original draft. **Eva Thomm:** Conceptualization, Writing – review & editing. **Christina Figalist:** Support Investigation, Methodology, Software, Writing – editing. **Laura Janssen:** Support, Investigation, Writing – review & editing. **Kristina Schick:** Support Investigation, Writing – critical review, Project administration. **Eva Dörfler:** Software. **Pascal O. Berberat:** Funding acquisition, Supervision, Support Investigation, Writing – review & editing. **Martin Gartmeier:** Funding acquisition, Supervision, Support Investigation, Writing – review & editing. **Johannes Bauer:** Conceptualization, Methodology, Supervision, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all the students for their feedback on our first test draft and their willingness to participate in our studies and the experts for their time and helpful feedback. We also thank our student assistants for their help in gathering data.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.pec.2021.08.020.

References

[1] Fallowfield L, Jenkins V, Farewell V, Saul J, Duffy A, Eves R. Efficacy of a Cancer Research UK communication skills training model for oncologists: a randomised controlled trial. Lancet 2002;359(9307):650–6. https://doi.org/10.1016/S0140-6736(02)07810-8

[2] Jünger J, Kommunikation Ärztliche. Praxisbuch zum Masterplan Medizinstudium 2020. [Medical Communication: Practice Book for the Master Plan for Medical Studies 2020]. Stuttgart: Schattauer; 2018.

[3] Ha JF, Longnecker N. Doctor-patient communication: A review. Ochsner J. 2010;10(1):38–43.

[4] Kelley JM, Kraft-Todd G, Schapira L, Kossowsky J, Riess H. The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2014;9(6):e94207https://doi.org/10.1371/journal.pone.0094207

[5] Ong LML, de Haes JCJM, Hoos AM, Lammes FB. Doctor-patient communication: a review of the literature. Soc. Sci. Med. 1995;40(7):903–18. https://doi.org/10.1016/0277-9536(94)00155-M

[6] Schick K, Reiser S, Mosene K, Schacht L, Janssen L, Thomm E, Dinkel A, Fleischmann A, Berberat PO, Bauer J, Gartmeier M. How can communicative competence instruction in medical studies be improved through digitalization? GMS J. Med. Educ. 2020;37(6):Doc57. 10.3205/zma001350.

[7] Swanson DB, van der Vleuten CPM. Assessment of clinical skills with standardized patients: State of the art revisited. Teach. Learn. Med. 2013;25(Suppl 1):17–25. https://doi.org/10.1080/10401334.2013.842916

[8] Fischer F, Helmer S, Rogge A, Arraras JI, Buchholz A, Hannawa A, et al. Outcomes and outcome measures used in evaluation of communication training in oncology - a systematic literature review, an expert workshop, and recommendations for future research. BMC Cancer 2019;19:808. https://doi.org/10.1186/s12885-019-6022-5

[9] Brown J. How clinical communication has become a core part of medical education in the UK. Med. Educ. 2008;42(3):271–8. https://doi.org/10.1111/j.1365-2923.2007.02955.x

[10] MFT / Medizinischer Fakultätentag der Bundesrepublik Deutschland e.V., Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM). [German national competence-based catalogue of learning objectives in medicine]. https://www.medstudek.uni-freiburg.de/studienganguebergreifende-bereiche/kompetenzzentrum/bmbf-verbundprojekt-merlin/nklm-final, 2015 (Accessed 16 June 2021).

[11] Burdick WP, Boulet JR, LeBlanc KE. Can We Increase the Value and Decrease the Cost of Clinical Skills Assessment? Acad. Med. 2018;93(5):690–2. https://doi.org/10.1097/ACM.0000000000001867

[12] Bauer J, Gartmeier M, Wiesbeck AB, Moeller GE, Karsten G, Fischer MR, et al. Differential learning gains in professional conversation training: A latent profile analysis of competence acquisition in teacher-parent and physician-patient communication. Learn. Individ. Differ. 2018;61:1–10. https://doi.org/10.1016/j.lindif.2017.11.002

[13] Koerfer, A., Köhle, K., Obliers, R., Sonntag, B., Thomas, W., Albus, C. Training und Prüfung kommunikativer Kompetenzen. Aus- und Fortbildungskonzepte zur ärztlichen Gesprächsführung [Training and testing of communicative competence. Training and further training concepts for medical conversation]. http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2008/ag-koerfer.pdf, 2008 (Accessed 16 June 2021).

[14] Rushforth HE. Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. Nurse Educ. Today 2007;27(5):481–90. https://doi.org/10.1016/j.nedt.2006.08.009

[15] Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med. Educ. 2004;38(2):199–203. https://doi.org/10.1111/j.1365-2923.2004.01755.x

[16] Cömert M, Zill JM, Christalle E, Dirmaier J, Härter M, Scholl I. Assessing Communication Skills of Medical Students in Objective Structured Clinical Examinations (OSCE) - A Systematic Review of Rating Scales. PLoS One 2016;11(3):e0152717https://doi.org/10.1371/journal.pone.0152717

[17] Lane C, Rollnick S. The use of simulated patients and role-play in communication skills training: A review of the literature to August 2005. Patient Educ. Couns. 2007;67(1–2):13–20. https://doi.org/10.1016/j.pec.2007.02.011

[18] Wiesbeck AB, Bauer J, Gartmeier M, Kiessling C, Möller GE, Karsten G, et al. Simulated conversations for assessing professional conversation competence in teacher-parent and physician-patient conversations. J. Educ. Res. 2017;9(3):82–101.

[19] Turner JL, Dankoski ME. Objective Structured Clinical Exams: A Critical Review. Fam. Med. 2008;40(8):574–8.

[20] Fröhlich M, Kahmann J, Kadmon M. Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. Int. J. Sel. Assess. 2017;25(1):94–110. https://doi.org/10.1111/ijsa.12163

[21] Ludwig S, Behling L, Schmidt U, Fischbeck S. Development and testing of a summative video-based e-examination in relation to an OSCE for measuring communication-related factual and procedural knowledge of medical students. GMS J. Med. Educ. 2021;38(3):70. https://doi.org/10.3205/zma001466

[22] Baribeau DA, Mukovozov I, Sabljic T, Eva KW, Delottinville CB. Using an objective structured video exam to identify differential understanding of aspects of communication skills. Med. Teach. 2012;34(4):e242–50. https://doi.org/10.3109/0142159X.2012.660213

[23] Humphris GM, Kaney S. The Objective Structured Video Exam for assessment of communication skills. Med. Educ. 2008;34(11):939–45. https://doi.org/10.1046/j.1365-2923.2000.00792.x

[24] Hulsman RL, Mollema ED, Oort FJ, Hoos AM, de Haes JCJM. Using standardized video cases for assessment of medical communication skills: Reliability of an objective structured video examination by computer. Patient Educ. Couns. 2006;60(1):24–31. https://doi.org/10.1016/j.pec.2004.11.010

[25] Kiessling C, Bauer J, Gartmeier M, Iblher P, Karsten G, Kiesewetter J, et al. Development and validation of a computer-based situational judgement test to assess medical students' communication skills in the field of shared decision making. Patient Educ. Couns. 2016;99(11):1858–64. https://doi.org/10.1016/j.pec.2016.06.006

[26] Graupe T, Fischer MR, Strijbos J-W, Kiessling C. Development and piloting of a Situational Judgement Test for emotion-handling skills using the Verona Coding Definitions of Emotional Sequences (VR-CoDES). Patient Educ. Couns. 2020;103(9):1839–45. https://doi.org/10.1016/j.pec.2020.04.001

[27] Husbands A, Rodgerson MJ, Dowell J, Patterson F. Evaluating the validity of an integrity-based situational judgement test for medical school admissions. BMC Med. Educ. 2015;15:144. https://doi.org/10.1186/s12909-015-0424-0

[28] Lievens F, Sackett PR. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. J. Appl. Psychol. 2012;97(2):460–8. https://doi.org/10.1037/a0025741

[29] Cucina JM, Su C, Busciglio HH, Harris Thomas P, Thompson Peyton S. Video-based Testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. Int J. Select. Assess. 2015;23(3):197–209. https://doi.org/10.1111/ijsa.12108

[30] Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. Med. Teach. 2016;38(1):3–17. https://doi.org/10.3109/0142159X.2015.1072619

[31] Whetzel DL, Sullivan TS, McCloy RA. Situational Judgment Tests: An Overview of Development Practices and Psychometric Characteristics. PAD 2020;6(1):1–16. https://doi.org/10.25035/pad.2020.01.001

[32] Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med. Educ. 2012;46(9):850–68. https://doi.org/10.1111/j.1365-2923.2012.04336.x

[33] Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. J Appl. Psychol. 2011;96(5):927–40. https://doi.org/10.1037/a0023496

[34] McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. Situational judgment tests, response instructions, and validity: A meta-analysis. Pers. Psychol. 2007;60(1):63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

[35] Koczwara A, Patterson F, Zibarras L, Kerrin M, Irish B, Wilkinson M. Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. Med. Educ. 2012;46(4):399–408. https://doi.org/10.1111/j.1365-2923.2011.04195.x

[36] Patterson F, Rowett E, Hale R, Grant M, Roberts C, Cousans F, et al. The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. BMC Med. Educ. 2016;16:87. https://doi.org/10.1186/s12909-016-0606-4

[37] Christian MS, Edwards BD, Bradley JC. Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. Pers. Psychol. 2010;63(1):83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

[38] Whetzel DL, McDaniel MA, Nguyen NT. Subgroup Differences in Situational Judgment Test Performance: A Meta-Analysis. Hum. Perform. 2008;21(3):291–309. https://doi.org/10.1080/08959280802137820

[39] Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med. Teach. 2007;29(7):642–7. https://doi.org/10.1080/01421590701746983

[40] Kurtz S, Silverman J, Benson J, Draper J. Marrying Content and Process in Clinical Method Teaching: Enhancing the Calgary–Cambridge Guides. Acad. Med. 2003;78(8):802–9. https://doi.org/10.1097/00001888-200308000-00011

[41] Gartmeier M, Bauer J, Fischer MR, Hoppe-Seyler T, Karsten G, Kiessling C, et al. Fostering professional communication skills of future physicians and teachers: effects of e-learning with video cases and role-play. Instr. Sci. 2015;43:443–62. https://doi.org/10.1007/s11251-014-9341-6

[42] Situational judgment tests: Theory, measurement, and application. In: Weekley JA, Ployhart RE, editors. first ed.New York: Psychology Press; 2006.

[43] Guenole N, Chernyshenko OS, Weekly J. On Designing Construct Driven Situational Judgment Tests: Some Preliminary Recommendations. Int. J. Test. 2017;17(3):234–52. https://doi.org/10.1080/15305058.2017.1297817

[44] Lievens F. Construct-Driven SJTs: Toward an Agenda for Future Research. Int. J. Test. 2017;17(3):269–76. https://doi.org/10.1080/15305058.2017.1309857

[45] Wilson M. Constructing measures: An Item Response Modeling Approach. New Jersey: Lawrence Erlbaum Associates; 2005.

[46] Schwibbe A, Lackamp J, Knorr M, Hissbach J, Kadmon M, Hampe W. Medizinstudierendenauswahl in Deutschland Messung kognitiver Fähigkeiten und psychosozialer Kompetenzen [Selection of medical students: Measurement of cognitive abilities and psychosocial competencies]. Bundesgesundheitsbl. 2018;61:178–86. https://doi.org/10.1007/s00103-017-2670-2

[47] Campion MC, Ployhart RE, MacKenzie Jr. WI. The State of Research on Situational Judgment Tests: A Content Analysis and Directions for Future Research. Hum. Perform. 2014;27(4):283–310. https://doi.org/10.1080/08959285.2014.929693

[48] Olson-Buchanan JB, Drasgow F. Multimedia situational judgment tests: The medium creates the message. In: Weekley JA, Ployhart RE, editors. Situational judgment tests: Theory, measurement, and application. first ed.New York: Psychology Press; 2006. p. 253–78.

[49] Chan D, Schmitt N. Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. J. Appl. Psychol. 1997;82(1):143–59. https://doi.org/10.1037/0021-9010.82.1.143

[50] Richman-Hirsch WL, Olson-Buchanan JB, Drasgow F. Examining the impact of administration medium on examinee perceptions and attitudes. J. Appl. Psychol. 2000;85(6):880–7. https://doi.org/10.1037/0021-9010.85.6.880

[51] Bardach L, Rushby JV, Kim LE, Klassen RM. Using video- and text-based situational judgement tests for teacher selection: a quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. Eur. J. Work. Organ. Psychol. 2020;30(2):251–64. https://doi.org/10.1080/1359432X.2020.1736619

[52] Arthur W, Glaze RM, Jarrett SM, White CD, Schurig I, Taylor JE. Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. J. Appl. Psychol. 2014;99(3):535–45. https://doi.org/10.1037/a0035788

[53] Bergman ME, Drasgow F, Donovan MA, Henning JB, Juraska SE. Scoring Situational Judgment Tests: Once You Get the Data, Your Troubles Begin. Int. J. Sel. & Assess. 2006;14(3):223–35. https://doi.org/10.1111/j.1468-2389.2006.00345.x

[54] Lievens F, Peeters H, Schollaert E. Situational judgment tests: a review of recent research. Pers. Rev. 2008;37(4):426–41. https://doi.org/10.1108/00483480810877598

[55] Cullen MJ, Sackett PR, Lievens F. Threats to the Operational Use of Situational Judgment Tests in the College Admission Process. Int. J. Sel. & Assess. 2006;14(2):142–55. https://doi.org/10.1111/j.1468-2389.2006.00340.x

[56] Weng Q, Yang H, Lievens F, McDaniel MA. Optimizing the validity of situational judgment tests: The importance of scoring methods. J. Vocat. Behav. 2018;104:199–209. https://doi.org/10.1016/j.jvb.2017.11.005

[57] Blömeke S, Gustafsson J-E, Shavelson RJ. Beyond Dichotomies: Competence Viewed as a Continuum. Z. Psychol. 2015;223(1):3–13. https://doi.org/10.1027/2151-2604/a000194

[58] Golubovich J, Seybert J, Martin-Raugh M, Naemi B, Vega RP, Roberts RD. Assessing Perceptions of Interpersonal Behavior with a Video-Based Situational Judgment Test. Int. J. Test. 2017;17(3):191–209. https://doi.org/10.1080/15305058.2016.1194275

[59] Goss BD, Ryan AT, Waring J, Judd T, Chiavaroli NG, O'Brien RC, et al. Beyond Selection: The Use of Situational Judgement Tests in the Teaching and Assessment of Professionalism. Acad. Med. 2017;92(6):780–4. https://doi.org/10.1097/ACM.0000000000001591

[60] Gartmeier M, Bauer J, Fischer MR, Karsten G, Prenzel M. Modellierung und Assessment professioneller Gesprächsführungskompetenz von Lehrpersonen im Lehrer-Elterngespräch [Modelling and assessment of teachers' professional communication skills in teacher-parent interviews]. In: Zlatkin-Troitschanskaia O, editor. Stationen empirischer Bildungsforschung: Traditionslinien und Perspektiven [Stations of empirical educational research: lines of tradition and perspectives], first. Wiesbaden: VS Verl. für Sozialwiss; 2011. p. 412–24https://doi.org/10.1007/978-3-531-94025-0_29.

[61] Prediger S, Schick K, Fincke F, Fürstenberg S, Oubaid V, Kadmon M, et al. Validation of a competence-based assessment of medical students' performance in the physician's role. BMC Med. Educ. 2020;20:6. https://doi.org/10.1186/s12909-019-1919-x

[62] Uebersax , J., Statistical methods for diagnostic agreement. http://john-uebersax.com/stat/agree.htm#recs, 2015 (Accessed 16 June 2021).

[63] LeBreton JM, Senter JL. Answers to 20 Questions About Interrater Reliability and Interrater Agreement. Organ. Res. Methods 2008;11(4):815–52. https://doi.org/10.1177/1094428106296642

[64] McDaniel MA, Psotka J, Legree PJ, Yost AP, Weekley JA. Toward an understanding of situational judgment item validity and group differences. J. Appl. Psychol. 2011;96(2):327–36. https://doi.org/10.1037/a0021983

[65] Gold B, Holodynski M. Development and Construct Validation of a Situational Judgment Test of Strategic Knowledge of Classroom Management in Elementary Schools. Educ. Assess. 2015;20(3):226–48. https://doi.org/10.1080/10627197.2015.1062087

[66] Rosman T, Mayer A-K, Krampen G. Measuring Psychology Students' Information-Seeking Skills in a Situational Judgment Test Format, Construction and Validation of the PIKE-P Test. Eur. J. Psychol. Assess. 2016;32(3):220–9. https://doi.org/10.1027/1015-5759/a000239

[67] Knafl K, Deatrick J, Gallo A, Holcombe G, Bakitas M, Dixon J, et al. The analysis and interpretation of cognitive interviews for instrument development. Res. Nurs. Health. 2007;30(2):224–34. https://doi.org/10.1002/nur.20195

[68] Furr RM. Psychometrics: An Introduction. third ed. California: SAGE; 2018.

[69] Murray E, Charles C, Gafni A. Shared decision-making in primary care: tailoring the Charles et al. model to fit the context of general practice. Patient. Educ. Couns. 2006;62(2):205–11. https://doi.org/10.1016/j.pec.2005.07.003

[70] VandeKieft GK. Breaking bad news. Am. Fam. Physician 2001;64(12):1975–9.

[71] Härtl A, Bachmann C, Blum K, Höfer S, Peters T, Preusche I, et al. Desire and reality - teaching and assessing communicative competencies in undergraduate medical education in German-speaking Europe - a survey. GMS J. Med. Educ. 2015;32(5):Doc56https://doi.org/10.3205/zma000998

[72] Lievens F. Adjusting medical school admission: assessing interpersonal skills using situational judgement tests. Med. Educ. 2013;47(2):182–9. https://doi.org/10.1111/medu.12089

[73] Cullen MJ, Zhang C, Marcus-Blank B, Braman JP, Tiryaki E, Konia M, et al. Improving Our Ability to Predict Resident Applicant Performance: Validity Evidence for a Situational Judgment Test. Teach. Learn. Med. 2020;32(5):508–21. https://doi.org/10.1080/10401334.2020.1760104