Development and piloting of a Situational Judgement Test for emotion-handling skills using the Verona Coding Definitions of Emotional Sequences (VR-CoDES)

Graupe, Tanja; Fischer, Martin R.; Strijbos, Jan-Willem; Kiessling, Claudia

Published in:
Patient Education and Counseling

DOI:
10.1016/j.pec.2020.04.001

Patient Education and Counseling 103 (2020) 1839–1845

Contents lists available at ScienceDirect

Patient Education and Counseling

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

Development and piloting of a Situational Judgement Test for emotion-handling skills using the Verona Coding Definitions of Emotional Sequences (VR-CoDES)

Tanja Graupe[a,*], Martin R. Fischer[a], Jan-Willem Strijbos[b], Claudia Kiessling[c]

[a] Institute for Medical Education, University Hospital, LMU Munich, Germany
[b] Faculty of Behavioural and Social Sciences, Department of Educational Sciences, University of Groningen, the Netherlands
[c] Lehrstuhl für die Ausbildung personaler und interpersonaler Kompetenzen im Gesundheitswesen, Fakultät für Gesundheit, Universität Witten/Herdecke, Witten, Germany

ARTICLE INFO

Article history:
Received 5 November 2019
Received in revised form 31 March 2020
Accepted 2 April 2020

Keywords:
Medical education
Assessment
Video-based assessment
Communication skills
Emotion-handling skills
Verona Coding Definitions of Emotional Sequences (VR-CoDES)
Situational Judgment Test (SJT)

ABSTRACT

Objective: Emotion-handling skills are key components for interpersonal communication by medical professionals. The Verona Coding Definitions of Emotional Sequences (VR-CoDES) appears useful to develop a Situational Judgment Test (SJT) for assessing emotion-handling skills.
Methods: In phase 1 we used a multi-stage process with expert panels ($n_{panel1} = 16$; $n_{panel2} = 8$; $n_{panel3} = 20$) to develop 12 case vignettes. Each vignette includes (1) video representing a critical incident containing concern(s) and/or cue(s), (2) standardized lead-in-question, (3) five response alternatives. In phase 2 we piloted the SJT to assess validity via an experimental study with medical students (n = 88).
Results: Experts and students rated most of the 'Reduce space' responses as inappropriate and preferred 'Explicit' responses. Women scored higher than men and there was no decline of empathy according to students' year of study. There were medium correlations with self-assessment instruments. The students' acceptance of the SJT was high.
Conclusion: The use of VR-CoDES, authentic vignettes, videos and expert panels contributed to the development and validity of the SJT.
Practice implications: Development costs were high but could be made up over time. The agreement on a proper score and the implementation of an adequate feedback structure seem to be useful.

1. Introduction

Emotion-handling skills are key components of professional communication in health care [1]. An empathic response to patients' emotional needs is central to patient-centered communication [2,3]. Mercer and Reynolds (2002) define physicians' empathy as the ability (1) to understand the patients' situation, perspective, and feelings (and their attached meanings), (2) to communicate that understanding and check its accuracy, and (3) to act on that understanding with the patient in a helpful (therapeutic) way [4]. Empathic accuracy is the degree of correctly identifying what another person is thinking or feeling [5].

Although empathy can have positive impact on medical encounters [6–9], physicians miss 70–90 % of opportunities to act in an empathic manner [10]. One reason could be that they are not able to recognize patients' emotions [11]. Patients mostly express emotions through an indirect hint of an underlying feeling [12]. Based on the Verona Coding Definitions of Emotional Sequences (VR-CoDES), a concern is a clear and unambiguous expression of an unpleasant current or recent emotion, where the emotion is explicitly verbalized. A cue is a verbal or non-verbal hint, which suggests an underlying unpleasant emotion but lacks clarity [12].

Eide et al. (2011) demonstrated the validity of VR-CoDES for recognizing patients' concerns and cues. They recommended to use this framework as a tool to foster physicians' empathic accuracy [13]. DelPiccolo et al. (2017) showed that VR-CoDES is useful to develop interventions to promote proper handling of patients' emotions in medical encounters [14], and Ortwein et al. (2017) demonstrated that VR-CoDES is beneficial for analysing medical students' written responses focusing on emotional issues [15].

* Corresponding author at: Pettenkoferstr. 8a, 80336, Munich, Germany.
E-mail address: tanja.graupe@med.uni-muenchen.de (T. Graupe).

https://doi.org/10.1016/j.pec.2020.04.001
0738-3991/

1.1. Assessment of emotion-handling skills

Hemmerdinger (2007) classified assessments of empathy into first-, second- and third person assessment [16]. First person assessment includes standardized self-rating instruments such as the Interpersonal Reactivity Index (IRI) [17] and the Jefferson Scale of Physician Empathy (JSPE) [18]. Second person assessment covers questionnaires answered by patients [16]. Third person assessment includes standardized instruments used by observer(s) to rate the learners' behavior in real or simulated clinical scenarios, e.g. Objective Structured Clinical Examination (OSCE). Running an OSCE is time and resource intensive [19]. Written and video-based tests might be an acceptable alternative for novice learners due to cost-value ratio. Van Dalen et al. (2002) pointed out that a paper-and-pencil-test of knowledge about communication skills showed good predictive validity for performing these skills in an OSCE [20]. Humphris and Kaney (2000) demonstrated that a video-based written examination is efficient, reliable and valid for testing cognitive aspects of communication skills [21].

In a Situational Judgement Test (SJT) participants are confronted with written or video-based hypothetical work-related scenarios and asked to evaluate alternative reactions within these scenarios [22]. Responses can be knowledge-based or behavioral-based [23,24] and can vary from single-best-response to multiple-response and ranking-response formats [25,26]. SJTs are based on behavioural consistency theory: anticipated behaviour is able to predict future behaviour [27]. SJTs typically compare students' responses with results from an expert panel. There is also growing evidence that during SJTs individuals develop beliefs about the effectiveness of different behaviours [28]. Finally, SJTs seem to be effective predictors of performance in practice [27,29–31].

1.2. The use of a Situational Judgement Test in medical education

SJTs in a medical context have moderate to good levels of reliability, regardless of the method used to measure reliability [22,29,32–35], as well as good levels of predictive validity in healthcare education and training [25,26,29,35,36]. SJTs have less adverse impact regarding ethnicity and gender compared to other selection tools like cognitive ability tests [35,37–40]. Participants reactions towards SJTs are positive [33,35,40,41]. Video-based SJTs evoke more favourable learners' reactions and represent a medium degree of fidelity compared to text-based SJTs, which are low in fidelity [35]. The initial development costs of video-based SJTs are higher, compared to questionnaires and OSCEs, but as they work without simulated patients and can be easily reused, costs decrease over time [42].

1.3. Aims

This multi-phase study aims to develop an user-oriented video-based SJT for assessing medical students' emotion-handling skills based on VR-CoDES, and to determine the SJTs' validity. Data analysis was performed as part of a larger study at the Ludwig-Maximilians-Universität in Munich with the overarching goal to test different measurement instruments of students' emotion-handling skills.

2. Methods

Developing and piloting the SJT consisted of two phases with different steps, where we used several expert panels, according to the specific expertise we needed. Fig. 1 provides an overview.

2.1. Phase 1: developing the Situational Judgement Test

2.1.1. Collection of scenarios

The critical incident technique was used to collect a realistic image of physicians' handling of patients' concerns and cues [43,44]. In semi-structured interviews, an expert panel$_1$ ($n_{panel1} = 16$) was asked to recall scenarios from daily medical life where they had to handle patients' and accompanying relatives' concerns and cues. The interviews were transcribed and transformed into 29 paper-based vignettes, each containing two consecutive scenarios.

2.1.2. Transformation of paper-based vignettes to video-based vignettes

To guarantee a well-balanced selection of vignettes a blueprint was developed (Appendix A). Additionally, the classification of health problems from the International Classification of Primary Care [45] was used. An expert panel$_2$ ($n_{panel2} = 8$) plus two members of the research team classified the paper-based vignettes and deemed that 21 vignettes covered the blueprint. They were transformed into screenplays and filmed with simulated patients and physicians/medical students. Videos varied between one and two minutes and represented an excerpt of a consultation including one or more triggers (concern/cue). Each scenario was introduced by a short text which was also read out loud. Subsequently, the expert panel$_2$ analyzed the videos according to the following inclusion criteria: relevance of represented situation, authenticity of actors, and existence of patients' or relatives' concern(s) and/or cue(s). Eighteen video-based vignettes satisfied all inclusion criteria.

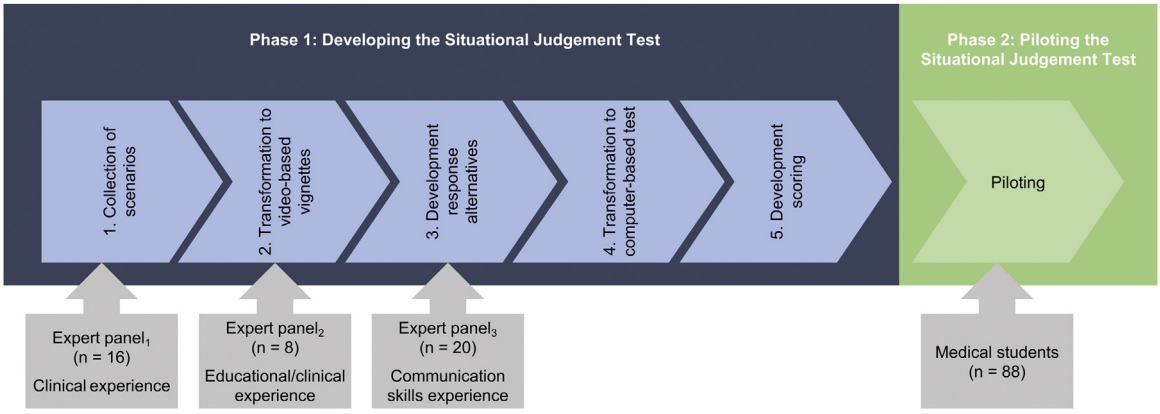

Fig. 1. Overview of the two phases of developing and piloting the SJT including the contribution of the expert panels.

2.1.3. Development and validation of response alternatives

The development of response alternatives was based on VR-CoDES [46]. Physicians` reactions to concerns and cues can generally be classified into 'Explicit' versus 'Non-explicit' and into 'Provide space' versus 'Reduce space'. The framework offers 17 strategies for physicians' possible action (e.g. Ignore (Non-explicit – Reduce space), Back Channel (Non-explicit – Provide space), Information-advice (Explicit – Reduce space), Empathy (Explicit – Provide space) [46]. Due to diversity we chose 5 response alternatives for each vignette and tried to distribute all strategies in a balanced manner, while avoiding over- or underrepresentation. Two members of the research team categorized each response alternative, resulting in acceptable interrater reliability (Cohen's kappa = 0.92). Remaining disagreements were resolved by discussion. An expert panel$_3$ (n_{panel3} = 20) was asked to complete the SJT to validate the responses. Afterwards the wording of some alternatives was changed due to ambiguousness. In the end we selected 11 video-based vignettes with two scenarios plus one vignette with only one scenario. As every scenario has 5 response alternatives, there were 115 responses in total. Of these, 28 were 'Non-explicit – Reduce space' (NR), 30 were 'Explicit – Reduce space' (ER), 16 were 'Non-explicit – Provide space' (NP), 41 were 'Explicit – Provide space' (EP) according to VR-CoDES.

2.1.4. The Situational Judgement Test as a computer-based instrument

The final 12 video-based vignettes were integrated into the online learning platform CASUS [47]. Fig. 2 illustrates an exemplary vignette. Each vignette consists of two scenarios with (1) a video representing a real-life physicians' critical incident and including one or more concern(s) and/or cue(s) expressed by a patient or relative, (2) a standardized lead-in-question, where the learner is asked to join the perspective of the physician/medical student, and (3) five response alternatives, each of which the learner rates on a slider-scale from 1 (*very inappropriate)* to 100 (*very appropriate*) with the testee not seeing the numeric values.

2.1.5. Scoring of learners' abilities

Two different scores were developed:

1 Expert-based-Score (ES): the expert panel$_3$ rated each of the response alternatives on a slider-scale from 1 to 100 and the median value was calculated for each response alternative. An answer was considered adequate if the median was 51 or more. For each scenario one "most appropriate" answer was defined among the five responses according to the highest median. Learners received a point when their answer was concordant with the expert panels' "most appropriate" answer. Given 12 vignettes with two scenarios each – except one vignette with only one scenario – the maximal ES was 23.
2 Providing-Space-based-Score (PSS): Although we know that VR-CoDES was developed for descriptive purpose we hypothesized that responses which provide space, explicitly or non-explicitly, invite patients to elaborate their concern(s) or cue(s) and are the "best" way to respond. Learners received a point if they identified (i.e. slider-scale value 51 or more) the response(s), which provided space as being appropriate. As there are 57 'Provide space' response alternatives (16 NP, 41 EP) out of 115 response alternatives in total the maximal PSS was 57.

2.2. Phase 2: piloting the Situational Judgement Test

2.2.1. Design

Medical students voluntarily participated, completing the SJT and a questionnaire. The questionnaire consisted of 13 items covering demographic data, the 28-item IRI comprising four subscales (Perspective Taking, Fantasy, Empathic Concern, Personal Distress) [17], the 20-item JSPE measuring students' perceived relevance of empathy [18,48], and 12 items on acceptance of the SJT (Appendix B).

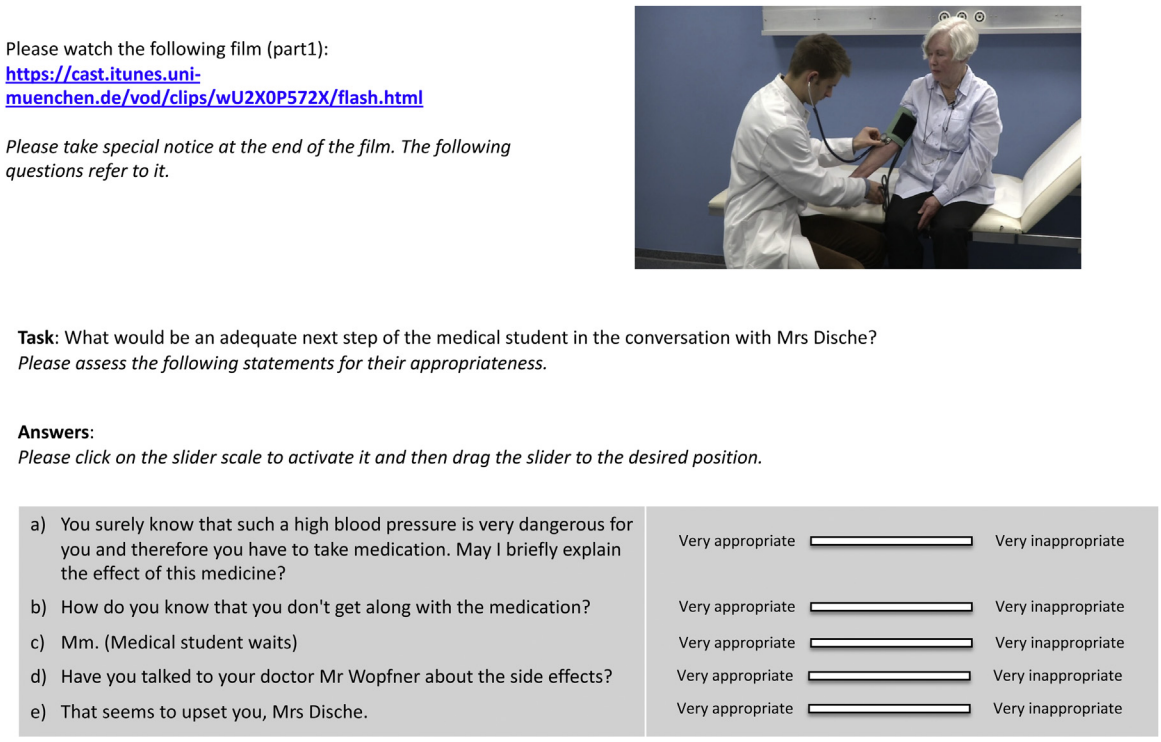

Fig. 2. Exemplary case vignette in CASUS.

2.2.2. Statistical analyses

Descriptive statistics were executed for the expert panel$_3$ and the student cohort. ES and PSS were calculated for each student. Internal consistency for both scores was determined via Cronbach's α using the student cohort. Subgroup-analysis of the student cohort was performed via t-tests. Correlations were computed using Pearson's Correlation Coefficient. Level of significance was set at 5 %. To control for multiple testing, the level of significance was set using the Bonferroni-method (p-value was set at 0.0125). All analyses were performed with SPSS 23.

3. Results

3.1. Phase 1: developing the Situational Judgement Test

3.1.1. Sample

Expert panel$_1$: 16 physicians participated in semi-structured interviews, eight (50 %) were female. The average age was 40.8 years. Eight physicians (50 %) worked in a medical practice and six (38 %) in rural regions. Their medical specialty was internal medicine (n = 5), general medicine (n = 3), surgery (n = 3) or others (n = 5).

Expert panel$_2$: Eight experts transformed the paper-based into video-based vignettes. Five experts (63 %) were female, professional background was medicine (n = 5) or educational sciences (n = 3).

Expert panel$_3$: 20 experts completed the SJT, eleven (55 %) were female. Experts' professional background was medicine (n = 13) or psychology (n = 7). All experts had experience in teaching communication skills. Two experts were additionally experienced in using VR-CoDES. These two completed the entire test. The other experts were randomly assigned into group A (n = 12) and B (n = 10) and filled in only one half of the SJT to reduce workload. Interrater reliability was determined with intra-class correlation (ICC2) for both groups (group A = 0.88; group B = 0.90). One expert from group A was a strong outlier and excluded from further analysis.

In all, 40 experts were involved. A few of them (n = 4) participated in two panels, the majority was only involved in one.

3.1.2. Descriptive statistics for the expert panel$_3$

The expert panel$_3$ rated most 'Reduce space' responses as inappropriate (NR = 97 %; ER = 80 %). However, several 'Provide space' responses were also rated as inappropriate by the experts with values $\leq$ 50 (NP = 56 %; EP = 37 %) (Table 1).

In 20 out of 23 scenarios, a 'Provide space' response (NP, EP) was judged as most appropriate. In the remaining scenarios, a 'Reduce space' response (NR, ER) was judged as most appropriate (Appendix A).

3.2. Phase 2: piloting the Situational Judgement Test

3.2.1. Sample

Of the eighty-eight participating students, 65 (74 %) were female. The average age was 24.3 years. Seventy-one participants (81 %) were born in Germany, 14 (16 %) were non-native German speakers, and 3 (4 %) did not disclose their origin. Thirty-three students (37 %) were in study years 1 or 2, and 55 (63 %) in study years 3 through 6. Forty-seven participants (53 %) had no previous experience with communication skills training. Because of data loss due to technical problems, one participant was excluded retrospectively.

3.2.2. Descriptive statistics for the student cohort

Students rated the majority of 'Reduce space' responses as inappropriate (NR = 82 %; ER = 60 %). However, students rated 40 % of 'Explicit – Reduce space' responses (ER) as appropriate. Only 31 % of 'Non-explicit – Provide space' responses (NP) were judged as appropriate (Table 2).

In 14 out of 23 scenarios, a 'Provide space' response (NP, NR) was judged as most appropriate. In the remaining scenarios a 'Reduce space' response (NR, ER) was judged as most appropriate.

With regard to ES the students' mean was 10.9 out of 23 points (SD = 0.4; min = 0, max = 19). Relating to item difficulty, there were five scenarios where less than 30 % of the students received a point. Internal consistency of the ES as measured by Cronbach's α was 0.75. With regard to PSS the students' mean was 28.8 out of 57 points (SD = 1.2; min = 0, max = 57). Internal consistency of the PSS as measured by Cronbach's α was 0.92.

3.2.3. Comparison of the expert panel$_3$ and the student cohort

Whereas experts rated 12 % of 'Reduce space' responses as adequate, students perceived 29 % as adequate. For experts, responses expressing empathy or affect acknowledgment (n = 19) were perceived as most adequate (average median$_{empathy}$ = 72; average median$_{acknowledgment}$ = 67). For students, responses expressing content exploration and post-poning were perceived as most adequate (each average median = 69). Both groups rated 'Explicit – Provide space' responses (EP) higher than 'Non-explicit – Provide space' responses (NP).

Experts' and students' ratings of the most appropriate response were congruent in 12 out of 23 scenarios. In seven scenarios, students' highest rating of the most appropriate response reflected experts' second highest rating. In four scenarios with no concordance, the experts voted for a 'Provide space' response (P) whereas the students voted in three scenarios for a 'Reduce space' response (R). Furthermore, in one scenario the students voted for 'Explicit – Provide space – Content – Acknowledge' (EPCAc), whereas the experts voted for 'Explicit – Provide space – Affect – Acknowledge' (EPAAc).

3.2.4. Evidence for the validity of the Situational Judgement Test

According to Downing (2003) we examined the degree of validity through hypothesis-driven subgroup-analyses [49].

3.2.4.1. Correlations between SJT and JSPE as well as IRI. We hypothesized positive correlations between the SJT, JSPE and IRI, as all supposedly measure (aspects of) empathy. Results showed that students' score on the JSPE correlated significantly positive with the ES (r = 0.326, p = 0.002), but their scores on the four IRI-subscales did not correlate with the ES. Students' scores on the JSPE and the four IRI subscales did not correlate with the PSS.

Table 1
Experts' rating of the SJTs' response alternatives according to VR-CoDES.

Category according to the Verona Coding Definitions of Emotional Sequences	Total number of responses (% of 115)	> 51 (% of total number)	≤ 50 (% of total number)
Non-explicit – Reduce space (NR)	28 (24 %)	1 (3 %)	27 (97 %)
Explicit – Reduce space (ER)	30 (26 %)	6 (20 %)	24 (80 %)
Non-explicit – Provide space (NP)	16 (14 %)	7 (44 %)	9 (56 %)
Explicit – Provide space (EP)	41 (36 %)	26 (63 %)	15 (37 %)

Table 2
Students' rating of the SJTs` response alternatives according to VR-CoDES.

Category according to the Verona Coding Definitions of Emotional Sequences	Total number of responses (% of 115)	> 51 (% of total number	≤ 50 (% of total number
Non-explicit – Reduce space (NR)	28 (24 %)	5 (18 %)	23 (82 %)
Explicit – Reduce space (ER)	30 (26 %)	12 (40 %)	18 (60 %)
Non-explicit – Provide space (NP)	16 (14 %)	5 (31 %)	11 (69 %)
Explicit – Provide space (EP)	41 (36 %)	26 (63 %)	15 (37 %)

3.2.4.2. Subgroup-analysis according to gender. We hypothesized that women (n = 65) would score higher than men (n = 23) because women generally show higher empathy values [50]. Women indeed scored descriptively higher in the ES (ES $mean_{men}$ = 9.0, SD = 4.0; ES $mean_{women}$ = 11.7, SD = 4.0; t(82) = 2.5, p = 0.014) and in the PSS than men, but not significantly (PSS $mean_{men}$ = 26.0, SD = 11.0; PSS $mean_{women}$ = 30.0, SD = 11.0; t(82) = 1.4, p = 0.115).

3.2.4.3. Subgroup-analysis according to study year. We hypothesized that advanced students (years 3 through 6; n = 55) would score lower than novice students (years 1 and 2; n = 33) because we expected a decline of empathy [51]. Results showed that advanced students scored significantly higher in ES and PSS than novice students (ES $mean_{1\ and\ 2}$ = 8.9, SD = 4.1; ES $mean_{3\ to\ 6}$ = 12.1, SD = 3.7; t(85) = 3.8, p ≤ 0.000; PSS $mean_{1\ and\ 2}$ = 24.8, SD = 10.5; PSS $mean_{3\ to\ 6}$ = 31.2, SD = 10.8; t(85) = 2.7, p = 0.009).

3.2.4.4. Subgroup-analysis according to grade of experience. We hypothesized that students with experience in communication skills training (n = 41) would score higher than students with no experience (n = 47) although it might be contradictive to the hypothesis in section 3.2.4.3. (students undergo a specific communication skills training with standardized patients at LMU Munich in years 2 and 3). Prior experience with communication skills training was measured with five numerical questions (participation in training, reading literature about communication, practical experience, formal qualification, other). Answers were rated as 0 or 1 and summed up (0 = no experience; 5 = rich experience). Increased experience with communication skills training correlated positively with both scores (ES r = 0.350, p = 0.001; PSS r = 0.271, p = 0.011).

3.2.4.5. Subgroup-analysis according to origin. We hypothesized non-native German speakers (n = 14) would score lower than native speakers (n = 71) due to language problems. Native speakers scored significantly higher in both scores than non-native speakers (ES $mean_{native}$ = 11.4, SD = 3.8; ES $mean_{non\text{-}native}$ = 8.4, SD = 4.9; t(85) = 2.6, p = 0.010; PSS $mean_{native}$ = 30.3, SD = 10.2; PSS $mean_{non\text{-}native}$ = 21.2, SD = 12.9; t(85) = 2.9, p = 0.004).

3.2.5. Acceptance of Situational Judgement Test

Of the 87 participants, 64 (73,5 %) rated the technical use of the SJT and the online learning environment as “good” or “very good”. Furthermore, 55 participants (63,2 %) rated the slider-scale as “very useful” on a 7-point Likert-scale. In all, 70 participants (80,5 %) expressed a very strong satisfaction with the format of the SJT (Likert-scale values ranging from 5 to 7) and 64 (73,5 %) deemed the SJT's content as very relevant for their clinical work. Finally, 86 (98,9 %) would regularly take part in formative or summative SJTs during their university career.

4. Discussion and conclusion

4.1. Discussion

We aimed to develop and pilot a video-based SJT measuring emotion-handling skills that is easy to apply and evaluate for clinical teachers. VR-CoDES was originally developed to describe and analyse provider-patient-encounters for research purposes [14], whereas we used this framework for a normative purpose. We hypothesized that physicians' ‘Provide space’ reactions to patients' concerns and cues are more appropriate than ‘Reduce space’ responses. Our results indicate that experts rated ‘Provide space’ responses more often as appropriate than students. Both groups preferred ‘Explicit’ responses in comparison to ‘Non-explicit’ responses. However, experts rated a ‘Reduce space’ response as most appropriate in three scenarios. In one scenario, the physician gave confusing information to the patient, which led to insecurity, and the expert panel decided that ‘Explicit – Reduce space – Information advise’ (ERIa) would be the most adequate response. The other two scenarios started with concerned relatives asking for information and the experts opted for the ‘Explicit – Reduce space – Post-poning’ (ERPp) response, talking with the relatives and the patient at a later point in time. These decisions by the experts seem plausible.

Consequential the approach behind PSS is not completely sufficient. ‘Provide space’ is not always the appropriate strategy and there are situations where ‘Reduce space’ responses appear more adequate for physicians. We recommend using the ES.

There were also several responses that were rated as inappropriate by the experts although they seem correct. A possible explanation could be their wording. It is very difficult to formulate responses that fit to everybody's use of language and personal style. Even single words or their order seem to have an impact. In one scenario, the highest judgement for the most appropriate answer was only 63, which is comparatively low. A rewording of the responses in this scenario is necessary for future use.

In the expert panel, responses expressing explicit empathy (EPAEm) and affect acknowledgment (EPAAc) and in the student cohort the codes ‘Explicit – Reduce space – Post-poning’ (ERPp) and ‘Explicit – Provide space – Content – Explore’ (EPCEx) were perceived as most appropriate. Affect-related codes played a minor role in the students' opinion. These findings indicate that it seemingly is not clear to students that dealing with emotions has a positive impact on patients' health. Therefore, the relevance of emotion-handling skills needs to be explicitly highlighted in communication skills curricula. Whether or not experts and students of varying educational levels (e.g. undergraduate vs. graduate) differ in their priorities based on their knowledge and/or experience in communication skills is an intriguing question for future research. In the expert panel the range of appropriateness was noticeably high regarding ‘Reduce space’ responses (Appendix A), which hints at some disagreement among the experts in using this kind of strategies.

The SJT showed different correlations with self-assessment instruments like JSPE [18] and IRI [17]. Only students' ES correlated significantly positivly with the JSPE. It seems that the construct underlying the JSPE appears closer to our SJT. This leads to the question whether emotion-handling skills are the same construct as empathy. The idea of VR-CoDES is to detect patients' concerns and cues and provide space to elaborate possible underlying emotions. The concept of empathy, according to Mercer and Reynolds [4], is very close to this construct. The difference between

VR-CoDES in our SJT and the JSPE is that with our SJT we measure cognitive ability whereas the JSPE measures attitude. Cognition, behavior and attitude are different facets of emotion-handling, and the interplay of these facets needs further investigation.

As hypothesized, females scored higher than males and students with prior experience in communication skills training scored higher than students with no experience. Against our hypothesis, we did not identify a decline of empathy according to year of study. Advanced students scored higher than novice students. This finding might be due to an improved communication skills training and more clinical experiences. German native speakers scored higher than non-native speakers. Perhaps there was some discrimination of non-native speakers within our SJT. In all, future research with a larger sample could provide more definitive information on subgroup comparisons.

Our study has some limitations. Students participated voluntarily and the cohort might be a selection of highly motivated students. Women were overrepresented in the sample. Some of the codes according to VR-CoDES include non-verbal behavior and were difficult to express in the style of written response alternatives, e.g. 'Non-explicit – Provide space – Silence' (NPSi) and 'Non-explicit – Provide space – Back channel' (NPBc). These response alternatives might be good strategies in real clinical life but were underrepresented in our set of responses. In relation to ES, one student managed to obtain no points. Learners had to move the sliders actively to rate their responses. Not moving the sliders was automatically translated into an "inappropriate" response. Therefore, it was not possible to identify whether this student decided that an answer was inappropriate or decided not to rate the response at all. To avoid ambiguity, we changed this feature of the sliders and students had to decide actively on each of the responses.

With the piloting of the SJT we aimed to test the tool according to its feasibility. As a consequence, we could not provide feedback on students' performance. Although acceptance of the SJT was high, students expressed their wish to receive feedback. There is a clear connection between assessment, feedback and continuous learning [52], which needs to be taken into consideration when implementing the SJT. For now, we would not recommend a pass-fail-decision when using the presented SJT, but rather recommend using this test as a formative assessment tool focusing on feedback alongside a communication skills training. Finally, as we aimed for a scoring system that is useful and easy to reproduce for a broad range of clinical teachers, we discovered that the scoring on the slider-scale might not be the best option. Future studies might apply a 5-point Likert-scale for each of the responses to allow a weighted scoring according to a Script Concordance Test [53] or a Graphic Rating Scale that combines the slider-scale with markers that depict 5-point Likert-scale type values [54].

4.2. Conclusion

VR-CoDES represents a feasible framework to develop a SJT for measuring medical students' emotion-handling skills. Development costs were initially high but should be made up over time because the instrument can be used repeatedly in different settings and stages of medical education. In order to help medical students to develop professional behavior, assessment needs to mimic realistic contexts [55]. The use of authentic scenarios, videos and expert panels are important components to achieve this goal. The continuous use of the SJT as a blended learning and assessment format, including feedback, will be a future step in our curriculum development efforts.

4.3. Practical implications

- A theoretical framework like VR-CoDES is a mandatory prerequisite for developing a SJT.
- Authentic real-life situations are an essential foundation for developing SJT content.
- Videos as stimulus for the SJT are costly but have a strong effect because they are authentic and highly accepted by learners.
- An expert-based score (ES) showed clearer results than a theory-based score (PSS).
- An adequate feedback structure seems to be a useful addition to a SJT.

CRediT authorship contribution statement

Tanja Graupe: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Project administration. **Martin R. Fischer:** Conceptualization, Methodology, Writing - review & editing. **Jan-Willem Strijbos:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Claudia Kiessling:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - review & editing, Supervision.

Acknowledgements

We thank all students for their willingness to participate in the study, all experts for their time and helpful feedback, all colleagues and research assistants who helped us to conduct our study and Peter Weichselbaum for proofreading this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.pec.2020.04.001.

References

[1] F. Ahrweiler, M. Neumann, H. Goldblatt, E.G. Hahn, C. Scheffer, Determinants of physician empathy during medical education: hypothetical conclusions from an exploratory qualitative survey of practicing physicians, BMC Med. Educ. 14 (2014) 122, doi:http://dx.doi.org/10.1186/1472-6920-14-122.
[2] D. Feldman-Stewart, M. Brundage, C. Tishelman, A conceptual framework for patient–professional communication: an application to the cancer context, Psychooncology 14 (10) (2005) 801–809, doi:http://dx.doi.org/10.1002/pon.950.
[3] T. Norfolk, K. Birdi, D. Walsh, The role of empathy in establishing rapport in the consultation: a new model, Med. Educ. 41 (7) (2007) 690–697, doi:http://dx.doi.org/10.1111/j.1365-2923.2007.02789.x.
[4] S.W. Mercer, W.J. Reynolds, Empathy and quality of care, Br. J. Gen. Pract. 52 (2002) 9–12.
[5] W. Ickes, Empathic accuracy, J. Pers. 61 (4) (1993) 587–610, doi:http://dx.doi.org/10.1111/j.1467-6494.1993.tb00783.x.
[6] J. Ogle, J.A. Bushnell, P. Caputi, Empathy is related to clinical competence in medical care, Med. Educ. 47 (8) (2013) 824–831, doi:http://dx.doi.org/10.1111/medu.12232.
[7] M. Neumann, F. Edelhäuser, D. Tauschel, M.R. Fischer, M. Wirtz, C. Woopen, H. Aviad, C. Scheffer, Empathy decline and its reasons: a systematic review of studies with medical students and residents, Acad. Med. 86 (8) (2011) 996–1009, doi:http://dx.doi.org/10.1097/ACM.0b013e318221e615.
[8] M. Neumann, C. Scheffer, D. Tauschel, G. Lutz, M. Wirtz, F. Edelhäuser, Physician empathy: definition, outcome-relevance and its measurement in patient care and medical education, GMS J. Med. Educ. 29 (1) (2012).
[9] S. Lelorain, A. Brédart, S. Dolbeault, S. Sultan, A systematic review of the associations between empathy measures and patient outcomes in cancer care, Psychooncology 21 (12) (2012) 1255–1264, doi:http://dx.doi.org/10.1002/pon.2115.
[10] I. Hsu, S. Saha, P.T. Korthuis, V. Sharp, J. Cohn, R.D. Moore, M.C. Beach, Providing support to patients in emotional encounters: a new perspective on missed

empathic opportunities, Patient Educ. Couns. 88 (3) (2012) 436–442, doi:http://dx.doi.org/10.1016/j.pec.2012.06.015.
[11] J.L. Coulehan, F.W. Platt, B. Egener, R. Frankel, C.T. Lin, B. Lown, W.H. Salazar, "Let me see if I have this right . . . ": words that help build empathy, Ann. Intern. Med. 135 (3) (2001) 221–227, doi:http://dx.doi.org/10.7326/0003-4819-135-3-200108070-00022.
[12] C. Zimmermann, L. Del Piccolo, J. Bensing, S. Bergvik, H. De Haes, H. Eide, I. Fletcher, C. Goss, C. Heaven, G. Humphris, Y.M. Kim, W. Langewitz, L. Meeuwesen, M. Nuebling, M. Rimondin, P. Salmon, S. Dulmen, L. Wissow, L. Zandbelt, A. Finset, Coding patient emotional cues and concerns in medical consultations: the Verona coding definitions of emotional sequences (VR-CoDES), Patient Educ. Couns. 82 (2) (2011) 141–148, doi:http://dx.doi.org/10.1016/j.pec.2010.03.017.
[13] H. Eide, T. Eide, T. Rustøen, A. Finset, Patient validation of cues and concerns identified according to Verona coding definitions of emotional sequences (VR-CoDES): a video-and interview-based approach, Patient Educ. Couns. 82 (2) (2011) 156–162, doi:http://dx.doi.org/10.1016/j.pec.2010.04.036.
[14] L. Del Piccolo, A. Finset, A.V. Mellblom, M. Figueiredo-Braga, L. Korsvold, Y. Zhou, C. Zimmermann, G. Humphris, Verona coding definitions of emotional sequences (VR-CoDES): conceptual framework and future directions, Patient Educ. Couns. 100 (12) (2017) 2303–2311, doi:http://dx.doi.org/10.1016/j.pec.2017.06.026.
[15] H. Ortwein, A. Benz, P. Carl, S. Huwendiek, T. Pander, C. Kiessling, Applying the Verona coding definitions of emotional sequences (VR-CoDES) to code medical students' written responses to written case scenarios: some methodological and practical considerations, Patient Educ. Couns. 100 (2) (2017) 305–312, doi:http://dx.doi.org/10.1016/j.pec.2016.08.026.
[16] J.M. Hemmerdinger, S.D. Stoddart, R.J. Lilford, A systematic review of tests of empathy in medicine, BMC Med. Educ. 7 (1) (2007) 24, doi:http://dx.doi.org/10.1186/1472-6920-7-24.
[17] M.H. Davis, Measuring individual differences in empathy: evidence for a multidimensional approach, J. Pers. Soc. Psychol. 44 (1) (1983) 113–126, doi:http://dx.doi.org/10.1037/0022-3514.44.1.113.
[18] M. Hojat, S. Mangione, T.J. Nasca, M.J. Cohen, J.S. Gonnella, J.B. Erdmann, J. Veloski, M. Magee, The Jefferson Scale of Physician Empathy: development and preliminary psychometric data, Educ. Psychol. Meas. 61 (2) (2001) 349–365.
[19] J. Turner, M. Dankoski, Objective structured clinical exams: a critical review, Fam. Med. 40 (8) (2008) 574–578.
[20] J. van Dalen, E. Kerkhofs, G.M. Verwijnen, B.W. van Knippenberg-van den Berg, H.A. van den Hout, A.J. Scherpbier, C.P. van der Vleuten, Predicting communication skills with a paper-and-pencil test, Med. Educ. 36 (2002) 148–153, doi:http://dx.doi.org/10.1046/j.1365-2923.2002.01066.x.
[21] G.M. Humphris, S. Kaney, The objective structured video exam for assessment of communication skills, Med. Educ. 34 (11) (2000) 939–945, doi:http://dx.doi.org/10.1046/j.1365-2923.2000.00792.x.
[22] F. Patterson, V. Ashworth, L. Zibarras, P. Coan, M. Kerrin, P. O'Neil, Evaluations of situational judgement tests to assess non-academic attributes in selection, Med. Educ. 46 (9) (2012) 850–868, doi:http://dx.doi.org/10.1111/j.1365-2923.2012.04336.x.
[23] M.A. McDaniel, N.T. Nguyen, Situational judgment tests: a review of practice and constructs assessed, Int. J. Sel. Assess 9 (1–2) (2001) 103–113, doi:http://dx.doi.org/10.1111/1468-2389.00167.
[24] N.T. Nguyen, M.A. McDaniel, Response instructions and racial differences in a situational judgment test, J. Hum. Resour. Manag. Res. 8 (1) (2003) 33–44.
[25] M.A. McDaniel, N.S. Hartman, D.L. Whetzel, W.L. Grubb III, Situational judgment tests, response instructions, and validity: a meta-analysis, Pers. Psychol. 60 (1) (2007) 63–91, doi:http://dx.doi.org/10.1111/j.1744-6570.2007.00065.x.
[26] M.S. Christian, B.D. Edwards, J.C. Bradley, Situational judgment tests: constructs assessed and a meta-analysis of their criterion-related validities, Pers. Psychol. 63 (1) (2010) 83–117, doi:http://dx.doi.org/10.1111/j.1744-6570.2009.01163.x.
[27] S.J. Motowidlo, A.C. Hooper, H.L. Jackson, Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items, J. Appl. Psychol. 91 (4) (2006) 749–761, doi:http://dx.doi.org/10.1037/0021-9010.91.4.749.
[28] S.J. Motowidlo, M.E. Beier, Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test, J. Appl. Psychol. 95 (2) (2010) 321–333, doi:http://dx.doi.org/10.1037/a0017975.
[29] F. Lievens, F. Patterson, 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. 96 (5) (2011) 927–940, doi:http://dx.doi.org/10.1037/a0023496.
[30] F. Patterson, E. Rowett, R. Hale, M. Grant, C. Roberts, F. Cousans, S. Martin, The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia, BMC Med. Educ. 16 (1) (2016) 87, doi:http://dx.doi.org/10.1186/s12909-016-0606-4.
[31] D. Chan, N. Schmitt, Situational judgment and job performance, Hum. Perform. 15 (3) (2002) 233–254, doi:http://dx.doi.org/10.1207/S15327043HUP1503_01.
[32] R.E. Ployhart, M.G. Ehrhart, Be careful what you ask for: effects of response instructions on the construct validity and reliability of situational judgment tests, Int. J. Sel. Assess. 11 (1) (2003) 1–16, doi:http://dx.doi.org/10.1111/1468-2389.00222.
[33] A. Koczwara, F. Patterson, L. Zibarras, M. Kerrin, B. Irish, M. Wilkinson, Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection, Med. Educ. 46 (4) (2012) 399–408, doi:http://dx.doi.org/10.1111/j.1365-2923.2011.04195.x.
[34] A. Husbands, M.J. Rodgerson, J. Dowell, F. Patterson, Evaluating the validity of an integrity-based situational judgement test for medical school admissions, BMC Med. Educ. 15 (1) (2015) 144, doi:http://dx.doi.org/10.1186/s12909-015-0424-0.
[35] F. Patterson, L. Zibarras, V. Ashworth, Situational judgement tests in medical education and training: research, theory and practice: AMEE Guide No. 10, Med. Teach. 38 (1) (2016) 3–17, doi:http://dx.doi.org/10.3109/0142159X.2015.1072619.
[36] F. Lievens, P.R. Sackett, The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance, J. Appl. Psychol. 97 (2) (2012) 460–468, doi:http://dx.doi.org/10.1037/a0025741.
[37] D.L. Whetzel, M.A. McDaniel, N.T. Nguyen, Subgroup differences in situational judgment test performance: a meta-analysis, Hum. Perform. 21 (3) (2008) 291–309, doi:http://dx.doi.org/10.1080/08959280802137820.
[38] K. Woolf, H.W. Potts, I.C. McManus, Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis, Br. Med. J. 342 (2011) d901, doi:http://dx.doi.org/10.1136/bmj.d901.
[39] R. Wakeford, M. Denney, K. Ludka-Stempien, J. Dacre, I.C. McManus, Cross-comparison of MRCGP & MRCP (UK) in a database linkage study of 2,284 candidates taking both examinations: assessment of validity and differential performance by ethnicity, BMC Med. Educ. 15 (1) (2015) 1, doi:http://dx.doi.org/10.1186/s12909-014-0281-2.
[40] M. Luschin-Ebengreuth, H.P. Dimai, D. Ithaler, H.M. Neges, G. Reibnegger, Situational judgment test as an additional tool in a medical admission test: an observational investigation, BMC Res. Notes 8 (1) (2015) 81, doi:http://dx.doi.org/10.1186/s13104-015-1033-z.
[41] C. Roberts, T. Clark, A. Burgess, M. Frommer, M. Grant, K. Mossman, The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training, BMC Med. Educ. 14 (1) (2014) 169, doi:http://dx.doi.org/10.1186/1472-6920-14-169.
[42] B.D. Goss, A.T. Ryan, J. Waring, T. Judd, N.G. Chiavaroli, R.C. O'Brien, G.J. McColl, Beyond selection: the use of situational judgement tests in the teaching and assessment of professionalism, Acad. Med. 92 (6) (2017) 780–784, doi:http://dx.doi.org/10.1097/ACM.0000000000001591.
[43] J.C. Flanagan, The critical incident technique, Psychol. Bull. 51 (4) (1954) 327–357.
[44] L.D. Butterfield, W.A. Borgen, N.E. Amundson, A.S.T. Maglio, Fifty years of the critical incident technique: 1954–2004 and beyond, Qual. Res. 5 (4) (2005) 475–497.
[45] M. Verbeke, D. Schrans, S. Deroose, J. De Maeseneer, International classification of primary care (ICPC-2): an essential tool in the EPR of the GP, Stud. Health Technol. Inform. 124 (2006) 809.
[46] L. Del Piccolo, H. De Haes, C. Heaven, J. Jansen, W. Verheul, J. Bensing, S. Bergvik, M. Deveugele, H. Eide, J. Flechter, C. Goss, G. Humphris, Y.M. Kim, W. Langewitz, M.A. Mazzi, T. Mjaaland, F. Moretti, M. Nübling, M. Rimondini, P. Salmon, T. Sibbern, I. Skre, S. van Dulmen, L. Wissow, B. Young, L. Zandbelt, C. Zimmermann, A. Finset, Development of the Verona coding definitions of emotional sequences to code health providers' responses (VR-CoDES-P) to patient cues and concerns, Patient Educ. Couns. 82 (2) (2011) 149–155, doi:http://dx.doi.org/10.1016/j.pec.2010.03.017.
[47] A.B. Simonsohn, M.R. Fischer, Evaluation of a case-based computerized learning program (CASUS) for medical students during their clinical years, DMW 129 (11) (2004) 552–556, doi:http://dx.doi.org/10.1055/s-2004-82054.
[48] I. Preusche, M. Wagner-Menghin, Rising to the challenge: cross-cultural adaptation and psychometric evaluation of the adapted German version of the Jefferson Scale of Physician Empathy for Students (JSPE-S), Adv. Health Sci. Educ. 18 (4) (2012) 573–587, doi:http://dx.doi.org/10.1007/s10459-012-9393-9.
[49] S.M. Downing, Validity. On the meaningful interpretation of assessment data, Med. Educ. 37 (9) (2003) 830–837, doi:http://dx.doi.org/10.1046/j.1365-2923.2003.01594.x.
[50] M. Hojat, Empathy in Health Professions Education and Patient Care, Springer, NY, New York, 2016, doi:http://dx.doi.org/10.1007/978-3-319-27625-0.
[51] M. Hojat, M.J. Vergare, K. Maxwell, G. Brainard, S.K. Herrine, G.A. Isenberg, J. Veloski, J.S. Gonnella, The devil is in the third year: a longitudinal study of erosion of empathy in medical school, Acad. Med. 84 (9) (2009) 1182–1191, doi:http://dx.doi.org/10.1097/ACM.0b013e3181b17e55.
[52] J. Norcini, B. Anderson, V. Bollela, V. Burch, M.J. Costa, R. Duvivier, R. Galbraith, R. Hays, A. Kent, V. Perrott, T. Roberts, Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 conference, Med. Teach. 33 (3) (2011) 206–214, doi:http://dx.doi.org/10.3109/0142159X.2011.551559.
[53] B. Charlin, L. Roy, C.G. Brailovsky, F. Goulet, C. van der Vleuten, The Script Concordance test: a tool to assess the reflective clinician, Teach. Learn. Med. 12 (4) (2000) 189–195, doi:http://dx.doi.org/10.1207/S15328015TLM1204_5.
[54] E. Svensson, Comparison of the quality of assessments using continuous and discrete ordinal rating scales, Biom. J. 42 (4) (2000) 417–434, doi:http://dx.doi.org/10.1002/1521-4036(200008)42:4<417::AID-BIMJ417>3.0.CO;2-Z.
[55] D.C. Taylor, H. Hamdy, Adult learning theories: implications for learning and teaching in medical education: AMEE guide No. 83, Med. Teach. 35 (11) (2013) e1561–e1572, doi:http://dx.doi.org/10.3109/0142159X.2013.828153.