

ORIGINAL ARTICLE

Assessing Empathy in Anesthesia Residents with the Chinese Version of the Consultation and Relational Empathy: A Pilot Standardized Patient Program

Yahong Gong, Lijian Pei, Xia Ruan, Xu Li,
Xuerong Yu, Ruiying Wang, Weijia Wang,
Gang Tan, Yuguang Huang*

Department of Anesthesiology, Peking Union Medical College Hospital,
Chinese Academy of Medical Sciences & Peking Union
Medical College, Beijing 100730, China

Key words: empathy; standardized patient encounter; assessment; rating modality

Objectives To validate the reliability of the Chinese version of the Consultation and Relational Empathy (CARE) in physician-standardized patient (SP) encounter. We also tried to examine the agreement between video-based ratings and in-room ratings, as well as the agreement between the faculty ratings and SP ratings.

Methods The CARE was translated into Chinese. Forty-eight anesthesia residents were recruited to make preoperative interview in SP-counter. Performance of each resident was graded by in-room raters, video raters and SP raters. Consistency between different raters was examined.

Results The Chinese-CARE measure demonstrated high scale reliability with a Cronbach's alpha value of 0.95 and high consistency in the in-room ratings in intraclass correlation (coefficient=0.888, $P<0.001$). Despite a good consistency in intraclass correlation, video ratings were significantly higher than in-room ratings (39.6 ± 7.1 vs. 24.0 ± 10.0 , $P<0.001$), and Wilcoxon signed-rank test indicated that the pass/fail rate was significantly higher based on video ratings than based on in-room ratings (45/48 vs. 22/48, $P<0.001$). SP ratings had a moderate consistency with in-room faculty ratings (coefficient=0.568, $P<0.001$), and there was no significant difference between the pass/fail rates based on the in-room ratings and SP ratings (22/48 vs. 28/48, $P=0.12$).

Conclusions The Chinese-CARE measure is reliable in the assessment of empathy during preoperative anesthesia interview. In-room and video ratings are not equivalent, while SP may provide a feasible alternative for the faculty rater in the assessment of communication skills with an appropriate measure.

Received August 12, 2020. Published online December 1, 2020.

*Corresponding author Tel: 86-10-69152001, Fax: 86-10-69155591, E-mail: garypumch@163.com

The study was funded by the Undergraduate Teaching Reform Project from Peking Union Medical College (Grant number: 2017zlgc0102).

EMPATHY is considered to be a basic component of therapeutic relationships^[1] and a key factor in patient's definitions of quality of care.^[2, 3] Empathy training for physicians has been demonstrated to enhance the doctor-patient relationships and improve patients' satisfaction.^[4] However, evaluation of empathy would be a big challenge without a reliable assessment tool. In 2004, Mercer developed and initially validated a measure in general practice called the Consultation and Relational empathy (CARE) measure, and they proved the validity and reliability of the CARE measure as a tool for measuring patients' perceptions of relational empathy in the consultation.^[5] They have also proved the ability of the CARE measure to discriminate effectively between doctors, in a large study of over 3000 patients in general practice.^[6] CARE has also been validated and translated into various languages and is being used by researchers in several other countries.^[7-10] Fung *et al.*^[11] made a Cantonese version of CARE and tested its validity and reliability in a primary care setting. While there are many differences in the culture and language between Hongkong and Mainland China, the Cantonese version of CARE could not be used directly in Mainland China.

Simulation-based assessment with standardized patients (SPs) has become a standard in evaluating technical and non-technical skills.^[12] It has been used as a component of some high stakes and licensure examinations.^[13] In our center, SP encounter has been long used in the training and appraisal of history-taking and physical examination skills for medical students. In the last two years, we adopted SP encounter in the training of communication skills including showing empathy. But a lack of reliable measurement made it difficult to assess the efficacy of training.

Besides a reliable tool, reliability and accuracy of the evaluation is also important to an accurate assessment. Video-based rating has gained popularity in the last few years because of limited resources to conduct in-room observations, but little is known about the equivalency of student performance scores using video-based ratings compared with in-room ratings. SP rating is a possible alternative to in-room faculty rating, but its reliability is not approved.

In this study, we translated the CARE measure into Chinese and tested its reliability in a physician-SP encounter. Meanwhile we examined the potential differences between video-based ratings and in-room ratings, as well as the differences between the faculty and SP ratings.

PATIENTS AND METHODS

The study was approved by the institutional review board of Peking Union Medical College Hospital and conducted at Peking Union Medical College Hospital.

Translation of the CARE measure

The CARE measure comprises 10 questions measuring empathy in the setting of the therapeutic rapport during a one-on-one consultation between a clinician and a patient. Each item uses a 5-point scale ranging from 'poor' to 'excellent', generating values between 1 and 5. There is also a choice 'does not apply' for each item. We translated the CARE measure into a Chinese version according to Streiner and Norman.^[14] Two bilingual (Chinese/English) people with Chinese as a mother tongue translated the scale into English, and another two bilingual (Chinese/English) people with English as a mother tongue then back-translated the draft to English. The differences between the separate versions were analyzed to get the first Chinese draft. This draft was then revised by a panel of experts with rich experience in medical education through Delphi technique. All the experts, including 6 anesthesiologists, 2 psychologists and 4 SPs, completed three rounds of iterative process. Questionnaire used in the Delphi process was shown in **Table 1**. Modifications were made based on the feedbacks after each round. On the 4th round, they agreed that all the items were accurately translated and the wording could be easily understood. Then the revised Chinese-CARE was tested with 25 preoperative patients in different age categories (18-75 years of age) and sex (10 females and 15 males). The patients scheduled for elective surgery were sequentially enrolled and written informed consents were obtained. The highest education was junior middle school for 3 patients, high middle school for 3 patients, senior school for 9 patients, and college education for 10 patients. The majority (92%) found the questionnaire generally easy to understand, and minor wording modifications were made based on their feedback.

SP encounter and assessment of empathy

A total of 48 second- and third-year residents of anesthesiology were recruited in the study. They were asked to make a preoperative interview with a 65-year old woman (acted by SP) with a huge retroperitoneal mass. The patient was scheduled to have an open surgery for tumor resection, and massive bleeding and

Table 1. Questionnaire used by the experts during the Delphi process

CARE measure item	Evaluation on the translation of each item			
	Poor	Good	Perfect	Any suggestions on revision
1. Making you feel at ease				
2. Letting you tell your story				
3. Really listening				
4. Being interested in you as a whole person				
5. Fully understand your concerns				
6. Showing care and compassion				
7. Being positive				
8. Explain things clearly				
9. Helping you take control				
10. Making a plan of action with you				

CARE: Consultation and Relational Empathy.

hemorrhagic shock were expected during the surgery. As the patient was a widow, she was wavering on the surgery in fear of being maimed after the surgery. The residents were required to explain the potential risks of anesthesia and surgery to her, comfort her and get a written informed consent. Residents were informed that the exam was a formative assessment, ratings would be de-identified, and no results would be shared with faculty or administrators.

Four experienced faculty raters with over 8 years of educational evaluation were recruited in the study. Two were assigned to the in-room ratings and another two to video ratings. The in-room raters sat 2 meters away from the encounter, side facing the residents who were seated facing the SP across a small table. The resident's and SP's faces were visible to the raters. Raters and residents were instructed not to move from their seats. Video recordings were captured with the camera positioned between the two raters. The in-room and video raters had a similar visual perspective. The interviewed SPs (two SPs with 3-5 years' experience of educational evaluation) also completed the same rating scale of CARE after each encounter.

Prior to the exam, the in-room faculty raters received one-hour training as a group on the scoring standard of each item. The video faculty raters received the same rater training and had prior rating experience as an in-room faculty rater. The SPs had the trainings and experiences of medical assessment, but did not receive the pre-exam rater training. To closely simulate conditions of in-room rating, the video rater viewed each encounter once and did not replay any video when scoring residents.

Rating forms for each resident were completed

independently on paper by each in-room rater. Video ratings and SP ratings were also completed on paper. The consistency of the individual resident scores between the two in-room raters and between the two video raters was examined. The consistency of the individual resident scores between the in-room and the video raters and between the faculty and SP raters was also examined. The average scores of the two in-room raters were regarded as the faculty ratings.

Data analysis

Ratings for each resident were then recorded into Microsoft Excel (Micro-soft, Redmond, WA) and analyzed using SPSS version 23 (IBM Corp., Armonk, NY). Cronbach's alpha was applied to evaluate the internal reliability of the Chinese-CARE. We considered the scores of the in-room raters as the gold standard for rating, and Cronbach's alpha value was calculated based on the scores of in-room raters. A Cronbach's alpha larger than 0.8 was regarded as evidence for good internal reliability.

Intraclass correlation coefficient was performed to determine the consistency of ratings between different raters. An intraclass correlation coefficient larger than 0.85 was regarded as good consistency and an intraclass correlation coefficient less than 0.4 was regarded as poor consistency. A Wilcoxon signed-rank test was conducted to compare pass/fail rates between different raters. A 2-sided *P* value less than 0.05 was regarded as statistically significant.

RESULTS

Totally, 48 residents participated in the study with

10 male and 38 female residents. Two in-room raters, two video raters and two SP raters submitted their independent ratings for the 48 residents. No missing item was found in all the ratings. In-room ratings on the 10 items of the Chinese-CARE measure were shown in **Table 2**. Over 50% residents scored the minimum score for the 9th and 10th items, demonstrating a floor effect. The frequency of endorsement was skewed towards the three lowest ratings (poor, fair and good) for the items from 3 to 8.

The internal reliability of the items of Chinese-CARE was high, with an overall Cronbach's alpha value of 0.95, which is equal to the original CARE (with a Cronbach's alpha value of 0.93). The Skewness of the ratings of two in-room raters was 0.657 and 0.710 separately, and the Kurtosis of the ratings of two in-room raters was -0.086 and -0.383.

Intraclass correlation coefficient revealed a good consistency in ratings between the two in-room raters (intraclass correlation coefficient=0.888, $P<0.001$), and a moderate consistency in ratings between the two video raters (intraclass correlation coefficient=0.516, $P<0.001$) (**Table 3**). Intraclass correlation coefficient

also indicated a good consistency in ratings between the in-room raters and video raters (intraclass correlation coefficient=0.767, $P<0.001$) and a moderate consistency in ratings between the in-room raters and SP raters (intraclass correlation coefficient=0.568, $P<0.001$) (**Table 3**).

When pass is determined by average score of 25 out of 50, there were 22 residents out of 48 passed the exam based on in-room ratings, 28 residents out of 48 based on SP ratings, 45 out of 48 based on video ratings. Wilcoxon signed-rank test indicated that there were no significant difference in the pass/fail rates based on the in-room ratings and SP ratings ($P=0.12$), but the pass/fail rates were significantly higher based on video ratings than that based on in-room ratings ($P<0.001$).

DISCUSSION

This study was conducted to test the reliability of the Chinese version of CARE and examine agreement between different ratings. The results revealed a high scale reliability of Chinese-CARE and high inter-rater

Table 2. Number of anesthesia residents who scored respective response option [n (%)]

CARE measure items	Poor (No.1/No.2)	Fair (No.1/No.2)	Good (No.1/No.2)	Very good (No.1/No.2)	Excellent (No.1/No.2)	Not applicable (No.1/No.2)	Total (No.1/No.2)
Making you feel at ease	9 (18.7)/ 6 (12.5)	14 (29.2)/ 14 (29.2)	16 (33.3)/ 13 (27.1)	6 (12.5)/ 8 (16.7)	3 (6.3)/ 7 (14.6)	0/0	48/48
Letting you tell your story	3 (6.3)/ 5 (10.4)	20 (41.7)/ 17 (35.4)	15 (31.2)/ 10 (20.8)	8 (16.7)/ 12 (25.0)	2 (4.2)/ 4 (8.3)	0/0	48/48
Really listening	13 (27.1)/ 12 (25.0)	15 (31.2)/ 8 (16.7)	12 (25.0)/ 8 (16.7)	5 (10.4)/ 12 (25.0)	3 (6.3)/ 8 (16.7)	0/0	48/48
Being interested in you as a whole person	12 (25.0)/ 13 (27.1)	15 (31.2)/ 14 (29.2)	12 (25.0)/ 12 (25.0)	7 (14.6)/ 4 (8.3)	2 (4.2)/ 5 (10.4)	0/0	48/48
Fully understand your concerns	13 (27.1)/ 9 (18.8)	18 (37.5)/ 16 (33.3)	9 (18.8)/ 13 (27.1)	5 (10.4)/ 6 (12.5)	3 (6.3)/ 4 (8.3)	0/0	48/48
Showing care and compassion	13 (27.1)/ 14 (29.2)	13 (27.1)/ 11 (22.9)	14 (29.2)/ 9 (18.8)	6 (12.5)/ 9 (18.8)	2 (4.2)/ 5 (10.4)	0/0	48/48
Being positive	12 (25.0)/ 17 (35.4)	13 (27.1)/ 13 (27.1)	13 (27.1)/ 10 (20.8)	9 (18.8)/ 4 (8.3)	1 (2.1)/ 4 (8.3)	0/0	48/48
Explain things clearly	16 (33.3)/ 11 (22.9)	16 (33.3)/ 22 (45.8)	4 (8.3)/ 3 (6.3)	8 (16.7)/ 7 (14.6)	4 (8.3)/ 5 (10.4)	0/0	48/48
Helping you take control	28 (58.3)/ 23 (47.9)	12 (25.0)/ 15 (31.2)	4 (8.3)/ 5 (10.4)	2 (4.2)/ 3 (6.3)	2 (4.2)/ 2 (4.2)	0/0	48/48
Making a plan of action with you	28 (58.3)/ 23 (47.9)	6 (12.5)/ 11 (22.9)	3 (6.3)/ 8 (16.7)	11 (22.9)/ 5 (10.4)	0 (0)/ 1 (2.1)	0/0	48/48

No. 1 indicates No. 1 in-room faculty rater; No.2 indicates No. 2 in-room faculty rater.

Table 3. Intraclass correlation coefficient between the ratings of each resident's performance by different raters and different modalities

Items	Score (mean±SD)	Intraclass correlation	
		Coefficient	P-value
In-room rater No. 1	23.2±9.7	0.888	<0.001
In-room rater No. 2	24.8±10.8		
Video rater No. 1	40.6±7.1	0.516	<0.001
Video rater No. 2	38.6±8.9		
Standardized patient rater	28.9±9.2	0.568	<0.001
Mean score of in-room raters	24.0±10.0		
Mean score of video raters	39.6±7.1	0.767	<0.001
Mean score of in-room raters	24.0±10.0		

correlation between in-room raters during the assessments, which is consistent with the English CARE, indicating that the Chinese CARE has been accurately translated and is suitable to be used in Chinese culture. And there was a good equivalence between the in-room faculty ratings and SP ratings during the assessments, although the SPs did not receive the pre-exam training. In other words, consistent inter-rater agreement between the faculty and SP was achieved even though the SPs did not know the scoring standard of CARE, highlighting that CARE was a good patient-centered measure.

Despite a good consistency, there was no equivalence between the in-room ratings and video ratings in current investigation. This concurs with previous patient-physician consultation studies which reported that the in-room ratings and video ratings did not result in equivalent pass/fail decisions despite high scale reliability and interclass correlation.^[15] However, the conclusions were different in studies of assessing procedural skills such as joint examination and airway insertion, which reported high congruence between in-room and video ratings.^[16, 17] We assumed that video-based rating might not be as reliable in communication skill assessment as in procedural skill assessment. There are several potential explanations for this. First, communication skills (verbal and nonverbal) are not easily captured on camera. The in-room rater was closer and had greater access to the finer nuances of communication and the connection between resident and SP, they could better perceive and rate those behaviors. Second, the in-room rater could get more information on body language as they had opportunity to change the observation perspective during the live

encounter without changing of the seated position, whereas the camera angle captured only one fixed distant perspective for the video rater.

Inter-rater variability within the in-room raters and video-raters were also perceived, and the reasons might be as follows: (1) A 5-point competency scale for communication assessment requires more judgment than the yes/no checklists used for procedural assessment, which resulted in a bigger variation between different raters. (2) The faculty's own clinical skills may impact their ratings for residents. While the assessment of empathy is in lack of a unified standard, their ratings could rely more on their own skills.

The reliability of SP rating is still controversial. Some studies showed that SPs who were not trained on how to consistently grade communication skills, resulting in a weak correlation among graders for the communication section despite overall total score correlation.^[17] Another study found moderate to substantial reliability between SP ratings and second rater ratings.^[18] In our study, we have found a moderate consistency between the SP ratings and in-room faculty ratings, and pass/fail rates were parallel based on the two ratings. We suggested that a well-trained SP could discriminate between high and low performing students just as well as the faculty raters in assessment of communication skills.

Our study has several limitations. First, as a pilot study, the number of raters was relatively limited to determine the inter-rater differences, and further studies were also needed to find the potential reasons for the differences. Second, although all the faculty raters had a pre-assessment training on how to score each of the items, the oral description of the scoring standard was very vague. We assume that if standardized encounter cases graded as excellent, good and poor for residents' performance be videotaped and shown to the raters as a reference, the raters could get a better understanding of the scoring standard. Third, the SPs in our study have been well trained to make educational assessment. Although they have no medical background before becoming an SP, their judgment might be influenced by doctors in previous experiences. Their ratings may not resemble exactly the real patients' perception.

In conclusion, the preliminary data in our study support the reliability of the Chinese version of the CARE measure in preoperative anesthesia interview in China. This study also brings us a step toward

understanding the non-equivalence of in-room and video ratings in the assessment of communication skills. SPs may provide a feasible alternative for faculty rater in the assessment of communication skills with an appropriate measure.

Acknowledgements

We would like to thank Haobo Ma from Harvard Medical School for helping us in study design, and Yuelun Zhang from Peking Union Medical College Hospital for helping us in data analysis.

Conflicts of Interest Statement

The authors have no competing interest to declare.

REFERENCES

1. Reynolds W, Scott B. Empathy: a crucial component of the helping relationship. *J Psychiatr Ment Health Nurs* 2010; 6(5):363-70. doi: 10.1046/j.1365-2850.1999.00228.x.
2. Menendez ME, Chen NC, Mudgal CS, et al. Physician empathy as a driver of hand surgery patient satisfaction. *J Hand Surg* 2015; 40(9):1860-5.e2. doi: 10.1016/j.jhsa.2015.06.105.
3. Chaitoff A, Sun B, Windover A, et al. Associations between physician empathy, physician characteristics, and standardized measures of patient experience. *Acad Med* 2017; 92(10):1464-71. doi: 10.1097/ACM.0000000000001671.
4. Boissy A, Windover AK, Bokar D, et al. Communication skills training for physicians improves patient satisfaction. *J Gen Inter Med* 2016; 31(7):1-7. doi: 10.1007/s11606-016-3597-2.
5. Mercer SW. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. *Fam Pract* 2004; 21(6):699-705. doi: 10.1093/fampra/cmh621.
6. Mercer SW, McConnachie A, Maxwell M, et al. Relevance and practical use of the Consultation and Relational Empathy (CARE) measure in general practice. *Fam Pract* 2005; 22(3):328-34. doi: 10.1093/fampra/cmh730.
7. Aomatsu M, Abe H, Abe K, et al. Validity and reliability of the Japanese version of the CARE measure in a general medicine outpatient setting. *Fam Pract* 2014; 31(1):118-26. doi: 10.1093/fampra/cmt053.
8. Wirtz M, Boecker M, Forkmann T, et al. Evaluation of the "Consultation and Relational Empathy" (CARE) measure by means of Rasch-analysis at the example of cancer patients. *Patient Educ Couns* 2011; 82(3):298-306. doi: 10.1016/j.pec.2010.12.009.
9. Neumann M, Wirtz M, Bollschweiler E, et al. Determinants and patient-reported long-term outcomes of physician empathy in oncology: a structural equation modelling approach. *Patient Educ Couns* 2007; 69(1-3):69-75. doi: 10.1016/j.pec.2007.07.003.
10. Fung CS, Mercer SW. A qualitative study of patients' views on quality of primary care consultations in Hong Kong and comparison with the UK CARE measure. *BMC Fam Pract* 2009; 10:10. doi: 10.1186/1471-2296-10-10.
11. Fung CS, Hua A, Tam L, et al. Reliability and validity of the Chinese version of the CARE measure in a primary care setting in Hong Kong. *Fam Pract* 2009; 26(5):398-406. doi: 10.1093/fampra/cmp044.
12. Adamo G. Simulated and standardized patients in OSCEs: achievements and challenges 1992-2003. *Med Teach* 2003; 25(3):262-70. doi: 10.1080/0142159031000100300.
13. Furman GE. The Role of standardized patient and trainer training in quality assurance for a high-stakes clinical skills examination. *Kaohsiung J Med Sci* 2008; 24(12):651-5. doi: 10.1016/S1607-551X(09)70031-2.
14. Health measurement scales: a practical guide to their development and use (5th edition). *Aust N Z J Public Health* 2016; 40(3):294-5. doi: 10.1111/1753-6405.12484.
15. Sturpe DA. Objective structured clinical examinations in doctor of pharmacy programs in the United States. *Am J Pharm Educ* 2010; 74(8):148. doi: 10.5688/aj7408148.
16. Vivekananda-Schmidt P, Lewis M, Coady D, et al. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. *Arthritis Rheum* 2007; 57(5):869-76. doi: 10.1002/art.22763.
17. Yuasa M, Nagoshi M, Oshiro-Wong C, et al. Standardized patient and standardized interdisciplinary team meeting: validation of a new performance-based assessment tool. *J Am Geriatr Soc* 2014; 62(1):171-4. doi: 10.1111/jgs.12604.
18. Bolstad AL, Xu Y, Shen JJ, et al. Reliability of standardized patients used in a communication study on international nurses in the United States of America. *Nurs Health Sci* 2012; 14(1):67-73. doi: 10.1111/j.1442-2018.2011.00667.x.

中文摘要

中文版诊疗关系共情量表（CARE）用于麻醉科住院医师共情能力的评估：一项标准化病人项目的初步研究

龚亚红，裴丽坚，阮侠，李旭，虞雪融，王锐颖，王维嘉，谭刚，黄宇光

中国医学科学院 北京协和医学院 北京协和医院麻醉科，北京 100730

摘要

目的 验证中文版诊疗关系共情量表（Consultation and Relational Empathy, CARE）在标准化病人模拟门诊中使用的可靠性。同时探索性地评估视频评分方式和现场评分方式之间的一致性，以及临床考官评分和标准化病人（physician-standardized patient, SP）评分之间的一致性。

方法 将 CARE 量表的内容翻译成中文。纳入 48 名麻醉住院医师，在标准化病人模拟门诊中进行术前访视。每位住院医师的表现由现场考官、视频考官及 SP 评分员进行评分。分析不同类别评分员之间的一致性。

结果 中文版 CARE 量表具有较高的信度（Cronbach α 值为 0.95）和良好的组内相关性（现场考官评分的组内相关系数为 0.888, $P<0.001$ ）。尽管组内相关性良好，但视频考官的评分显著高于现场考官的评分（ 39.6 ± 7.1 比 24.0 ± 10 , $P<0.001$ ）。Wilcoxon 符号秩检验表明，基于视频评分的住院医师考核通过率显著高于现场评分的通过率（45/48 比 22/48, $P<0.001$ ）。SP 评分与现场考官评分之间具有中度一致性（系数 = 0.568, $P<0.001$ ），现场考官评分和 SP 评分的考核通过率之间没有显著性差异（22/48 比 28/48, $P=0.12$ ）。

结论 中文版 CARE 量表能客观、准确评估麻醉术前访视中住院医师的共情能力。现场评分和视频评分之间存在一定的差异，而 SP 可能能够替代临床考官对住院医师的临床沟通能力进行恰当的评估。

关键词：共情；标准化病人模拟门诊；评估；评分方式