

BMJ Open How culture affects validity: understanding Japanese residents' sense-making of evaluating clinical teachers

Makoto Kikukawa ¹, Renée E Stalmeijer ², Takahiro Matsuguchi,³ Miyako Oike,⁴ Emura Sei,⁵ Lambert W T Schuwirth,^{2,6} Albert J J A Scherpbier⁷

To cite: Kikukawa M, Stalmeijer RE, Matsuguchi T, *et al.* How culture affects validity: understanding Japanese residents' sense-making of evaluating clinical teachers. *BMJ Open* 2021;**11**:e047602. doi:10.1136/bmjopen-2020-047602

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-047602>).

Received 05 December 2020
Accepted 04 August 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Makoto Kikukawa;
kikukawa.makoto.892@m.kyushu-u.ac.jp

ABSTRACT

Objectives Traditionally, evaluation is considered a measurement process that can be performed independently of the cultural context. However, more recently the importance of considering raters' sense-making, that is, the process by which raters assign meaning to their collective experiences, is being recognised. Thus far, the majority of the discussion on this topic has originated from Western perspectives. Little is known about the potential influence of an Asian culture on raters' sense-making. This study explored residents' sense-making associated with evaluating their clinical teachers within an Asian setting to better understand contextual dependency of validity.

Design A qualitative study using constructivist grounded theory.

Setting The Japanese Ministry of Health, Labour and Welfare has implemented a system to monitor the quality of clinical teaching within its 2-year postgraduate training programme. An evaluation instrument was developed specifically for the Japanese setting through which residents can evaluate their clinical teachers.

Participants 30 residents from 10 Japanese teaching hospitals with experience in evaluating their clinical teachers were sampled purposively and theoretically.

Methods We conducted in-depth semistructured individual interviews. Sensitising concepts derived from Confucianism and principles of response process informed open, axial and selective coding.

Results Two themes and four subthemes were constructed. Japanese residents emphasised the awareness of their relationship with their clinical teachers (1). This awareness was fuelled by their sense of hierarchy (1a) and being part of the collective society (1b). Residents described how the meaning of evaluation (2) was coloured by their perceived role as senior (2a) and their experienced responsibility for future generations (2b).

Conclusions Japanese residents' sense-making while evaluating their clinical teachers appears to be situated and affected by Japanese cultural values. These findings contribute to a better understanding of a culture's influence on residents' sense-making of evaluation instruments and the validity argument of evaluation.

Strengths and limitations of this study

- Through qualitative methodology, this study enables a deeper understanding of residents' response process when evaluating their clinical teachers.
- The qualitative methodology enabled a more nuanced understanding of the role of culture in residents' response process.
- Social desirability bias in residents' answer during interviews cannot be fully ruled out.
- Japan has a very distinct culture which may limit the transferability of these results to other contexts.

INTRODUCTION

'Evaluation' refers to the process of determining the amount, number or value of something.¹ It is a core component of medical education and is used in quality assurance, care for patients and improving educational programmes.^{2–5} The significance of an evaluation depends on the validity of the measurement as well as the validity of the outcome of measurement, which has been widely discussed by educators and educational researchers.⁶ 'Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests'⁷ (p9) and several frameworks exist that explain which types of validity evidence need to be gathered for an instrument to be valid^{8,9} (see online supplemental tables 1 and 2). Traditionally, psychometric models stemming from a more (post) positivistic perspective, like Messick's framework for validity,⁸ have played a major role in the validation process within evaluation.^{8–11} These (post) positivistic models consider validity mainly as residing within the instrument, without taking into account interactions between participants, cultures and physical environments.^{7,12–16} A growing acceptance of cultural diversity within psychometric models like

that of Kane⁹ has led to several cross-validation studies in many academic fields (eg, health, social and educational sciences) since the last two decades.^{17–23} However, these studies have thus far focused on the differences between different cultures and not why these differences exist and to what extent they reside in the raters' interpretation of the evaluation.

Furthermore, increasing evidence exists to support the notion that the use of a *valid* instrument—particularly if it is based on observation and interpretation—is not sufficient per se. The ability/expertise of the raters (staff members and students) is equally important to ensure that the evaluation yields meaningful outcomes.^{24–25} This perspective must be broadened by moving beyond the interrogation of validity using psychometric models to generate more meaningful evaluations.^{26–27} A theoretical lens that supports the broadening of the validity discussion is the situativity theory,^{28–30} which is a theoretical framework which poses that human's sense-making is situated in experiences; and experiences comprise the participants, the culture and the physical environment. This notion of situativity distinguishes between affordances—what the instrument will allow the rater to do—and effectivities—what the rater is able to do with the instrument.²⁴ Both need to be optimal and aligned with each other; a poorly equipped rater cannot handle a perfect instrument and an imperfect instrument does not allow an expert rater to perform well.²⁴ Evaluation of clinical teachers in postgraduate education is often based on questionnaires by residents.³ Results from teaching evaluation instruments must be supported by a variety of validity evidence to truly improve clinical teaching. Many instruments for evaluating clinical teaching have been developed and validated but almost all in Western countries.³¹ Therefore, in a previous study, we developed an instrument with content validity in Japanese clinical setting to improve clinical teaching of clinical teachers in Japan.³² The purpose of the instrument was to provide feedback to clinical teachers on their teaching skills in order to help them improve. The instrument was not intended for promotion or accreditation purposes. Although the medical education field is gradually embracing the importance of acknowledging raters' sense-making as part of the validity discussion, most validity arguments on evaluation in this field have so far originated from a Western perspective.³³ Therefore, little is known in medical education regarding non-Western cultural influences on evaluation, particularly concerning the raters' sense-making.

Therefore, in this study, we explore the influence of culture on residents' process of 'sense-making'^{8–9} associated with the evaluation activities where residents are asked to evaluate their clinical teachers. Sense-making is described as the process by which raters assign meaning to their collective experiences, which is considered as response process or scoring on validity framework.^{8–9} In other words, we explore theories that support a better understanding of the sociocultural context that frames residents' responses to questions in clinical teaching

evaluations for validation. We use the case of an Asian culture, more specifically of Japan, to study the influence of culture on the process of sense-making. Japan has been described as having a strong value of respect for seniors, which makes it inappropriate for students to evaluate teachers.³⁴ Considering this background, our research question is the following: To what extent do Japanese cultural values influence residents' sense-making of an evaluation instrument?

METHODS

Overview

This is a qualitative study that was designed to employ a constructivist grounded theory (CGT) methodology. A constructivist epistemology considers knowledge as actively constructed and co-created as a product of human interactions and relationships.^{35–36} Within the constructivist view, the goal of the research is to develop better understanding and adequate models for specifically situated purposes.³⁷ The CGT allows for an open approach to better understand the psychological processes that underlie phenomena of interest.³⁵ In our study, we explore theories that support a better understanding of the sociocultural context that frames residents' responses to questions in clinical teaching evaluations.

Setting of the study

Japanese postgraduate medical education starts with a 2-year initial postgraduate clinical training and comprises several rotations of varying length in the following disciplines: mandatory rotations in internal medicine (more than 6 months), emergency medicine (more than 3 months), community medicine (more than 1 month), and at least two rotations in the following specialties: surgery, anaesthesiology, paediatrics, obstetrics and gynaecology, and psychiatry. During this process, residents are supervised by their clinical teachers, and residents' performance and competence are assessed by their clinical teachers through work-based assessments; that is, direct observation and case-based discussion. After the training, residents can apply for a specialised training programme. Under Japanese official postgraduate training programme regulations, physicians should have more than 7 years of clinical experience and take a Faculty Development Workshop certified by the Ministry of Health, Labour and Welfare to become clinical teachers.

Many teaching hospitals introduced the evaluation of clinical teachers since a 2-year postgraduate training programme was launched in 2004, the use of the results depended on each hospital. Although it was officially introduced in 2020 by the Ministry of Health, Labour and Welfare to monitor the quality of clinical teaching, it is not always mandatory that results are given to clinical teachers currently and the instrument has not been validated.

The Japanese cultural context of this study

Definitions of what culture entails vary. Japanese culture has been discussed from a variety of perspectives,³⁸

however, Confucianism has been most influential in shaping the behavioural patterns and structures of communities within Japan.³⁹ Confucianism—established by Confucius and his disciples—is a philosophy that considers proper human relationships as the basis of society. There are five virtues that have to be aimed for, namely, Ren (benevolence, humanism), Yi (integrity, uprightness), Li (rite, propriety and decorum), Chi (moral understanding, wisdom), and Shin (trust). Fostered by Confucian values are the five principal relationships through which each person defines identity, duty and responsibility.^{40 41} The five basic relationships are as follows: loyalty between the ruler and subject (government and citizen), closeness between father and son, distinction in duty between husband and wife, obedience toward others (between elders and youngsters), and mutual faith between friends.⁴²

Confucianism influences Japanese orientations toward high power distance and collectivism significantly.⁴³ In a culture influenced by Confucianism, high power distance is accentuated via communication behaviours that support hierarchical relationships. In the education field, teachers are expected to endeavour to pursue an ideal. In addition, youngsters should treat their elders with due respect. Power distance relates to the degree of inequality in power between a less powerful Individual (I) and a more powerful Other (O), in which *I* and *O* belong to the same social system.⁴⁴ Individualism, a value more highly upheld in Western countries,⁴³ emphasises the particular self and its uniqueness, while collectivism, conversely, is the opposite where members of a collectivistic society play a role in group interests, conformity and cohesion. Therefore, Japanese culture has been characterised by the roles that hierarchy and collectivity play within day-to-day life.⁴³

Patient and public involvement

Patients or the public were not involved in the design, conduct, reporting or dissemination plans of this study.

Data collection

The present study undertook a purposive sampling of Japanese residents engaged in 2-year initial postgraduate clinical training who had had experience with using an evaluation instrument to evaluate their clinical teachers on at least three occasions (see online supplemental appendix 1).³² The content validity of this Japanese instrument was established in a previous study.³² The instrument was sent to participating hospitals with a written explanation asking residents to anonymously fill out the questionnaire on paper for their clinical teachers. At this point in the validation of the instrument, results of the evaluation were not yet communicated to clinical teachers as full validity of the instrument was yet to be established. Questionnaires were returned to the researchers by mail. We asked key informants from teaching hospitals to recruit residents. Prospective participants received either a letter or an email from the researchers, in which the purpose of the study was explained, and a request to participate

in the interview was affixed. MK visited their hospitals and conducted face-to-face, in-depth and semistructured interviews in accordance with an interview guide with all individual residents in Japanese (see online supplemental appendix 2). All residents gave written informed consent after their own review of the study details. They were also informed that they could, at any time, freely withdraw from the study. We further explained that all content of interviews would be treated anonymously. The residents received a small financial incentive (2000 yen≈€16; exchange rate as of 28 September 2020). All interviews were audio-recorded and transcribed verbatim. The anonymity of transcripts and discussions was exclusive within the research team, and strict confidentiality was maintained. No member of the research team was involved in the residency programmes to which the participating residents belonged. MK used the interview guide as a starting point and pursued interesting new information that emerged in the process. This new information was included in the subsequent interviews. Analyses from open coding to selective coding were informed by sensitising concepts of the Japanese cultural perspective (Confucianism), and response process as explained above.

Consistent with CGT, the process of data collection and analysis was iterative, with analysis occurring alongside and informing data collection: themes that were identified in the examination of initial transcripts were explored in more depth in subsequent interviews. This process included adding questions to elaborate on emerging areas and theoretically sampling missing voices of residents.

Data analysis

MK and TM independently used open coding and constant comparative techniques to develop a preliminary coding structure.³⁶ After coding three interviews, MK and TM compared the findings and discussed differences until a consensus was reached, after which MK continued the analysis and could identify recurring themes. Regular discussion of coding and interpretation of the data via Skype were done with MK and TM, after which coding was discussed with MO, an expert qualitative researcher. Codes were established in an iterative fashion, in accordance with the principles of Confucianism. Coding informed subsequent discussions with the Japanese research team members.

Through constant comparison of codes and themes, the data were organised into aggregate themes. Explanatory memos were iteratively developed and tested by MK, TM and MO to link memos to the axial coding model and original data. Interviews and data analyses were conducted until theoretical saturation was met. MK, TM and MO inductively examined the interpretation from the themes to the data to deductively confirm the interpreted data.

These interviews were conducted in 10 teaching hospitals between March 2014 and March 2015. Analysis of

interviews attained saturation after 25 interviews. The saturation was confirmed after performing additional five other interviews, which did not result in the construction of new themes. Thus, 30 interviews in total were performed.

Finally, five of the participating residents were asked to read and comment on the results of the analysis to determine whether the data and conclusions accurately reflected the content of the interviews, member checking.⁴⁵ All residents agreed with the interpretations.

After the analysis, MK translated the results and the illustrative quotes into English with an expert bilingual translator. MK explained his interpretation of the results to RS, AJJAS and LWTS, who are non-Asian coauthors, via email and/or Skype. MK discussed with them how the themes appeared to be associated with the Asian perspective. Through discussion between Asian and non-Asian authors, we attempted to understand how Confucianist values impacted the process of residents' sense-making. Finally, all authors discussed and agreed with the results and the interpretations.

RESULTS

There were 18 male and 13 female residents aged 25–35 years (mean: 26 years). They had 9–24 months (mean: 22 months) of experience in training and 3–20 sessions (mean: 4 sessions) of evaluation before our interviews.

We identified two themes and four subthemes: (A) awareness of relationship which was subdivided in (A1) hierarchy and (A2) collectivity, and (B) meaning of evaluation, subdivided in (B1) responsibility as a senior and (B2) resignation (see online supplemental figure 1).

During the interviews, residents provided descriptions of their sense-making while evaluating their clinical teacher. Most interestingly, residents spent the majority of their time talking about how they were continuously (A) aware of their relationship with their clinical teachers and (B) how they saw the meaning of evaluation as related to their setting. The awareness seemed to be fuelled by their sense of the hierarchy of the resident–clinical teacher relationship and being a part of the collective from a societal perspective.

Awareness of relationship

Residents were very much aware of their relationship with clinical teachers during the process of evaluation. This awareness was fuelled by a drive to build and maintain a connection with their clinical teachers.

Hierarchy

During the interviews, residents talked about ratings of clinical teachers being strongly influenced by the awareness of being part of a 'relationship between a superior and a subordinate.' Residents were aware of the 'vertical nature' of this relationship and the inequality it created. Being part of a hierarchical relationship based on seniority in Japanese society, residents mentioned the

reciprocity of that relationship—they were expected to be polite and respectful to clinical teachers (superiors) and clinical teachers owed residents' (subordinates) protection and consideration under these same social norms. This is best understood from the residents' conviction that there are social responsibilities based on five principles of Confucianism and that the relationship is maintained by actively shouldering each responsibility. The responsibilities encompass teaching for clinical teachers as superiors and respect and obedience for residents as subordinates. Residents respected clinical teachers' attitudes toward shouldering their own responsibility based on seniority. Residents particularly paid respect to clinical teachers who treated them with consideration and felt remorse that they would not fulfil their duty as dictated by their position in the hierarchy if they did not show politeness and respect.

If I rate my clinical teachers with a low grade, I feel I disdain them and would feel very guilty. (P10)

Propriety as subordinates

Through their descriptions of the evaluation process, residents demonstrated Li (propriety, decorum) as a strong value. Within the Japanese culture, propriety is a social norm that maintains social order and keeps harmony among people. Residents (subordinates in this context) considered that it would be rude to rate their clinical teachers (superiors in this context) low. Therefore, residents felt bad about evaluating teachers while receiving their guidance.

I feel bad...there must be a case where I rate them low while simultaneously considering that they were so kind to teach us, but still, I feel it would be impudent, so I evaluate them wondering if it would be okay [because I am a resident]. (P9)

Residents talked about how even the mere process of evaluation of clinical teachers in itself was disrespectful. Residents were reluctant to judge the values of teaching by clinical teachers, because the judgement itself is contrary to Confucian preaching, it is improper and it lacks decorum.

Evaluating clinical teachers is impertinent..., how dare we, the extreme beginners, evaluate clinical teachers, I'd say. (P1)

A fear to deny the clinical teachers' competence

Residents felt that if they, as subordinates, would provide a low rating of their superiors, they may infer a denial of the clinical teachers' expertise or role as a teacher (superior). Residents realised this action was against the 'proper' hierarchical relationship. Taking an attitude that does not respect clinical teachers meant that residents would destroy delicate relationships that had been established by mutual effort. Therefore, residents tended to avoid rating behaviour critically.

I believe that my clinical teacher did his/her best in teaching. I feel sorry if my clinical teacher understands that I deny him/her by my low evaluation. (P5)

Collectivity

Japanese people tend to value their role in groups rather than their role as individuals. Consequently, residents did not consider filling out an evaluation as the individual activity of either a resident or a clinical teacher but as a result of the interaction between residents and their clinical teachers. During the evaluation process, residents put themselves in the other person's (clinical teacher's) shoes. This was fuelled by a feeling of collective responsibility for the results of the evaluation and consideration for the feelings of clinical teachers.

Collective responsibility for the results of the evaluation

Residents acknowledged that the results of the evaluation would not only be an indication of clinical teachers' behaviours but also of their (the residents') own behaviours. From a collectivist perspective, learning is predominantly the result of an interaction between the members of a collective rather than a combination of individual performances. Therefore, clinical teachers and residents are interactively related to one another. Because of this, residents felt that rating clinical teachers was not distinguishable from rating the residents' own performances.

I think if we'd rather have to have such a thing [evaluation], it's not our clinical teachers' fault but mine. (P12)

Consideration for the feelings of clinical teachers

Residents did not seem to want to or be able to completely separate their own feelings from those of their clinical teachers. They thought that clinical teachers surely had the same feeling. Residents recognised that they do not want to do to other people what they naturally hate having done (rated a low scale in this case) to them. Through considerations for the feelings of clinical teachers, residents thought that confrontations and conflicts should be avoided in relationships with clinical teachers.

I mean...if I am a clinical teacher and have been giving my best, but despite that effort, residents say the teacher is not really proper, or something like that, it would make me upset, honestly. (P5)

Meanings of evaluation

Residents understood the meaning of evaluation in several different ways, which was influenced by their awareness of relationship. It seemed that the meaning that residents ascribed to this evaluation depended on what they were aware of more: a superior's or a subordinate's stance.

Responsibility as a senior

Evaluating their clinical teachers to improve teaching for their future subordinates was also seen as a duty. Improvement of teaching through this evaluation was a responsibility (including providing a critical evaluation) toward themselves as future seniors of future subordinates. The responsibility meant that superiors owed subordinates protection and consideration, coming from Confucianism philosophy.

I do not want my junior residents to pass through the same difficulties I passed through. So I evaluated earnestly. (P27)

Resignation

Conversely, if some residents did not see the significance or benefit of this evaluation; therefore, motivation tended to stay low.

Well, after all, I guess so. If asked whether things would change after filling this questionnaire, uh, well, honestly, I don't think in our generation, so, that is one point that lowers motivation. (P12)

This resignation seemed to come from the notion of hierarchy that superiors would not accept subordinates' opinions because clinical teachers make an effort to pursue an ideal by themselves.

If a certain young resident evaluates this one clinical teacher...I think, honestly, that the clinical teacher's basic teaching style would not change. (P5)

These subthemes illustrate the following paradox: on the one hand, evaluation with a view of improving education for the whole group and the group of future residents made the activity acceptable within the value framework. However, evaluation with a focus to just improve the teaching of a superior was not only seen as probably futile (someone in power might not change) but also as misaligned with cultural values (disrespect for seniority).

DISCUSSION

This study aimed to better understand the extent to which Japanese culture factors into residents' sense-making of an evaluation instrument and, thus, to validate the evaluation of clinical teachers, and consequently to improve clinical teaching. Our findings paint a picture of how cultural values appeared to be interwoven in residents' sense-making of evaluating clinical teachers.

The fact that the relationship between the resident and the supervisor could influence residents' rating has been previously reported in a Dutch setting by Fluit *et al.*⁴⁶ However, the exact nature of the relationship's influence on residents' ratings was not further explored in the Fluit *et al.*'s study. In our study, we further explored the influence of the relationship and found that residents' awareness of relationship appeared to be

based on Confucian philosophy that emphasises proper human relationships as a basis for society. Moreover, our results indicate that it was not just the relationship with their supervisor that influenced residents' sense-making but also the residents' relationship with society at large and more specifically those residents that would follow in their footsteps.

It might also suggest that if the raters of an instrument have different cultural perspectives, sense-making of the evaluation process and its items might differ from one country to the next, even if the scores are the same. For a deeper understanding of the exact nature of the influence of cultural contexts on the use of evaluation instruments, raters' sense-making should be explored in a variety of cultural settings.

An evaluation activity in which a resident is requested to evaluate their clinical teacher has the potential to generate a paradoxical situation for residents. Residents wanted to evaluate their clinical teachers with an outlook to improve education for future generations and their own future role as supervisors. This represents the Confucianism virtue of Shin (trust) toward their (future) subordinates. However, the act of evaluating was an act of being critical toward their supervisors. This is directly at odds with Li (rite, propriety and decorum). Therefore, residents' sense-making was neither linear nor simple. Instead, it appeared to be a complex interaction between the residents, clinical teachers, the instrument and culture, which is better explained from a situativity theory perspective²⁴ rather than a traditional psychometric model, which requires a numerical and reductionist approach.⁹ Judging from our findings, a reductionist approach might have limitations in presenting the complexity of a rating as a single number at least in Japan. Our findings support the notion of affordances of the instrument: what the instrument will allow the rater to do and the effectivities of the rater; and what the rater is able to do with the instrument are equally as important as the utility that resides in the interaction of affordances and effectivities.

Although both of Messick's five validity framework and Kane's validity inferences are comprehensive conceptualisation of validity, Kane's validity framework is often favoured⁴⁷ because his framework does not rely heavily on psychometric and numerical data and is more coherent when it comes to rational interpretation by raters who are situated in their own cultural contexts.⁹ Therefore, Kane's framework opens the door to the inclusion of more complex and situated views on validity.⁹ Kane's four inferences that bridge the gap between observation and construct can actually be interpreted from a situativity perspective, as each of the connections between affordances of the situation and effectivities of the actors in the process is essential in underpinning the clarity, coherence and plausibility of the underlying rationale for the inference.⁹ Therefore, it is not only essential to discuss the validity of evaluating clinical teachers with researchers and medical

educators who have a variety of cultural backgrounds, but also discuss the appropriateness of such an instrument within a certain context and what can be done to stimulate its meaningful implementation.

There are several limitations to this study. First, because of cultural aspects, residents might have given the interviewer socially desirable answers during the interviews. To mitigate this issue, we explained and promised that we would never disclose their answers. Additionally, the interviewer was not involved in residency training of any of the participating residents. However, although we might have taken all possible precautions against this bias, we cannot be sure that it did not occur. Second, Japan is only one of many countries that are identified as Asian countries. Other Asian countries have, of course, different cultural perspectives. Although our study was not set up to uncover the ultimate answer to all the possible influences different cultures may have on the phenomenon of questionnaire-based evaluation, we still cannot rule out the fact that the Japanese context is the single outlier and that all other cultural contexts would be aligned with each other. We therefore think that this study should be replicated in various contexts to investigate the possible culture-related impacts on evaluation practices in a variety of cultural settings. Third, our perspectives are primarily from the Confucian perspective. It is argued that Japanese culture is influenced by several religions. Investigations from different lenses would also be needed for better understandings of raters' sense-making.

CONCLUSION

We investigated how cultural contexts could influence raters' sense-making of evaluation instruments from an Asian perspective. Japanese residents' sense-making seemed to be influenced by their own culture and appeared to be situated in the experiences of interaction with their clinical teachers, the instrument and their cultural understanding of the relationship with their teacher. We have constructed a more detailed picture of the influence of relationship and the culture they are situated in on residents' sense-making. Their sense-making was not linear but complex. Our findings contribute to a better understanding of culture's influence not only on residents' sense-making of evaluation instruments but also on the validity argument of evaluation.

Author affiliations

¹Department of Medical Education, Kyushu University, Fukuoka, Japan

²Department of Educational Development and Research, Maastricht University Faculty of Health, Medicine and Life Sciences, Maastricht, The Netherlands

³Department of Gastroenterology, Kitakyushu Municipal Medical Center, Fukuoka, Japan

⁴Department of Nursing, Fukuoka International University of Health and Welfare, Fukuoka, Japan

⁵Saga Medical Career Support Center, Saga University Hospital, Saga, Japan

⁶Prideaux Centre for Research in Health Professions Education, Flinders University, Adelaide, South Australia, Australia

⁷Institute for Medical Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

Correction notice This article has been corrected since it was published. Middle initial has been added for Renée E Stalmeijer.

Twitter Lambert W T Schuwirth @LSchuworth

Acknowledgements We thank the residents who participated in the interviews. Moreover, the support of Akiko Imaizumi with data management is also acknowledged.

Contributors The international research team included Japanese educators (MK, ES), a physician (TM), a qualitative educational scientist (MO), educational scientists from the Netherlands (RS, AJJAS) and Australia (LWTS). MK conducted all interviews. MK, TM, ES and MO analysed and coded the Japanese data from a Japanese cultural perspective. RS, AJJAS and LWTS were involved in the interpretation of the themes and quotes, which were translated into English. Using their non-Asian perspective, they specifically aimed to understand how Japanese culture influenced the sense-making of residents. Discussions continued until a consensus was reached within all of the research members.

Funding This work was supported by JSPS KAKENHI (grant number 17H04097).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study was approved by the Institutional Review Board of Kyushu University.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Makoto Kikukawa <http://orcid.org/0000-0002-7903-6430>

Renée E Stalmeijer <http://orcid.org/0000-0001-8690-5326>

REFERENCES

- Oxford University Press. *Definition of evaluation*. English Oxford Living Dictionaries. (n.d). 2019. <https://en.oxforddictionaries.com/definition/evaluation>.
- Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. *Acad Med* 2000;75:161–6.
- Snell L, Tallett S, Haist S, et al. A review of the evaluation of clinical teaching: new perspectives and challenges. *Med Educ* 2000;34:862–70.
- Epstein RM. Assessment in medical education. *N Engl J Med* 2007;356:387–96.
- Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015;49:560–75.
- Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. *Acad Med* 2016;91:785–95.
- American Educational Research Association APA, National Council in Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, 1999.
- Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;50:741–9.
- Kane M. *Educational measurement*. 4th edn. Westport, CT: Praeger Publishers, 2006: 17–64.
- Schuworth LWT, van der Vleuten CPM. Challenges for educationalists. *BMJ* 2006;333:544–6.
- Strand P, Sjöborg K, Stalmeijer R, et al. Development and psychometric evaluation of the undergraduate clinical education environment measure (UCEEM). *Med Teach* 2013;35:1014–26.
- Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;37:830–7.
- Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–166.e16.
- Fromme HB, Karani R, Downing SM. Direct observation in medical education: a review of the literature and evidence for validity. *Mt Sinai J Med* 2009;76:365–71.
- Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009;302:1316–26.
- Sullivan GM. A primer on the validity of assessment instruments. *J Grad Med Educ* 2011;3:119–20.
- Bult MK, Verschuren O, Gorter JW, et al. Cross-cultural validation and psychometric evaluation of the Dutch language version of the children's assessment of participation and enjoyment (CAPE) in children with and without physical disabilities. *Clin Rehabil* 2010;24:843–53.
- Sperber AD. Translation and validation of study instruments for cross-cultural research. *Gastroenterology* 2004;126:S124–8.
- Berthoz S, Wessa M, Kedia G, et al. Cross-cultural validation of the empathy quotient in a French-speaking sample. *Can J Psychiatry* 2008;53:469–77.
- Sousa VD, Rojjanasirak W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *J Eval Clin Pract* 2011;17:268–74.
- Ndosi M, Tennant A, Bergsten U, et al. Cross-cultural validation of the educational needs assessment tool in RA in 7 European countries. *BMC Musculoskelet Disord* 2011;12:110.
- Wong Ngai-Ying, Lin Wen-Ying, Watkins D, et al. Cross-cultural validation of models of approaches to learning: an application of confirmatory factor analysis. *Educ Psychol* 1996;16:317–27.
- McInerney DM, Ali J. Multidimensional and hierarchical assessment of school motivation: cross-cultural validation. *Educ Psychol* 2006;26:717–34.
- Durning SJ, Artino AR. Situativity theory: a perspective on how participants and the environment can interact: AMEE guide No. 52. *Med Teach* 2011;33:188–99.
- Schuworth LWT, van der Vleuten CPM. General overview of the theories used in assessment: AMEE guide No. 57. *Med Teach* 2011;33:783–97.
- Moss PA, Girard BJ, Haniford LC. Chapter 4: validity in educational assessment. *Rev Educ Res* 2006;30:109–62.
- Patel VL, Yoskowitz NA, Arocha JF. Towards effective evaluation and reform in medical education: a cognitive and learning sciences perspective. *Adv Health Sci Educ Theory Pract* 2009;14:791–812.
- Hickey DT, Zuiker SJ. A new perspective for evaluating innovative science programs. *Sci Educ* 2003;87:539–63.
- Mislevy RJ. How developments in psychology and technology challenge validity argumentation. *J Educ Meas* 2016;53:265–92.
- Mislevy RJ. On integrating psychometrics and learning analytics in complex assessments. In: *Data Anal Psychometrics: inform assess Pract*. 1, 2018.
- Beckman TJ, Ghosh AK, Cook DA, et al. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med* 2004;19:971–7.
- Kikukawa M, Stalmeijer RE, Emura S, et al. An instrument for evaluating clinical teaching in Japan: content validity and cultural sensitivity. *BMC Med Educ* 2014;14:1–8.
- Doja A, Horsley T, Sampson M. Productivity in medical education research: an examination of countries of origin. *BMC Med Educ* 2014;14:243.
- Hodges BD, Maniate JM, Martimianakis MAT, et al. Cracks and crevices: globalization discourse and medical education. *Med Teach* 2009;31:910–7.
- Watling CJ, Lingard L. Grounded theory in medical education research: AMEE guide No. 70. *Med Teach* 2012;34:850–61.
- Charmaz K. *Constructing grounded theory*. SAGE, 2014.
- Bryant A. Re-grounding grounded theory. *J Inform Technol Theory Appl* 2002;4:25.
- Varley HP. *Japanese culture*. University of Hawaii Press, 2000.
- Park IH, Cho L-J. Confucianism and the Korean family. *J Comp Fam Stud* 1995;26:117–34.

- 40 Yum JO. The impact of Confucianism on interpersonal relationships and communication patterns in East Asia. *Commun Monogr* 1988;55:374–88.
- 41 Craig PE, Craig E. *Concise Routledge encyclopedia of philosophy*. Taylor & Francis, 2013.
- 42 Kincaid DL. *Communication theory: Eastern and Western perspectives*. Academic Press, 2013.
- 43 Hofstede G. *Culture's consequences: comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: Sage Publications, Inc, 2001.
- 44 Mulder M, Ritsema van Eck JR, de Jong RD. An organization in crisis and non-crisis situations. *Hum Relat* 1971;24:19–41.
- 45 Lincoln YS, Guba EG, Pilotta JJ. Naturalistic inquiry. *International Journal of Intercultural Relations* 1985;9:438–9.
- 46 Fluit C, Bolhuis S, Grol R, et al. Evaluation and feedback for effective clinical teaching in postgraduate medical education: validation of an assessment instrument incorporating the CanMEDS roles. *Med Teach* 2012;34:893–901.
- 47 Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015;49:560–75.

Supplementary Table 1. Five source of validity evidence

Evidence Source	Definition
Content	The relationship between a test's content and the construct it is intended to measure. Refers to themes, wording, and format of items on an evaluation instrument. Includes analysis by experts of how adequately items measure the content to ensure appropriate content representation.
Internal structure	The degree to which individual items within the instrument fit the underlying constructs. Items measuring a unidirectional construct should be homogeneous, while items measuring complex constructs should not. Most often assessed through measurements of internal consistency reliability and factor analysis.
Response process	Analyses of responses include the actions, strategies, and thought processes of individual responses or observers. Differences in response processes may reveal sources of variance that are irrelevant to the construct being measured. Also includes instrument of security, scoring, and reporting of results.
Relation to other variables	The relationship between scores and other variables relevant to the construct being measured. Relationship may be positive (convergent or predictive) or negative (divergent or discriminating), depending on the construct being measured.
Consequences	Evaluations are intended to have some desired effect (e.g., improve teaching performance), but they also have unintended effects. Evaluating such consequences can support or challenge the validity of score interpretations.

Supplementary Table 2. Kane's four inferences

Inferences	Explanations
Scoring	<u>From observed performance to an observed score</u> Examples: items, observation format, scoring rubric, rater selection and training
Generalization	<u>From observed score to the test-world universe</u> Examples: sampling strategy, sample size
Extrapolation	<u>From test-world universe to the real world</u> Scope of test, authenticity of evaluation context, authenticity of item/scenario
Implications	<u>From the real world to the final decision</u> Examples: Pass/Fail standard, planned action based on assessment results

Appendix 1. Teaching Evaluation Sheet

Teaching Evaluation Sheet					
Hospital:		Terms of rotation:		month,	Course:
Teacher's Name:		Date:		(mm/dd/yy)	
Please assess your teacher. Encircle the appropriate response on the rating scale.					
My clinical teacher:	Strongly disagree	Dis-agree	Neutral	Agree	Strongly agree
1. Shows enthusiasm for teaching.	1	2	3	4	5
2. Is accessible.	1	2	3	4	5
3. Encourages residents to call him or her at any time for any reason.	1	2	3	4	5
4. Displays reasoning processes.	1	2	3	4	5
5. Provides sufficient support.	1	2	3	4	5
6. Stimulates residents' interest in learning and/or subject.	1	2	3	4	5
7. Treats residents with respect.	1	2	3	4	5
8. Actively involves residents in patient care.	1	2	3	4	5
9. Sets clear roles for residents.	1	2	3	4	5
10. Stimulates residents to think critically when solving a problem.	1	2	3	4	5
11. Is a good role model for relationships with medical staff.	1	2	3	4	5
12. Is a good role model for doctor-patient relationships	1	2	3	4	5
13. Is a good clinical supervisor at all times.	1	2	3	4	5
14. Gives residents opportunities to practice.	1	2	3	4	5
15. Does not pretend to know all things.	1	2	3	4	5
16. Demonstrates the importance of safety.	1	2	3	4	5
17. Thinks speculatively about areas of uncertainty with residents.	1	2	3	4	5
18. Explains clearly which aspects are important and why.	1	2	3	4	5
19. Gives concrete indications about what should be improved.	1	2	3	4	5

20. Contributes additional clinical information or advice about diagnosis and management plan for each active problem.	1	2	3	4	5
21. Encourages residents to consider psychosocial problems.	1	2	3	4	5
22. Teaches residents how to conduct clinical research.	1	2	3	4	5
23. Discusses clear training goals with residents during rotations.	1	2	3	4	5
24. Shows social common sense.	1	2	3	4	5
25. I would like to work with this teacher again.	1	2	3	4	5

Kikukawa M, Stalmeijer RE, Emura S, *et al.* An instrument for evaluating clinical teaching in Japan: content validity and cultural sensitivity. *BMC Med Educ* 2014;**14**:1–8.

Appendix 2. Interview guideline

I. Introductory conversation

1. The respondents are thanked for participating in this study.
2. The purpose of this study is stated.
3. Informed consent is obtained from the participants.
4. Permission to record on audiotape is obtained from the participants.

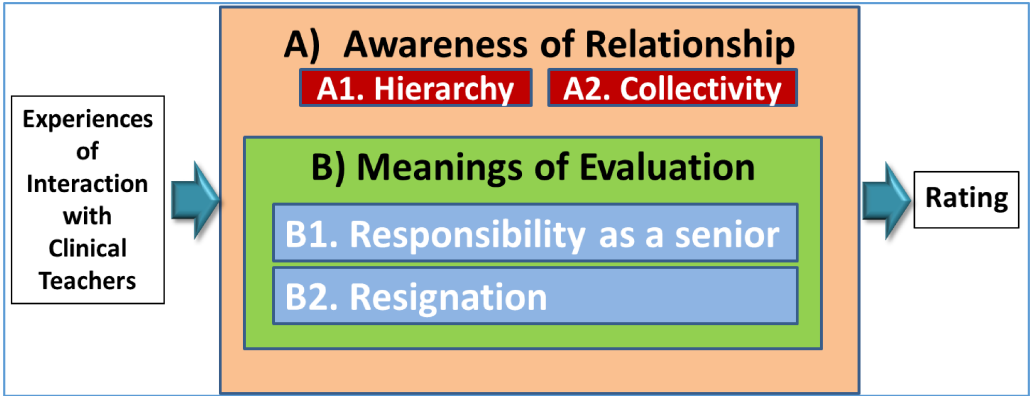
II. Remember when you fill in this form

1. Tell me your process of filling the form in detail.
2. Do you remember specific situations or general impressions?
3. Is your criterion different in each item? If yes, how different is it?
4. Which things influence you when you fill out such a rating scale?
5. How do the factors influence you to assign rates to your clinical teachers?
6. To what degree are you influenced by these factors when filling a form?
7. Which situation do you remember in filling in a form?
8. How do you decide the rating from 1 to 5?
9. This form is anonymous. If it is not anonymous, would you change your rating behavior?
10. Is there anything else that influenced your rating behavior?
Is yes, what is that? And how?

III. Conclusion

Would you like to say anything else about this topic?

Supplementary Figure 1. Residents’ Sense Making



Supplementary Table 1. Five source of validity evidence

Evidence Source	Definition
Content	The relationship between a test's content and the construct it is intended to measure. Refers to themes, wording, and format of items on an evaluation instrument. Includes analysis by experts of how adequately items measure the content to ensure appropriate content representation.
Internal structure	The degree to which individual items within the instrument fit the underlying constructs. Items measuring a unidirectional construct should be homogeneous, while items measuring complex constructs should not. Most often assessed through measurements of internal consistency reliability and factor analysis.
Response process	Analyses of responses include the actions, strategies, and thought processes of individual responses or observers. Differences in response processes may reveal sources of variance that are irrelevant to the construct being measured. Also includes instrument of security, scoring, and reporting of results.
Relation to other variables	The relationship between scores and other variables relevant to the construct being measured. Relationship may be positive (convergent or predictive) or negative (divergent or discriminating), depending on the construct being measured.
Consequences	Evaluations are intended to have some desired effect (e.g., improve teaching performance), but they also have unintended effects. Evaluating such consequences can support or challenge the validity of score interpretations.

Supplementary Table 2. Kane's four inferences

Inferences	Explanations
Scoring	<u>From observed performance to an observed score</u> Examples: items, observation format, scoring rubric, rater selection and training
Generalization	<u>From observed score to the test-world universe</u> Examples: sampling strategy, sample size
Extrapolation	<u>From test-world universe to the real world</u> Scope of test, authenticity of evaluation context, authenticity of item/scenario
Implications	<u>From the real world to the final decision</u> Examples: Pass/Fail standard, planned action based on assessment results

Appendix 1. Teaching Evaluation Sheet

Teaching Evaluation Sheet					
Hospital:		Terms of rotation:		month,	Course:
Teacher's Name:		Date:		(mm/dd/yy)	
Please assess your teacher. Encircle the appropriate response on the rating scale.					
My clinical teacher:	Strongly disagree	Dis-agree	Neutral	Agree	Strongly agree
1. Shows enthusiasm for teaching.	1	2	3	4	5
2. Is accessible.	1	2	3	4	5
3. Encourages residents to call him or her at any time for any reason.	1	2	3	4	5
4. Displays reasoning processes.	1	2	3	4	5
5. Provides sufficient support.	1	2	3	4	5
6. Stimulates residents' interest in learning and/or subject.	1	2	3	4	5
7. Treats residents with respect.	1	2	3	4	5
8. Actively involves residents in patient care.	1	2	3	4	5
9. Sets clear roles for residents.	1	2	3	4	5
10. Stimulates residents to think critically when solving a problem.	1	2	3	4	5
11. Is a good role model for relationships with medical staff.	1	2	3	4	5
12. Is a good role model for doctor-patient relationships	1	2	3	4	5
13. Is a good clinical supervisor at all times.	1	2	3	4	5
14. Gives residents opportunities to practice.	1	2	3	4	5
15. Does not pretend to know all things.	1	2	3	4	5
16. Demonstrates the importance of safety.	1	2	3	4	5
17. Thinks speculatively about areas of uncertainty with residents.	1	2	3	4	5
18. Explains clearly which aspects are important and why.	1	2	3	4	5
19. Gives concrete indications about what should be improved.	1	2	3	4	5

20. Contributes additional clinical information or advice about diagnosis and management plan for each active problem.	1	2	3	4	5
21. Encourages residents to consider psychosocial problems.	1	2	3	4	5
22. Teaches residents how to conduct clinical research.	1	2	3	4	5
23. Discusses clear training goals with residents during rotations.	1	2	3	4	5
24. Shows social common sense.	1	2	3	4	5
25. I would like to work with this teacher again.	1	2	3	4	5

Kikukawa M, Stalmeijer RE, Emura S, *et al.* An instrument for evaluating clinical teaching in Japan: content validity and cultural sensitivity. *BMC Med Educ* 2014;**14**:1–8.

Appendix 2. Interview guideline

I. Introductory conversation

1. The respondents are thanked for participating in this study.
2. The purpose of this study is stated.
3. Informed consent is obtained from the participants.
4. Permission to record on audiotape is obtained from the participants.

II. Remember when you fill in this form

1. Tell me your process of filling the form in detail.
2. Do you remember specific situations or general impressions?
3. Is your criterion different in each item? If yes, how different is it?
4. Which things influence you when you fill out such a rating scale?
5. How do the factors influence you to assign rates to your clinical teachers?
6. To what degree are you influenced by these factors when filling a form?
7. Which situation do you remember in filling in a form?
8. How do you decide the rating from 1 to 5?
9. This form is anonymous. If it is not anonymous, would you change your rating behavior?
10. Is there anything else that influenced your rating behavior?
Is yes, what is that? And how?

III. Conclusion

Would you like to say anything else about this topic?

Supplementary Figure 1. Residents’ Sense Making

