BMJ Open

Assessing doctors' competencies using multisource feedback: validating a Japanese version of the Sheffield Peer Review Assessment Tool (SPRAT)

Hatoko Sasaki,[1,4] Julian Archer,[2] Naohiro Yonemoto,[3] Rintaro Mori,[4] Toshihiko Nishida,[5] Satoshi Kusuda,[5] Takeo Nakayama[1]

To cite: Sasaki H, Archer J, Yonemoto N, *et al*. Assessing doctors' competencies using multisource feedback: validating a Japanese version of the Sheffield Peer Review Assessment Tool (SPRAT). *BMJ Open* 2015;**5**:e007135. doi:10.1136/bmjopen-2014-007135

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2014-007135).

Received 7 November 2014
Revised 30 April 2015
Accepted 7 May 2015

For numbered affiliations see end of article.

Correspondence to
Hatoko Sasaki;
hatokos@hotmail.com

ABSTRACT

Objective: To assess the validity and reliability of the Sheffield Peer Review Assessment Tool (SPRAT) Japanese version for evaluating doctors' competencies using multisource feedback.

Methods: SPRAT, originally developed in the UK, was translated and validated in three phases: (1) an existing Japanese version of SPRAT was back-translated into English; (2) two expert panel meetings were held to develop and assure content validity in a Japanese setting; (3) the newly devised Japanese SPRAT instrument was tested by a multisource feedback survey, validity was tested using principal component factor analysis, and reliability was assessed using generalisability and decision studies based on generalisability theory.

Results: 86 doctors who had been practising for between 2 and 33 years participated as assessees and were evaluated with the SPRAT tool. First, the doctors identified 1019 potential assessors who were each sent SPRAT forms (response rate, 81%). The mean number of assessors per doctor was 9.7 (SD=2.5). The decision study showed that 95% CIs of ±0.5 were achieved with only 5 assessors. 85 of the 86 doctors achieved scores that could be placed with 95% CI above the 4 expected standard. Doctors received lower scores from more senior assessors ($p<0.001$) and higher scores from those they had known longer ($p<0.001$). Scores also varied with the job role ($p<0.05$).

Conclusions: Following translation and content validation, the Japanese instrument behaved similarly to the UK tool. Assessor selection remains a primary concern, as the assessment scores are affected by the seniority of the assessor, the length of the assessor–assessee working relationship, and the assessor's job role. Users of the SPRAT tool need to be aware of these limitations when administering the instrument.

Strengths and limitations of this study

- Established methods were used to translate and assess the scale's content validity.
- The findings show that the Japanese version of Sheffield Peer Review Assessment Tool (SPRAT) behaved similarly to the original English version.
- The Japanese SPRAT can be used to assess and provide feedback on the performance of Japanese doctors, and to compare doctors' performance with that of peers in Japan and the UK.
- The assessor's characteristics can affect overall scores.
- Further research is needed to investigate the generalisability of the results beyond paediatricians.

INTRODUCTION

Evaluation of physicians' interpersonal and communication skills, professionalism and teamwork behaviours is a critical and universal issue for the development of professional human resources in healthcare. Workplace-based peer assessment is widely used and is known to be a reliable technique in order to provide feedback and guide performance.[1 2] Multisource feedback (MSF) or 360° evaluation is a survey-based method in which assessees are evaluated by supervisors, peers (coworkers) and patients. MSF has been adopted by licensing authorities[3] and healthcare facilities[1 4] to assess a broad range of physician competencies, including performance, teamwork behaviours, teaching, interpersonal and communication skills.[2 5] Even though individual factors, context of feedback and administration of the survey have a fundamental effect on assessees' responses, MSF can lead to performance improvement.[6] A recent systematic review[7] has shown that MSF, if implemented correctly, can have a positive effect on performance.

The Sheffield Peer Review Assessment Tool (SPRAT) was originally developed to assess the competencies of paediatricians based on good medical practice (GMP)[8] in the UK. SPRAT informs the quality assurance process when assessing doctors' work-based performance. The tool encompasses five domains of GMP:

good clinical care; maintaining GMP; teaching and training, assessing and appraising; relationships with patients and working with colleagues. SPRAT consists of 24 questions with a six-point scale ranging from 'very poor' to 'very good' and includes the option to select 'unable to comment'. A space for 'strengths' and 'suggestions for development' is also provided.

A tool modelled on SPRAT was introduced in Japan to assess doctors' clinical skills. However, validity and reliability assessments of the tool for Japanese subjects were not performed prior to its introduction. We believe it is important to take cultural adaptivity into account when any established instrument is introduced into a different culture. In this study, we went beyond a simple translation and examined the validity (including reliability) evidence of the Japanese version of SPRAT as part of the Improvement of NICU Practice and Team-Approach Cluster randomised controlled trial (INTACT).[9] Translation and validation were conducted in three phases. In the first phase, we conducted back-translation of the existing Japanese SPRAT tool into English. In the second phase, a panel of experts met to assess the content validity of the instrument. In the third phase, we performed pilot testing of the MSF survey for Japanese patients, and tested the validity and reliability of the Japanese version using psychometric methods. This paper mainly focuses on the statistical results of the pilot testing.

METHODS

Ethics approval

This study did not involve patients, and therefore written consent was not required. Author HS and collaborators of the participating hospitals gave all participants an explanation of the pilot study and an instruction sheet of MSF. Participating in the study was voluntary and consent was obtained orally or by email. Anonymity and confidentiality of the data were assured to all participants.

Translation and back-translation

Permission to use an existing SPRAT Japanese translation was obtained from the translator. In order to assess the quality of the translation, back-translation into English was performed by a professional translator. This translation was then compared with the original tool by its author (JA).

Expert panel

We recruited an expert panel of 18 members including medical educators, neonatologists, paediatricians, internists, paediatric nurse specialists, other health professionals and family patient representatives to assess the content validity of the Japanese translation. We searched for suitable panellists using two of the largest paediatric mailing lists in Japan: the Japan Pediatric Mailing List Conference (https://jpmlc.org/index.php?mod=Jpmlc&act=GuestIndex) and Nicu-Forum.Net (http://www.nicu-forum.net/). The original author, JA, was also invited to join the panel. Two panel meetings were held: one facilitated by JA in English and the other held in Japanese in order to maximise opportunities to gather a wide range of experts from Japan. The panel first assessed the relevance of Japanese expression and then compared SPRAT questions with established performance criteria[10 11] in Japan for paediatricians and board-certified perinatal medicine physicians. A mapping sheet was used to examine whether SPRAT-response items covered the established criteria. Finally, demographic data to be collected as part of the study were added to the tool and the scale was validated.

Pilot testing of the instrument: MSF survey

We conducted a pilot test of the MSF survey from October to December 2012 using the newly developed tool to investigate its validity and reliability.

Study population

Four neonatal intensive care units (NICUs) located in different areas of Japan that were involved in INTACT, and one department of paediatrics that was not involved in INTACT, participated in the pilot study. All doctors working at the units and the department were recruited as study participants.

Questionnaire distribution

Each consenting doctor or 'assessee' was asked to select at least 10 assessors from his/her supervisors, peers, junior residents, nurses and other health professionals with whom they worked closely. The target number of assessors was between 8 and 12 in order to achieve reasonable levels of reliability.[1]

Data analysis

Data were anonymised and responses of 'unable to comment' were removed prior to analysis. We did not replace the missing values. All statistical analyses were undertaken in SPSS V.21.0 (IBM Corporation, USA). Feasibility was evaluated using response rates and response time. The mean score per SPRAT form was used for all analyses. Scores of self-assessment were excluded for all analyses.

Item analysis

We calculated mean ratings of individual and overall items and the percentage of missing values.

Factor analysis

We conducted a principal component factor analysis with an extraction criterion of Eigenvalue >1 by a scree plot and with varimax rotation, using the Kaiser-Meyer-Olkin (KMO) and Bartlett tests to explore the validity of SPRAT in line with previous studies.[12] The KMO and Bartlett tests measured the strength of the relationship among variables. Field[13] recommends that KMO values greater than 0.7 are acceptable. We used the guideline for identifying significant factor loading based on

sample size.[14] The cut-off value of this study was set at 0.3, as per the guideline. If a variable had several high-factor loadings, we selected the larger size of the factor loading to interpret the factor matrix as having importance in a practical sense. This is because the majority of factor solutions do not lead to a simple structure solution (a single high loading for each variable on only one factor).[14] We also performed congruence analysis to calculate a congruence coefficient using the free software, Orthosim 2.1. The congruence coefficient is an indicator of the similarity between the factor loadings for the Japanese sample and that for the UK sample. The coefficient varies between 0 and 1 with–absolute identity.

Demographic data analysis: assessee

Frequency, mean and SD were calculated for gender, length of clinical experience, board certification, specialty and seniority. Length of clinical experience was divided into two categories: ≥5 years and <5 years. This cut-off was determined because a minimum of 5 years' training is required for medical graduates to be eligible for board certification as paediatricians in Japan.

Demographic data analysis: assessor

The job roles or job descriptions of assessors were classified into six groups: consultant (eg, director, professor, head physician, associate professor), specialist (eg, house/medical staff, fellow, lecturer, assistant professor), resident (eg, junior residents with 1–2 years of experience in paediatric residency training, senior residents with 3–5 years of experience), managerial nurse, nurse and other. We calculated mean scores for each job role. Demographic data on assessors were analysed using hierarchical regression to calculate potential influences on assessees' ratings. This was undertaken with controls for the seniority of assessees (≥5 years and <5 years), as it was accepted that performance would be affected by training. Other characteristics included assessors' gender, occupation, length of working relationship with assessees, educational background and year of graduation. P values ($p<0.01$) were reported as a measure of the relative importance of each potential confounder.

Reliability

Reliability can be assessed in several ways including internal consistency with Cronbach's α coefficients and test-retest reliability, considered as classical test theory. Generalisability theory[14] is more suitable for this study than classical test theory by means of focusing on improving assessment and providing models and methods that allow a multifaceted perspective on measurement error and its components. Generalisability theory comprises two studies: a generalisability study (G study) and a decision study (D study). A G study estimates variance components of the facets (assessee and assessor). The D study investigates the degree of reliability of assessment using a generalisability coefficient by estimating variance components. A generalisability coefficient is similar to an intraclass correlation. This analysis gives an investigator the estimated number of assessors required to obtain a reliable assessment per assessee. Assessors are nested with assessed doctors in this study. Each doctor was rated by unequal numbers of assessors. Variance components were calculated using VARCOMP (Minimum Norm Quadratic Unbiased Estimation—the MINQUE procedure) in SPSS using SPSS syntax.[15] The estimated variance components for both assessees and the interaction of assessees and assessors (error) were extracted to generate a generalisability coefficient (Ep2)=a ratio of the estimated variance components for assessees over the sum of the estimated variance components for assesses, plus the interaction of assessees and assessors (error).[16] Mushquash and O'Connor[17] provide a more in-depth discussion about generalisability theory analysis.

We attained a measure of precision by producing the 95% CI around each mean rating as described below. We used the square root of the measurement error as the SE of measurement (SEM), and determined the SEM for 2–13 assessors ($\sqrt{}$error/number of assessors). The 95% CIs were equal to the SEM multiplied by 1.96, and were added to and subtracted from a mean rating.[12 18] If the 95% CI around this score was still above or below the cut-off score, then we can be 95% certain that they have indeed 'passed' or 'failed'.

Free-text comments

We analysed free-text comments using EKWords V.2.0.1 (DJ Soft Co, Ltd), a type of free software for qualitative text analysis of the Japanese language. Frequent words were counted first, and then synonyms and related terms for the top three frequent words were extracted to generate themes of keywords.

RESULTS

Back-translation and expert panels

No major difference was observed between the back-translation and the original English instrument. Although the expert panel had some questions that they did not map directly to any of the documents, the panel considered that all items of the Japanese tool were relevant, and therefore no items were removed and no new items were developed. However, the panel members agreed that some items needed to be rephrased and reworded to be faithful to the original text as well as to incorporate more natural phrasing in Japanese. For example, two similar terms were used for 'ability' in the Japanese translation, so for consistency we ensured that only one single term was used throughout. Also, the panel decided that the term 'self-improvement' was more suitable than the term 'learning' in the context of the Japan Pediatric Association training handbook, which encourages paediatricians to actively improve and develop their professional skills throughout their working life. Panellists generated footnotes for five items of the tool to help assessors better understand the items, and discussed the validity of the scale. The panel

decided that the required demographic data to be collected from assessees would include gender, job role, years of practice, board certification and specialty. Demographic data for assessors included gender, occupation, job role, specialty, length of working relationship with assessees, educational background and year of graduation. In the existing Japanese translation, no descriptors for each point of the scale were included. Since descriptors can help assessors to understand the meaning of point scales, they were added to each point scale. After two panel meetings, the panel came to a consensus and the Japanese version was finalised (see online supplementary appendix 1).

Pilot testing of the instrument

The characteristics of assessed doctors and assessors are shown in table 1. Eighty-six assessees (years of practice: mean=9.0, SD=8.0) identified 1019 potential assessors who were each distributed SPRAT forms. Out of these, 826 completed forms (years of practice: mean=9.7, SD=7.9) were returned (response rate, 81%). The mean number of assessors per assessee was 9.7 (range 2 to 13). Seventy-three (84.8%) assessees received their feedback from more than eight assessors. The mean time required for each assessor to complete the form was 6 min (range 0.5–30 min).

Item analysis

The mean ratings of the individual items ranged from 4.67 (SD=1.02) to 5.13 (SD=0.89). The lowest rating was given for 'Leadership skills' and the highest rating was given for 'Accessibility/reliability'. Among 86 assessees, 85 (99%) scored an overall mean of 4 or more. The percentage of missing values among the 25 items ranged from 0.5% to 7%.

Table 1 Characteristics of assessed doctors and assessors

	Assessed doctors (N=86) n (%)	Assessors (N=826) n (%)
Gender		
Male	57 (66.3)	408 (49.5)
Female	29 (33.7)	417 (50.5)
Year of practice		
5 years and above	56 (65.1)	511 (62.0)
Less than 5 years	26 (30.2)	284 (34.0)
Unknown	4 (4.7)	31 (4.0)
Board-certified specialist		
Yes	38 (44.2)	–
No	31 (36.0)	–
Unknown	17 (19.8)	–
Specialty		
General	45 (52.0)	–
Paediatrics		
Neonatology	41 (48.0)	–
Job role		
Consultant	–	104 (12.9)
Specialist	–	269 (33.3)
Resident	–	247 (30.6)
Managerial nurse	–	44 (5.4)
Nurse	–	142 (17.6)
Other	–	2 (0.2)

Factor analysis

The whole instrument was found to be suitable for factor analysis (KMO=0.96, p<0.001). The principal components factor analysis returned a two-factor solution accounting for 69% of the variance (table 2). One factor is related to questions about aspects of clinical care in medical practice, and the other is related to psychosocial skills. There was no factor loading lower than 0.3, while several items were coloaded on both factor components. The overall solution congruence was 0.99. The similarity of factor loadings between the Japanese sample and the UK sample is proved.

Demographic data analysis: assessees

The overall mean score achieved by assessees on SPRAT was 4.87 (SD=0.43; figure 1). No difference in ratings was observed between gender (male n=57, mean=4.89, SD=0.47, female n=29, mean=4.82, SD=0.34, p=0.382). The length of clinical experience did not affect scores (≥5 years n=53, mean=4.93, SD=0.37 and <5 years n=28, mean=4.79, SD=0.50, p=0.154). Board-certified specialists did not score differently from non-holders (holders n=38, mean=4.96, SD=0.37, non-holders n=31, mean=4.81, SD=0.44, p=0.142). No difference was observed by specialty (general paediatrics n=45, mean=4.85, SD=0.48, neonatology n=41, mean=4.89, SD=0.37, p=0.626). However, physicians (clinical experience ≥5 years) scored significantly higher than residents (clinical experience <5 years; physicians n=48, mean=4.97, SD=0.37, residents n=38, mean=4.73, SD=0.46, p=0.009).

Demographic data analysis: assessor

The mean ratings for each assessor's job role are shown in figure 2. Both consultants (eg, director, professor, head physician, associate professor) and specialists (eg, house/medical staff, fellow, lecturer, assistant professor) rated significantly lower than residents (consultants n=104, mean=4.88, SD=0.68, resident n=247, mean=5.05, SD=0.56, p=0.03; specialists n=269, mean=4.90, SD=0.69, p=0.007, respectively). No difference was observed between consultants and specialists. Managerial nurses were assigned significantly lower scores than nurses (managerial nurses n=44, mean=4.37, SD=0.52, nurses n=142, mean=4.89, SD=0.72, p<0.001). Assessment scores were also affected by the seniority of assessors (year of graduation; p<0.001) and length of working relationships (p<0.001).

Reliability

Little difference was observed between the reliability coefficients for all assessees, that is, the two categories of clinical experience (≥5 years and <5 years) or clinical care and psychosocial skills (figure 3). Figure 4 shows that 74 of the 86 assessees scored an overall mean of 4.5

Table 2 Principal components factor analysis

	Japanese version of SPRAT questions	Component 1	Component 2
1	Ability to diagnose patient problems	0.806	0.349
2	Ability to formulate appropriate management plans	0.826	0.319
3	Ability to manage complex patients	0.766	0.360
4	Awareness of their own limitations	0.609	0.434
5	Ability to respond to psychosocial aspects of illness	0.375	0.720
6	Appropriate utilisation of resources, for example, ordering investigations	0.610	0.419
7	Ability to assess risks and benefits when treating patients	0.793	0.345
8	Ability to coordinate patient care	0.730	0.442
9	Technical skills (appropriate to current practice)	0.784	0.213
10	Ability to apply up-to-date/evidence-based medicine	0.827	0.220
11	Ability to manage time effectively/prioritise	0.763	0.265
12	Ability to deal with stress	0.462	0.351
13	Commitment to learning	0.654	0.372
14	Willingness and effectiveness when teaching/training colleagues	0.703	0.402
15	Ability to give feedback (private, honest and supportive)	0.613	0.538
16	Communication with patients	0.276	0.866
17	Communication with carers and/or family	0.263	0.879
18	Respect for patients and their right to confidentiality	0.279	0.841
19	Verbal communication with colleagues	0.327	0.783
20	Written communication with colleagues	0.440	0.683
21	Ability to recognise and value the contribution of others	0.397	0.769
22	Accessibility/reliability	0.491	0.645
23	Leadership skills	0.763	0.374
24	Management skills	0.765	0.358

SPRAT, Sheffield Peer Review Assessment Tool.

or more. When investigating the 95% confidence levels around the mean score, we observed 95% CIs of ±0.5 when the number of assessors was 5. Of the 86 assessees, only 5 assessors would then be required to obtain a reliable score. However, little difference was observed between the two categories of clinical experience. For participants with ≥5 years of clinical experience, 95% CIs of ±0.5 can be achieved with six assessors while those

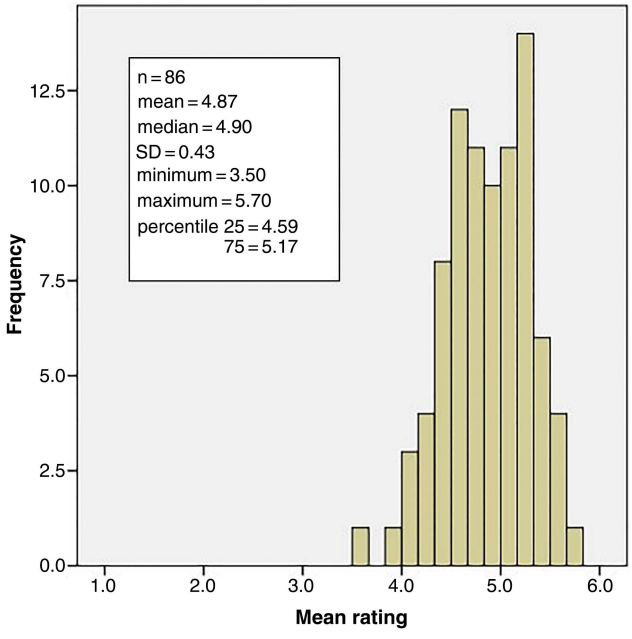

Figure 1 Distribution of aggregate scores for assessees. Histogram with a normal distribution curve shows distribution of aggregate means for assessees. Except for one assessee, all aggregate scores were above 4.0 if they met the expected standard.

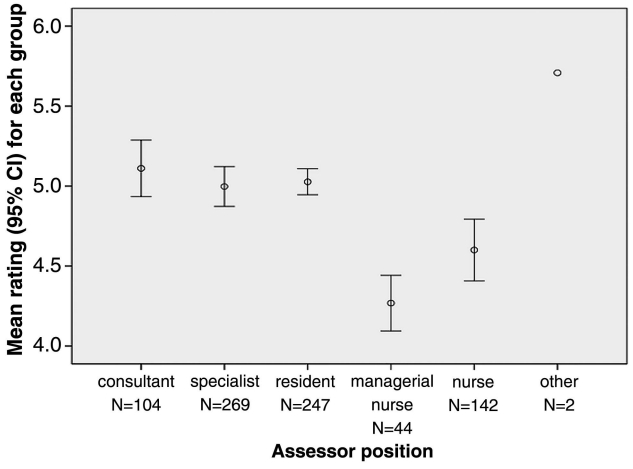

Figure 2 Mean and 95% CI for assessors in position groups. Error plot shows mean and 95% CI for assessors in position groups. Others (researcher and midwife) rated the highest mean (mean=5.50, SD=0.29). The managerial nurse rated the lowest mean (mean=4.37, SD=0.52). Both consultants (eg, director, professor, head physician, associate professor) and specialists (eg, house/medical staff, fellow, lecturer, assistant professor) rated significantly lower (consultants' mean=4.88, SD=0.68, p=0.03; specialists' mean=4.90, SD=0.69, p=0.007) compared with residents (mean=5.05, SD=0.56).

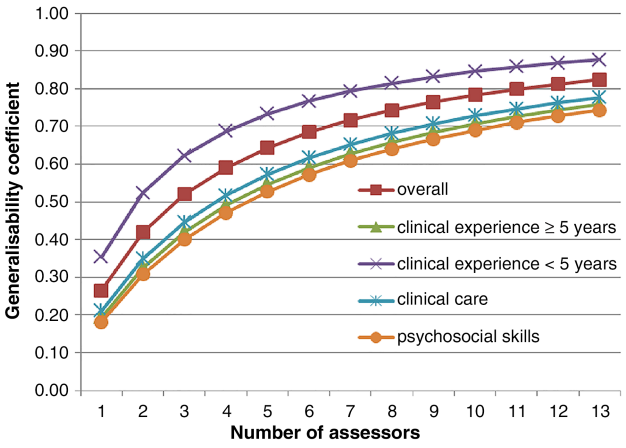

Figure 3 Predicted reliability of ratings. Decision studies showing how sampling affects the predicted reliability of ratings in the cohort as a whole, for each clinical experience group and for each factor identified. Red represents the overall cohort; green represents the cohort of clinical experience ≥5 years; purple represents the cohort of clinical experience <5 years; blue represents the component of clinical care, and orange represents the component of psychosocial skills. The greater generalisability coefficient indicates greater reliability.

with <5 years of clinical experience can achieve 95% CIs of ±0.5 with only four assessors. If 4 is the expected score in the Japanese sample, 99% of assessees scored an overall mean of 4 or more and only one doctor had an overall mean of 4 or below.

Free-text comments

We summarised free-text comments into seven themes: in areas of strength, themes included good communication with patients/their family/medical staff, sympathy with patients, and accessibility; in areas of weakness, themes were lack of respect for others, lack of self-healthcare management, lack of leadership and communication and lack of work efficiency.

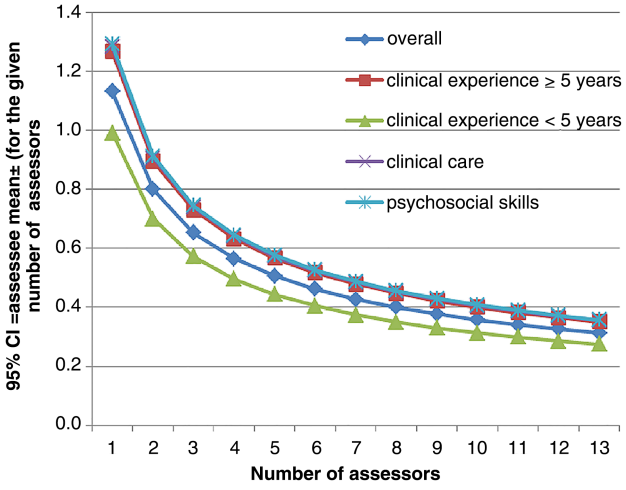

Figure 4 95% CI generated from the SE of measure. The decision study shows 95% CI generated from the SE of measure by different numbers of assessors. Blue represents the overall cohort; red represents the cohort of clinical experience ≥5 years; green represents the cohort of clinical experience <5 years; purple represents the component of clinical care, and aqua blue represents the component of psychosocial skills.

DISCUSSION

Main findings

We have developed and validated the Japanese version of SPRAT for assessing doctors' competencies using 360° evaluation. Our findings show that the Japanese version of SPRAT behaved similarly to the original English version. In this study, reliability of the present version was assessed using the generalisability theory. We found that senior doctors required more assessors than junior doctors to obtain a reliable assessment: a 95% CI with four assessors was ±0.5 for junior doctors, whereas a 95% CI with six assessors was ±0.5 for senior doctors. The two-factor solution was obtained from the Japanese sample, which was similar to the original UK sample (the congruence coefficient=0.99). Nurses assigned doctors lower scores, and in particular the mean score of managerial nurses was significantly lower than any other job roles, which is similar to previous studies.[19] Assessees received lower scores from more senior assessors, which was similar to findings by Davies *et al*[5] where consultants scored trainees lower using the histopathology MSF tool, PATH-SPRAT. However, assessees received higher scores from those they had known longer, which was consistent with the UK studies using SPRAT,[12 20] and implies that scores may be affected by familiarity between the assessor and assessee.[2] Mean response time was 6 min, which is consistent with previous studies.[20]

Explanation and interpretation

The lowest and highest rated items were consistent with results from the UK sample. This implies that basic physician competencies are common across cultures and countries. Although the factor analysis returned two components with a high value of KMO and a high-congruence coefficient, most factor solutions did not result in a simple factor solution (a single high loading for each item on only one factor). This may be because questions that considered clinical care components in medical practice focused on general clinical skills rather than specialty techniques, and therefore they may overlap or closely correlate with questions on psychosocial skills. There is scope in the scale to consider modifying items. However, SPRAT does not report the subscale score but the mean score per form. The intended purpose of the factor analysis is to better understand the internal structure of the scale, instead of justification for reporting subscale scores that correspond to two factors.

In this study, nurses assigned assessees low scores and managerial nurses rated assessees significantly lower than any other job roles, which is in contrast to previous UK studies using SPRAT[19 21] and PATH-SPRAT[5] where

consultants rather than managerial nurses rated assessees significantly lower. This disparity might be explained by cultural differences. A multicentre, cross-sectional study of professionalism using 360° assessments for Japanese residents showed that the mean score of nurses was the lowest among evaluator subgroups.[22] Japanese nurses may have high expectations of doctors' clinical and psychosocial skills.

Seniority of assessors and the length of working relationships also contributed to the variability of the mean score. Assessees received lower scores from more senior assessors. As highlighted by Archer *et al*,[12] assessors' self-confidence in their own skills and experience may change their ability to accurately rate assessees, and this ability may help distinguish evaluative categories. In other words, it might be difficult for junior doctors to assess peers, especially seniors, as junior doctors have less self-confidence in their own skills and experience. The fact that senior doctors generally spend more time in administration and less time in practice might also explain why senior doctors may need more assessors than junior doctors.

Length of the assessor–assessee working relationship was also a confounding factor, which was consistent with previous studies.[12] Assessors seem to more positively evaluate physicians with whom they have worked longer compared with those with shorter working relationships. A broad range of experience established through working with an individual may support the assessor's confidence of their evaluation rather than just personal attachment or familiarity.

Limitations

As SPRAT was originally developed for paediatricians, our sample was drawn from paediatric medicine; however, the sample mainly included the single specialty of NICU. Although items in SPRAT cover the fundamental competencies of doctors rather than special clinical skills, the psychometric properties of the assessment may behave differently in other specialties.

Our findings support the reliability and validity of the MSF instrument for doctors in Japan; however, several factors may affect the scores, including seniority of the assessor, length of the assessor–assessee working relationship and assessor's job role. SPRAT was originally designed to assess the competencies of paediatricians based on GMP, which provides national standards of practice for doctors in the UK. Postgraduate training has been standardised to meet GMP requirements and MSF is also undertaken based on GMP. However, in Japan, there is no such national standard that assessors can refer to, and therefore peer assessment tends to rely on the subjective opinion of the assessors.

Although assessees were asked to select at least 10 assessors with 2 from each job role category, the number of assessors selected actually ranged from 2 to 13. A balanced sample of assessors should be sought when conducting MSF. Inviting a third party to select assessors may be one solution to reduce this bias, although this may not be without its own challenges.[12 20 23 24]

Implications

SPRAT is a tool like other 360° assessments in which assessor characteristics have been shown to have an impact on scores.[12 20 21 23 24] Researchers and investigators using this instrument in the Japanese context should be aware of its potential limitations. Further investigation of the reliability and validity of the instrument in different specialties and in a large sample is warranted in order to assess Japanese physicians in general. Peer assessment for hospital-based physicians has not been conducted systematically in Japan, although some hospitals, especially university-based hospitals, have advanced systems for assessing physicians' competencies to improve educational and professional development. Others are faced with an 'organisational culture' in which doctors feel uncomfortable assessing each other. Even consultants feel inadequate in assessing younger doctors. This unfamiliarity or resistance to peer assessment is another challenge to conducting the survey and may be a cultural difference as compared with those European and North American countries where MSF tools are being widely used. It is important for trainers, administrators and researchers to first make clear the purpose of peer assessment. It may be necessary to emphasise that feedback will not impact their employment but is undertaken to support professional development and to help establish developmental plans with consultants or trainers.

The Japanese version of SPRAT is a much-needed validated instrument that can be used to assess and provide feedback on the performance of Japanese doctors, and to compare doctor performance with peers in Japan and the UK. At the same time, the standing question of international validity and whether the validity of instruments differs by culture remains. Further research is needed to explore this challenge. Free-text comments can also provide valuable information for assessees to understand the overall meaning of their assessment results, rather than simply receiving a numerical score.

CONCLUSIONS

This is the first validation study of SPRAT to be conducted in a country where the official language is not English. The Japanese version demonstrates similar content validity and reliability with the UK sample. However, the instrument is limited by assessor selection, in which assessor seniority, length of the assessor–assessee working relationship and assessor job role can affect overall scores, and lead to the same assessee receiving higher or lower scores depending on the assessor's characteristics. As well as being a valuable professional development tool for doctors in Japan, the Japanese SPRAT may also be a useful instrument in future research into peer assessment practices. However, actual administration of the tool will require a careful consideration of assessor selection.

Author affiliations
[1]Department of Health Informatics, School of Public Health, Kyoto University, Kyoto, Japan
[2]The Collaboration for the Advancement of Medical Education Research & Assessment (CAMERA), Plymouth University Peninsula Schools of Medicine & Dentistry, Plymouth University, Plymouth, UK
[3]Department of Neuropsychopharmacology, National Center of Neurology and Psychiatry, Kodaira, Japan
[4]Department of Health Policy, National Center for Child Health and Development, Tokyo, Japan
[5]Department of Neonatology, Tokyo Women's Medical University, Maternal and Perinatal Center, Tokyo, Japan

Acknowledgements The authors would like to thank Dr Hajime Higashi (Amagasaki Medical Care Co-op Hospital) for permission to use his translation of SPRAT. They also thank Dr Akira Ishiguro (National Center for Child Health and Development), Dr Atsushi Uchiyama (Tokyo Women's Medical University), Dr Yushi Ito (National Center for Child Health and Development), Dr Shinichi Watabe (Kurashiki Central Hospital) and Dr Shigeharu Hosono (Nihon University Itabashi Hospital, Division of Neonatology) for data acquisition, expert panels for their contribution on validating contents of the tool, and all physicians, nurses and other health professionals who generously participated in this study. They also thank Ms Emma Barber (National Center for Child Health and Development) for her editorial support.

Contributors HS performed statistical analysis, interpreted results and drafted the manuscript. JA contributed to the methodology of the study, interpretation of the data and editing of the manuscript. NY provided supervision of data analysis and interpretation. TNi assisted with the recruitment of experts for the panel, and participated in the expert panel. RM assisted with the recruitment of experts for the panel, participated in the expert panel and provided intellectual contribution to the study. SK participated in the expert panel and provided intellectual contribution to the study. TNa critically revised the manuscript for important intellectual content. All authors were engaged in critical commentary and approved the final version of the manuscript.

Funding Health and Labour Sciences Research Grants in FY2012 (H23-Iryo・Shitei-008) were funded by the Ministry of Health, Labour and Welfare, Japan. The funder had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests None declared.

Patient consent Not obtained.

Ethics approval Ethical approval was obtained on 18 October 2012 from the independent review board of INTACT (UMIN000007064) which has its administrative office in Tokyo Women's Medical University.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Ramsey PG, Wenrich MD, Carline JD, *et al*. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269:1655–60.
2. Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof* 2003;23:4–12.
3. Wenghofer EF, Way D, Moxam RS, *et al*. Effectiveness of an enhanced peer assessment program: introducing education into regulatory assessment. *J Contin Educ Health Prof* 2006;26:199–208.
4. Ramsey PG, Carline JD, Blank LL, *et al*. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. *Acad Med* 1996;71:364–70.
5. Davies H, Archer J, Bateman A, *et al*. Specialty-specific multi-source feedback: assuring validity, informing training. *Med Educ* 2008;42:1014–20.
6. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ* 2010;341:c5064.
7. Saedon H, Salleh S, Balakrishnan A, *et al*. The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. *BMC Med Educ* 2012;12:25.
8. General Medical Council. *Good medical practice*. London: GMC, 2001.
9. Nishida T, Morib R, Toyoshimac K,*et al*. Collaborative quality improvement of clinical practice for very low birth weight infants in Japan [INTACT]—study protocol. 2013. http://www.evidencelive.org/posters/2013/collaborative-quality-improvement-of-clinical-practice-for-very-low-birth-weight-infant (accessed 16 Jul 2014).
10. Specialist in Perinatal Medicine. Secondary specialist in perinatal medicine. 2010. http://www.jspnm.com/topics/data/topics110113.pdf
11. Attainable Goals of Pediatricians. Secondary attainable goals of pediatricians. 2010. http://www.jpeds.or.jp/uploads/files/mokuhyo_5.pdf
12. Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in a national programme. *Arch Dis Child* 2010;95:330–5.
13. Field A. Discovering statistics using SPSS. In: Wright DB, ed. *ISM introducing statistical methods*. London: SAGE Publications, 2005:647–59.
14. Hair JF, Anderson RE, Tatham RL, *et al*. *Black (1998), multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall, 1998.
15. Putka DJ, McCloy RA. Estimating Variance Components in SPSS and SAS: An Annotated Reference Guide. 2008.
16. Brennan RL. *Coefficients and indices in generalizability theory. Center for Advanced Studies in Measurement and Assessment, CASMA Research Report* 2003;1:1–44.
17. Mushquash C, O'Connor BP. SPSS and SAS programs for generalizability theory analyses. *Behav Res Methods* 2006;38:542–7.
18. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ* 2005;331:903.
19. Wenrich MD, Carline JD, Giles LM, *et al*. Ratings of the performances of practicing internists by hospital-based registered nurses. *Acad Med* 1993;68:680–7. (1040–2446 (Print)).
20. Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005;330:1251–3.
21. Archer J, Norcini J, Southgate L, *et al*. Mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? *Adv Health Sci Educ Theory Pract* 2008;13:181–92.
22. Tsugawa Y, Ohbu S, Cruess R, *et al*. Introducing the Professionalism Mini-Evaluation Exercise (P-MEX) in Japan: results from a multicenter, cross-sectional study. *Acad Med* 2011;86:1026–31.
23. Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ* 2011;45:886–93.
24. Brinkman WB, Geraghty SR, Lanphear BP, *et al*. Evaluation of resident communication skills and professionalism: a matter of perspective? *Pediatrics* 2006;118:1371–9.