



OPEN ACCESS

EDITED BY

Jacqueline G. Bloomfield,
The University of Sydney, Australia

REVIEWED BY

Anna Siri,
University of Genoa, Italy
Ali Dabbagh,
Shahid Beheshti University of Medical
Sciences, Iran

*CORRESPONDENCE

Jianguang Qi
✉ qjg2006@126.com
Zhehan Jiang
✉ jiangzhehan@bjmu.edu.cn

[†]These authors have contributed equally to
this work and share first authorship

RECEIVED 24 September 2023

ACCEPTED 12 December 2023

PUBLISHED 08 January 2024

CITATION

Li S, Qi X, Li H, Zhou W, Jiang Z and
Qi J (2024) Exploration of validity evidence
for core residency entrustable professional
activities in Chinese pediatric residency.
Front. Med. 10:1301356.
doi: 10.3389/fmed.2023.1301356

COPYRIGHT

© 2024 Li, Qi, Li, Zhou, Jiang and Qi. This is
an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Exploration of validity evidence for core residency entrustable professional activities in Chinese pediatric residency

Shan Li^{1†}, Xin Qi^{2†}, Haichao Li³, Wenjing Zhou⁴, Zhehan Jiang^{5*}
and Jianguang Qi^{1*}

¹Department of Paediatrics, Peking University First Hospital, Beijing, China, ²Department of Plastic Surgery and Burns, Peking University First Hospital, Beijing, China, ³Department of Respiratory and Critical Medicine, Peking University First Hospital, Beijing, China, ⁴School of Public Health, Peking University, Beijing, China, ⁵Institute of Medical Education and National Center for Health Professions Education Department, Peking University, Beijing, China

Introduction: This study seeks to explore validity and reliability evidence for core residency entrustable professional activities (CR-EPAs) that were developed by Peking University First Hospital (PKUFH) in 2020.

Methods: A prospective cohort study was conducted in PKUFH. Trainers (raters) assessed pediatric residents on CR-EPAs over 1 academic year, bi-annually. Critical components within a validity evidence framework were examined: response process (rater perceptions), the internal structure (reliability and contributions of different variance sources), and consequences (potential use of a cutoff score).

Results: In total, 37 residents were enrolled, and 111 and 99 trainers' ratings were collected in Fall 2020 and Spring 2021, respectively. For rater perceptions, all the raters considered CR-EPAs highly operational and convenient. In all ratings, individual EPAs correlate with total EPA moderately, with Spearman correlation coefficients spanning from 0.805 to 0.919. EPA 2 (select and interpret the auxiliary examinations), EPA 5 (prepare and complete medical documents), EPA 6 (provide an oral presentation of a case or a clinical encounter), and EPA 7 (identify and manage the general clinical conditions) were EPAs correlated with other EPAs significantly. The results of the generalizability theory indicated that the variability due to residents is the highest (nearly 78.5%), leading to a large size of the reliability estimates. The matching results indicate that the lowest error locates at 5.933.

Conclusion: The rating showed good validity and reliability. The ratings were reliable based on G-theory. CR-EPAs have a magnificent internal structure and have promising consequences. Our results indicate that CR-EPAs are a robust assessment tool in workplace-based training in a carefully designed setting.

KEYWORDS

competency, postgraduate medical education, entrustable professional activities, pediatrics, China, validity

Introduction

Competency-based medical education (CBME) was widely used in postgraduate medical education (PGME) and has become a symbolic approach to reform medical education (1).

Competence is a complex, interrelated, multidimensional construct to be acquired by medical professionals for safe and effective clinical practice. By adhering to the two core principles of CBME, which involve redefining the criteria for a competent physician and emphasizing the achievement of competent graduates, CBME has demonstrated an enhancement in clinical practice and an improvement in patient safety (2). Given China's large population, it has established a series of policies to ensure a safe health system with only qualified professionals practicing, emphasizing the implementation of CBME in PGME (3). In its capacity as a trailblazer in pediatric residency training within China, Peking University First Hospital (PKUFH) initiated its pediatrics residency program as far back as 1946. Remarkably, this initiative preceded the Chinese National standardized resident training guidelines by nearly seven decades (3). The pediatric residency program at PKUFH was accredited by the Chinese Medical Doctor's Association (CMDA) and provisional accreditation by the Royal College of Physician and Surgeon of Canada (RCPSC). As proposed by Melle et al., we tried to implement five core components of CBME in the pediatric residency program (Appendix 1) (4) since 2012. The establishment of the Core Competency Framework for Residency Education by the China Consortium of Elite Teaching Hospitals in 2018 marked a significant milestone. Prior to this development, a comprehensive framework for residency education had not been in place (5).

The CanMEDS framework of RCPSC has been implemented in our institute as outcome competencies (Appendix 1). A stratified rotation is systematically conducted for each resident, wherein a highly specific competency-based assessment approach, encompassing both formative and summative evaluations, is meticulously delineated. Throughout the course of their rotations in various subspecialties, residents can expect to receive continuous feedback from their trainers and the program director, typically occurring at regular intervals of approximately every 2–3 months. The clinical competency committee (CCC) was held semi-annually. The rotation of certain subspecialties might be extended if a resident fails to meet the minimum criteria.

Through our practice, competency itself somehow is not easy to be understood nor to be observed by trainers, which creates a gap between competency and the daily tasks of doctors. As a tool to close the gap, the milestones initiative showed satisfactory validity as an assessment tool for competency (6); however, it is a rather complex system. Entrustable professional activities (EPAs), initiated by Ten Cate (7), provided a sound alternative approach for assessment in CBME (8–10), which might be the solution for sequenced progression of competency and programmatic assessment. EPAs are a bundle of clinical tasks vital to competency assessment (11, 12); in that way, EPAs can link routine clinical tasks to competency assessment. For instance, the commonplace task of “history taking” necessitates the demonstration of multiple competencies, encompassing roles such as medical expertise, effective communication, advocacy for health, and professionalism. Through the supervision of this “history-taking” process, trainers gain valuable insights into the trainee's overall competency. After supervision, entrustment decision-making was performed accordingly so that the competency could be translated into trusted behavior, which is easier for implementation in a nationwide fashion.

Based on the foundation of CBME, we further began exploring the development of EPA in PGME. In 2020, a modified Delphi method consisted of two iterative rounds and one consensus meeting to develop the core residency entrustable professional activities (CR-EPAs; Figure A1 in Appendix 2). A 15-item competency assessment framework mainly focused on generic and core professional activities, as in Table 1 (13). The supervision rating scale of nine-point Likert items was initially set up according to Chen et al. in EPA implementation studies (14, 15). Considering that the nine-point span is too sparse to be informative in assessing the least requirement for residents (16), we modified the nine-point scale into an eight-point scale (Table 2). To minimize the need for faculty training, we use the same supervision scale level across all the CR-EPAs.

The aforementioned CBME has been implanted in the PKUFH pediatric residency program. Aiming at understanding CR-EPAs in a practical assessment setting, revealing how it would integrate into our previous CBME system and explicit training perils and problems, we conducted a prospective cohort study in the pediatric residency training program of PKUFH. This research aimed to gather validity evidence for CR-EPAs in the pediatric residency training setting and inspect if CR-EPAs could provide reliable and meaningful data for evaluating residents. The study was organized into a three-component investigation: response process (rater perception), internal structure (variance components reliability), and consequences (potential use of a cutoff score).

Materials and methods

Study design

This study aimed to gather validity evidence for CR-EPAs. To gather information ahead and follow the trainees for a period of time,

TABLE 1 CR-EPAs^a.

EPA 1	Gather history and perform physical examination during patient encounter
EPA 2	Select and interpret the auxiliary examinations
EPA 3	Provide diagnosis and differential diagnosis
EPA 4	Develop the comprehensive management plan for patients
EPA 5	Prepare and complete medical documents
EPA 6	Provide oral presentation of a case or a clinical encounter
EPA 7	Identify and manage the general clinical conditions
EPA 8	Identify clinical emergency and critical illness and provide initial management
EPA 9	Transit and hand over the patient
EPA 10	Obtain informed consent for tests and/or procedures
EPA 11	Perform general procedures of a physician
EPA 12	Provide patient education and health advocacy
EPA 13	Deliver bad news to patients and/or family members
EPA 14	Deliver clinical teaching and instruct near-peers
EPA 15	Prepare and respond to public health events

^aPeking University First Hospital; EPAs, entrustable professional activities; CR-EPAs (13), core residency EPAs.

TABLE 2 EPAs supervision scales as used in Peking University First Hospital Pediatric Department.

Description of competence	Original code (17)	Coded in this study
Trusted to observe only	1b	1
Trusted to practice EPA only under proactive full supervision as coactivity with rater	2a	2
Trusted to practice EPA only under proactive, full supervision with rater in room ready to step in as needed	2b	3
Trusted to practice EPA only under reactive/on-demand supervision with rater immediately available, all findings double checked	3a	4
Trusted to practice EPA only under reactive/on-demand supervision with rater immediately available, key findings double checked	3b	5
Trusted to practice EPA only under reactive/on-demand supervision with rater distantly available, findings reviewed	3c	6
Trusted to practice EPA unsupervised	4	7
Trusted to supervise others in practice of EPA	5	8

EPAs, entrustable professional activities.

a prospective cohort study that implemented CR-EPAs in the PKUFH pediatric residency training program over 1 academic year (July 2020 to June 2021) was conducted. The rating was achieved in January 2021 and July 2021, denoting residents' performance in Fall 2020 and Spring 2021, respectively.

The SPSS (version 23.0.0) and R (version 4.0.1) were used for statistical analysis. The Prism (version 9.0.0) was used for visualizing the analysis.

Ethics approval and consent to participate

In accordance with the Declaration of Helsinki, the institutional review board at PKUFH granted this study (2021-107). Before participating, informed written consent were obtained from each resident and trainer in accordance with relevant guidelines and regulations.

Study participants

As forementioned well-launched CBME in PKUFH Department of Pediatrics, both the trainers and residents were familiar with the concept of competency and competency assessment. Residents in the PKUFH Department of Pediatrics residency training program were selected as the subjects of this study. It would maximumly reduce systemic errors to a certain extent. Inclusion criteria were (a) having finished at least 9 months of rotation during the study period and (b) being willing to be assessed by trainers. Residents were divided into different postgraduate year (PGY) groups according to their rotation year in July 2020. The demographic information was collected.

Response process

Trainers rated each resident according to their previous rotated subspecialties. These trainers were, therefore, the raters throughout this article. A series of tutorials on the EPAs' concept was delivered to the trainers to minimize the inter-rater difference.

The rating processes were performed through an online survey platform (<http://www.wjx.cn/>). The trainers could complete the ratings by either computer or mobile devices. Each form contained 15 items of CR-EPAs (Appendix 3). A link containing CR-EPA supervision rating and the list of trainees to be assessed was sent to each trainer every 6 months. For each trainee, multiple times of ratings would be conducted by different trainers according to their previous rotation and performance. To avoid missing data, the survey could only be submitted after all EPAs of a target trainee were fully observed. The raters were allowed to select "unable to rate" certain EPAs if they felt inadequate supervision/observation opportunities or insufficient qualifications. The time consumed for each rating was automatically recorded. For each resident, overall performance was the average of multiple trainers' ratings.

After completing the rating, we compared EPAs across PGYs and assessment periods. A focus interview with eight raters was conducted to document comments and thoughts about using CR-EPAs in practice, ensuring the completeness of the response process investigation. All interviewees' questions were open-ended: "how do you feel when you were using CR-EPAs in practice?"

Internal structure

The internal structure was investigated through (1) associations between any pairs of EPAs via Spearman correlation analysis, (2) discrimination of each EPA via Spearman correlation analysis, and (3) variance and reliability estimation via generalizability theory (G-theory) (18). Specifically, G-theory was used to decompose variance components of the assessment, and the estimation was achieved by using restricted maximum likelihood (REML) (19). As psychometrics theory indicates, the correlation between items (e.g., EPAs in our context) provides evidence for validity: they are expected to have a moderate-to-high correlation to show a good measurement structure. The correlation between an item (again an EPA in our context) and its sum/mean (as an overall performance for the entire assessment) is essentially "discrimination."

Many performance-based assessments are investigated through Cronbach's α , inter-rater reliability, inter-rater agreement, or concordance that all belong to classical test theory (CTT) (20–22). This study, however, utilized G-theory as it is more proper for the setting of our CR-EPAs. The reasons for using G-theory instead of others are listed below:

1. Instigating qualities of education assessment (i.e., validity) also most always involves measurement theories and their quantifying frameworks, including CTT, G-theory, and item response theory (23). Therefore, G-theory is a candidate for the study.
2. Compared with CTT that simply assumes that observed performance consists of true ability effect and error effect (i.e.,

the well-known $X=T+E$ and each effect correspond to variance such that $\sigma^2(X)=\sigma^2(T)+\sigma^2(E)$, G-theory is compatible with designs with *multiple facets* such as raters, items, groups, and occasions (24, 25), each of which is an effect affecting the observed scores. For instance, in performance assessment where a is the $p \times i \times r$ design present (each person p is graded by every rater r on each task/item i), G-theory can decompose observed response data as $X_{pri} = \mu + v_p + v_i + v_r + v_{pi} + v_{ir} + v_{pr} + \epsilon_{pri}$, where an observed score, X_{pri} , for person p on item i rated by rater r is made of the grand mean μ , person effect v_p , item effect v_i , rater effect v_r , interaction terms of any two random effects, and error effect ϵ_{pri} . Each of these effects involves variance as CTT does, and their values can indicate the proportion of an effect contributing to the data. To illustrate, the proportion of rater and item effects count for 80 and 10% of the total variance of the data and then intuitively one would consider the inconsistency between raters is high, while the items are more homogenous.

3. IRT is used more in large-scale standardized (multiple-choice) assessment (26), where the sample sizes are generally large. In certain simplified scenarios, such as scoring with the rating scale, IRT as G-theory can be used interchangeably (27). However, when *multiple facets* are available and non-large-scale scenarios are present, G-theory makes a more appropriate and reliable choice, especially when the designs are complex such as random-distributed and/or nested structures (28).
4. Competence/performance-related investigations through G-theory in the field of medical education have been seen more in the literature (24, 25) also conveying that our methodological adoption is a strong fit for the present study, which involves different EPAs, raters, and randomly crossed structure between raters and residents.

Variance estimates of G-theory allow calculating the level of (1) dependability (criterion- or domain-referenced) and (2) generalizability (e.g., norm-referenced interpretations of test scores), which are akin to reliability in CTT. G-theory enables researchers to make decisions on how to alter the reliability coefficient to a specific level. For instance, if G-theory shows a large variance in the rater effect implying a lack of consistency among themselves, a decision study (namely, D-study) will be informative to the prediction: how many raters are demanded to reach a specific coefficient level. In our study, the effects of raters, items (i.e., EPAs), residents, and their interactions were considered. Their estimation was achieved via the R software (29).

Consequence

Finally, consequence analysis was defined to investigate the potential use of EPA scores in future competency screening. In practice, administrators and raters tend to use an observed (mean or sum) score to evaluate if a resident meets the minimal requirements of the competency assessment. This implementation involves setting a cutoff score that theoretically consists of the least measurement errors or makes the highest sense through scientific reasoning. In this study, we aligned the observed (mean) EPA scores of each resident with true scores (i.e., the ability estimates from G-theory modeling

after excluding other noises such as rater effect and item effect), and the scores' uncertainty/errors yielded from the aforementioned G-theory analysis. Ideally, the cutoff observed score should correspond to the true score level with the lowest uncertainty/errors, namely, the most reliable threshold setting from a data-driven perspective (30).

Results

Descriptive statistics

Thirty-eight pediatric residents were enrolled in this study; one was excluded due to incomplete rotation. The demographics of 37 residents are shown in Table 3. In total, 23 raters (trainers) participated in the assessment, and their demographics are shown in Table 4.

In total, 111 and 99 ratings were received for the two investigation periods, respectively. Each resident received 3.0 ± 1.2 ratings in Fall 2020 and 2.7 ± 0.7 in Spring 2021 from trainers, respectively. Since each EPA needs to be completed before submission, no data for the EPA assessment was missing. The supervision rating results of each EPA through PGY1 to PGY3 in Fall 2020 and Spring 2021 are shown in Figure 1.

Response process

All the raters (trainers) fulfilled the supervision rating within 2 weeks of the assignment. None reported "unable to rate." The rating data were

TABLE 3 Demographic of 37 participating pediatric residents.

Demographic	No. (%), except where noted
<i>Gender</i>	
Female	10 (27)
Male	27 (73)
Age, average \pm SD	26.8 \pm 3.1
<i>Level of training</i>	
PGY-1 resident	13 (35)
PGY-2 resident	11 (30)
PGY-3 resident	13 (35)
Number of ratings of each resident in fall 2020, average \pm SD	3.0 \pm 1.2
1	3 (8)
2	12 (32)
3	9 (24)
4	8 (22)
5	5 (14)
Number of ratings of each resident in in spring 2021, average \pm SD	2.7 \pm 0.7
1	2 (5)
2	12 (33)
3	19 (51)
4	4 (11)

PGY, postgraduate year.

TABLE 4 Dermographics of 23 trainers (raters).

Demographic	No. (%), except where noted
<i>Gender</i>	
Female	18 (78.2)
Male	5 (21.8)
<i>Age, median (range)</i>	
	33 (29–47)
<i>Degree</i>	
Medical doctors	21 (91.3)
Master's degree	3 (8.7)
<i>Year engaged in teaching</i>	
<1 year	4 (17.5)
1 to <2 years	7 (30.4)
2 to <3 years	5 (21.7)
≥3 years	7 (30.4)
<i>Subspecialty</i>	
General pediatric	1 (4.3)
Respiratory	2 (8.6)
Pediatric intensive care	2 (8.6)
Neurology	4 (17.2)
Nephrology	3 (13.0)
Cardiology	2 (8.6)
Neonatology	4 (17.2)
Hematology	2 (8.6)
Neonatal intensive care	3 (13.0)

analyzed, and the results were reviewed with the raters (Figure A2 in Appendix 2). The emergent theme of both strengths and limitations was listed during the focus interview in Table 5. All the raters considered CR-EPAs to be highly operational and convenient. The conclusion is consistent with the time consumed to complete one rating—the minimum time is 37 s, while the maximum is 1700 s, with a median of 143 s. The raters believed that the proposed CR-EPA supervision rating was consistent with their clinical observations of a specific resident and a realistic reflection of the resident's clinical competence. If using traditional assessment tools (such as 360° assessment), raters often overestimate residents' "actual behavior" and cannot genuinely differentiate students' levels. However, it would be more mutual and objective when using supervision decision-making as the measurement scale, thus achieving good consistency among different raters.

Internal structure

In all ratings, individual EPAs correlate with total EPA moderately, with Spearman correlation coefficients spanning from 0.805 to 0.919, recorded in Figure 2, indicating that, overall, items nested within the target assessment possess good power in distinguishing residents' competency. Spearman correlation coefficient between all EPA pairs ranged from 0.541 to 0.926, with a median of 0.759 (Q1 0.697, Q3 0.827), recorded in Table A1 in Appendix 2. None of the EPA pairs' correlation was below 0.3. In total, 20 out of 105 (19%) individual EPA

pairs' correlation coefficient was above 0.85. Among those EPA pairs, EPA 2 (select and interpret the auxiliary examinations), EPA 7 (identify and manage the general clinical conditions), EPA 6 (provide an oral presentation of a case or a clinical encounter), and EPA 5 (prepare and complete medical documents) were the ones that significantly correlated with other counterparts.

The results of generalizability theory modeling were presented in Table 6, showing the estimates of the variance components for residents, raters, EPAs, the interactions between residents and EPAs, the interactions between raters and EPAs, the interactions between raters and EPAs, as well as the residuals. Variability due to residents is the highest (nearly 78.5%), leading to a large size of the reliability estimates (i.e., G-Coefficient and Φ -Coefficient estimates are both higher than 0.75). Two interaction terms—the one between residents and EPAs and the other between raters and residents—are lower than 1%, implying that these "noisy signals" are barely influential to the assessment. On the other hand, the interaction between raters and residents, although not high in a relative sense, shows the total variability related to raters is 8.4% (i.e., 0.099, 0.130, and 0.005). The contributions from EPAs are low and indicate a high consistency between the item levels.

Consequences

At each given observed mean score (i.e., overall performance), bootstrapping the G-theory yielded a set of true score estimates allowing the construction of an uncertainty range, which reflects the estimation precision. The matching results are contained in Figure 3, indicating that the lowest error is located at 5.933. On the other hand, when the overall performance is 2.700, the errors become the largest—1.069.

Discussion

To the best of our knowledge, this is the first study in China to implement EPAs in the pediatric residency program. In our study, rating data sets on CR-EPAs during 1 academic year were employed to verify their reliability and validity, emphasizing the educational and clinical outcome of CR-EPA-based assessment in China's 3-year standardized pediatric residency training.

Competency assessments require good reliability. The generalizability coefficients of CR-EPAs are 0.871, the Φ -coefficient of CR-EPAs is 0.785, and both coefficients are above 0.75, which indicates that the use of instruments in the given context is reliable. In comparison with other assessments of this kind, the reliability estimates are sufficiently high for performance-based tasks. For example, Meyer et al. (31) showed that the generalizability coefficients of a pilot preclinical entrustment assessment in undergraduate medical education fall between 0.66 and 0.52. The coefficients are decomposed into different sources, of which the resident effect dominates the variance. Surprisingly, the rater and the rater-related interaction effects consume a larger partition than others, which is usual because ratings are too subjective to be highly consistent among all raters. However, the proportions of the rater and the rater-related effects are still low, reflecting positive signs of the series of rater training prior to the present study.

Supervision decision-making is widely used in pediatrics throughout the world (32). PKUFH Pediatrics Department has a long

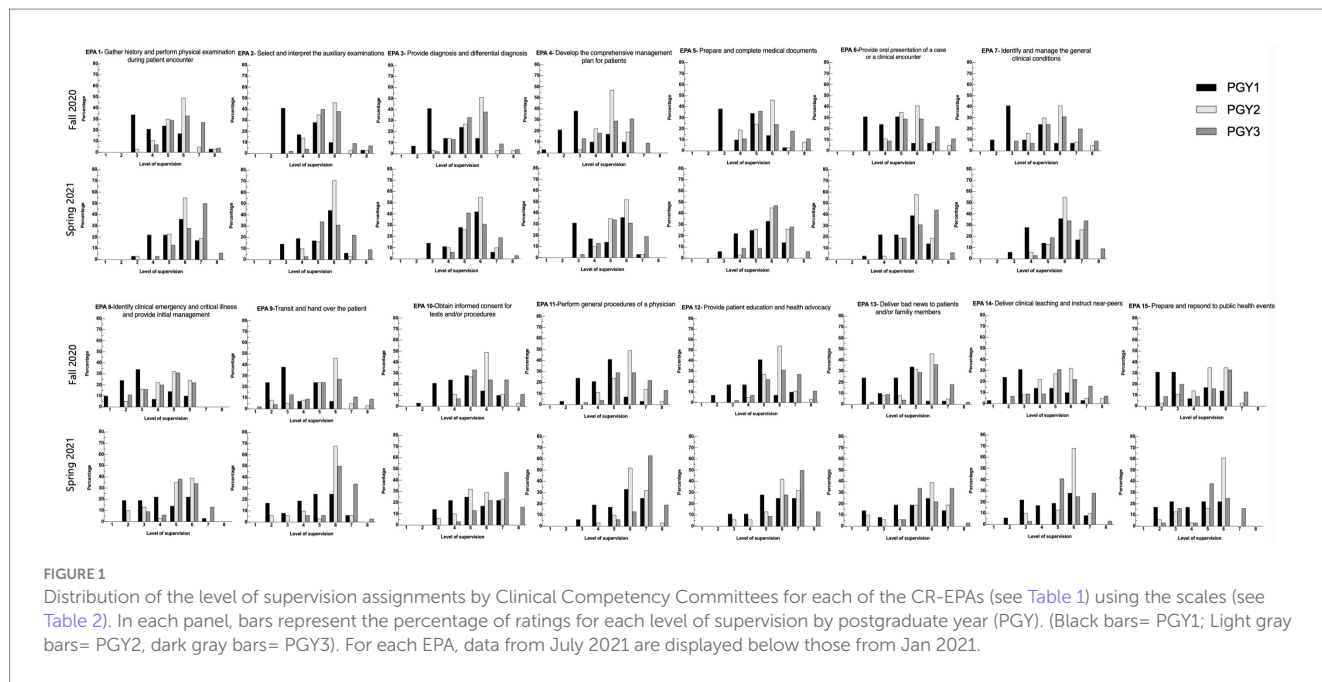


TABLE 5 Aspects of the CR-EPAs supervision decision-making.

Emergent theme	Description
<i>Favorable aspects</i>	
Characteristics of CR-EPAs	CR-EPAs covered major generic clinical behavior and was a clinical-based, bed-side based assessment. Raters could easily make a supervision decision based on residents' clinical behavior, and CR-EPAs supervision level could reflect residents' actual clinical competency.
Comparing to other assessment tools	CR-EPAs was a more mutual and objective assessment tool, compared with 360° assessment, and is clinical-based assessment compared with traditional structural case-interview. CR-EPAs illustrated more clinical competencies compared with mini-CEX and DOPS.
<i>Areas to improve</i>	
Assessment Platform	CR-EPAs should become a regular assessment with an interval of 2 to 3 months based on the rotation of residents. The assessment should have a specified electronic platform which is easy to review the previous results both for residents and raters and should have longitudinal data for certain residents and raters.
Lack of discriminations between PGY2 and PGY3 residents	Data were lack of discriminations between PGY residents. CR-EPAs was a generic clinical behavior. PGY-2 and PGY-3 residents shared similar responsibilities. The growth trajectory would be flat in the next 2 years. EPAs based on general pediatric training should be developed in future to fulfill the gap.

history of CBME. Hence, reliability could reach such a high level. Quoted for the focus review, supervisors describe CR-EPAs as a clinical-based, bed-side-based assessment tool, and supervisors could easily make a supervision decision based on clinical behavior. CR-EPAs were more reciprocal and objective when using supervision decision-making as the measurement scale, thus achieving good consistency among different raters.

In this study, CR-EPAs also showed relatively good validity evidence. The assessments made in Spring 2021 were higher than those made in Fall 2020 in most EPAs (Figure A2 in Appendix 2). These results were consistent with the expectation that residents require less supervision as their skills improve with more experience and teaching. The developmental trajectories are evident in aggregate data, and similar results were found in other studies based on pediatric fellows. However, there was no upward trajectory from Fall to Spring for EPA 15 (apply and respond to public health events) possibly due to a lack of practice and observation. It is a common problem the world is facing in the post-pandemic era. A specialized training course

should be implemented to enhance the training. As for the internal structure validity evidence, individual EPAs moderately correlate with the EPA total score, indicating that items nested within the target assessment possess good power in distinguishing residents' competency. As for the inter-EPA correlation analysis, most EPA pairs were moderately correlated, while a few EPAs were significantly correlated with each other. Those EPAs are EPA 2 (select and interpret the auxiliary examinations), EPA 5 (prepare and complete medical documents), EPA 6 (provide an oral presentation of a case or a clinical encounter), and EPA 7 (identify and manage the general clinical conditions). Those EPAs were the most fundamental meta-EPAs and correlated with other EPAs significantly. Those EPAs included common clinical scenarios were that more observable for supervisors and, hence, more comfortable in judgment.

Although serving as an initial exploration in the present study, the consequence part provides insights into the decision-making use of EPAs. The data-driven result shows that supervision Level 6 (5.933) nearly corresponds to the lowest error for assigning residents to binary

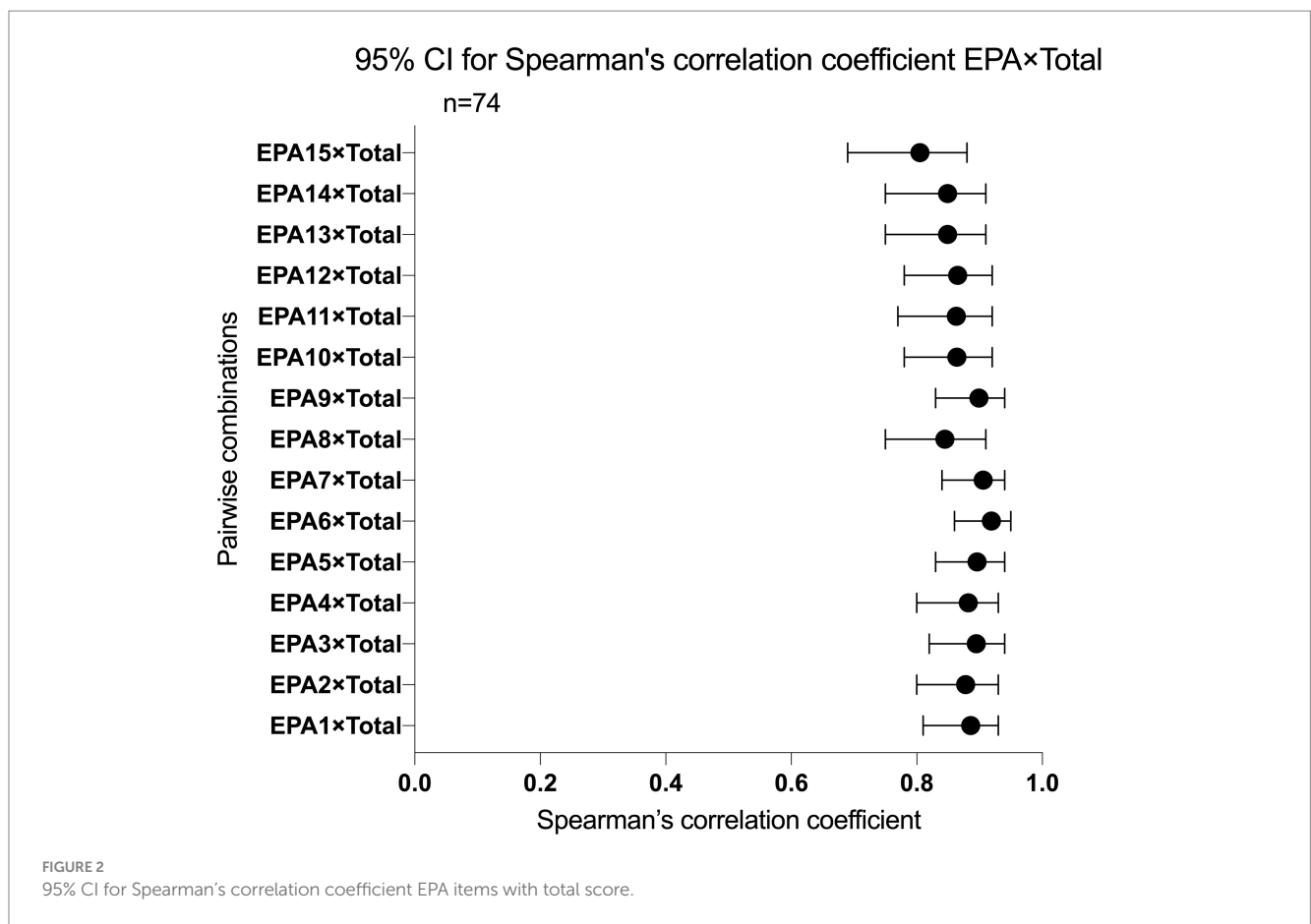


TABLE 6 Variance decomposition and reliability estimates via generalizability theory.

Effects	Variance component (VC)	% VC	df	G-Coefficient	Φ-Coefficient
Resident	0.925	78.5	1	0.871	0.785
Rater	0.099	8.4	3		
EPA	0.013	1.1	15		
Resident:EPA	0.000	0	15		
Rater:EPA	0.005	0.4	45		
Rater:Resident	0.130	11	3		
Residual	0.007	0.6	45		

classes, which are conventionally interpreted as “pass/fail,” “competent/incompetent,” and “qualified/unqualified.” The cutoff score is well aligned with Chen (17) who claims that level 6 (Chen’s Level 3c) should be the threshold when residents graduate from a program (the end of PGY-3 training). Hence, from the perspective of G-theory, Level 6 should be the cutoff value for supervision levels of CR-EPAs for residents.

residents. The study was conducted only within 1 academic year in a single center. Furthermore, as the lack of an electronic platform specified for CR-EPAs ratings, our ratings were conducted on a survey platform, and all the EPAs were listed in a single survey; This may lead to a halo effect. Fortunately, an e-portfolio specified for CR-EPAs ratings will launch in our center soon, which can solve the problem in future.

Limitations

Our study has several limitations. It is a single-center, small sample-size study. The supervision rating timespan was half a year. It would be less likely to reflect a real-time improvement in the

Conclusion

We developed an eight-level supervision scale for CR-EPAs and implemented it in a pediatric residency training program of

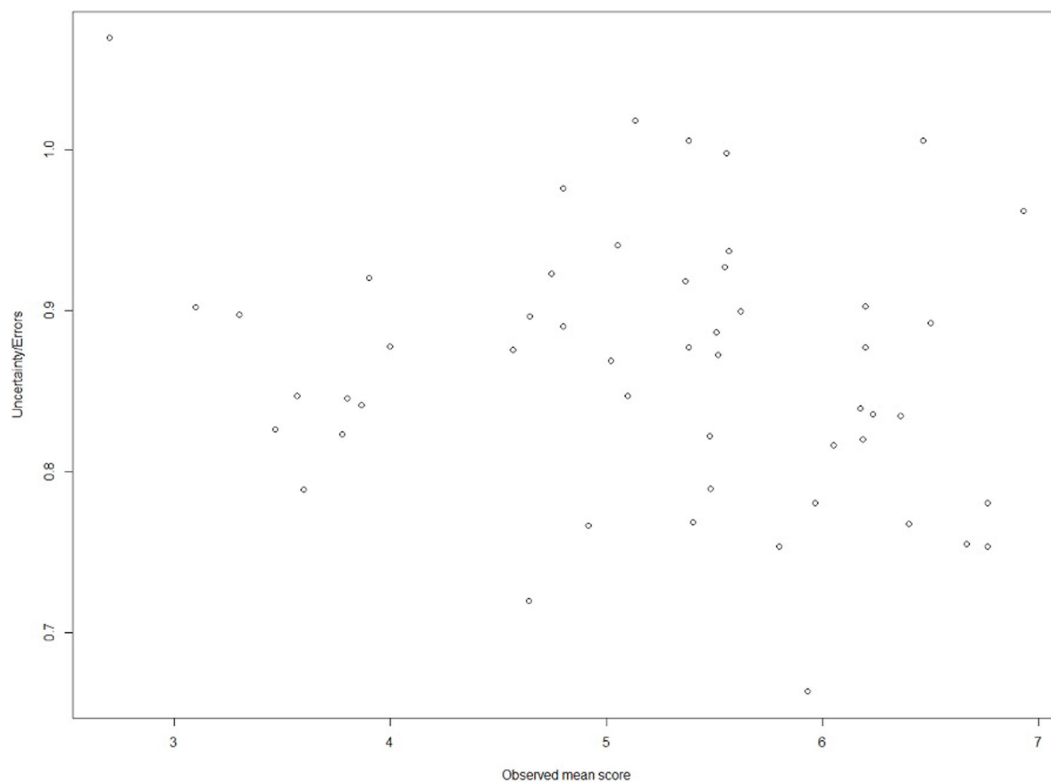


FIGURE 3
Relation between observed mean scores and corresponding uncertainty in a cut-off setting.

PKUFH. The ratings were reliable based on G-theory. CR-EPAs have a fine internal structure and the consequences of using them for binary decision shows reasonable utility. Our results indicate that CR-EPAs can serve as a robust assessment tool in workplace-based training in a carefully designed setting.

editing. WZ: Formal analysis, Validation, Writing – original draft. ZJ: Data curation, Formal analysis, Methodology, Validation, Writing – review & editing. JQ: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Visualization, Writing – review & editing.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the institutional review board at Peking University First Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SL: Formal analysis, Investigation, Project administration, Writing – original draft, Software. XQ: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. HL: Conceptualization, Resources, Visualization, Writing – review &

Funding

The author(s) declare financial support was received for the research, authorship, and /or publication of this article. This study was supported by the educational Teaching Research Project in Peking University Health Science Center under Grant [2020YB31] and [2022YB34]; the Medical Specialist Standardized Training Research Project in Peking University Health Science Center under Grant [2019ZP04] and [2022ZP02]; Peking University Research Project in 2021; The Center for Institutional Research under Grant [2021ZZ04]; National Natural Science Foundation of China under Grant [72104006] and National Medical Examination Center.

Acknowledgments

We are grateful for the excellence advice from Li Li at Zhongnan Hospital of Wuhan University, Wuhan, China. We are also grateful for all hardworks of all the residents and teachers at Peking University First Hospital.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1301356/full#supplementary-material>

References

- Wang W. Medical education in China: progress in the past 70 years and a vision for the future. *BMC Med Educ.* (2021) 21:453. doi: 10.1186/s12909-021-02875-6
- Touchie C, Ten Cate O. The promise, perils, problems and progress of competency-based medical education. *Med Educ.* (2016) 50:93–100. doi: 10.1111/medu.12839
- Guiding Opinions on Establishing the Standardised Residency Training (2013). 56. Available at: http://www.gov.cn/gzdt/2014-01/17/content_2569096.htm.
- van Melle E, Frank JR, Holmboe ES, Stockley D, Sherbino J |International Competency-based Medical Education Collaborators. A core components framework for evaluating implementation of competency-based medical education programs. *Acad Med.* (2019) 94:1002–9. doi: 10.1097/ACM.0000000000002743
- Education CCOETHFR. Consensus on Core competency framework for residency education among China consortium of elite teaching hospitals for residency education. *Med J Peking Union Med Coll Hosp.* (2022) 13:17–23. doi: 10.12290/xhyxzz.2021-0755
- Schuwirth LWT, Van Der Vleuten CPM. A history of assessment in medical education. *Adv Health Sci Educ Theory Pract.* (2020) 25:1045–56. doi: 10.1007/s10459-020-10003-0
- Ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ.* (2005) 39:1176–7. doi: 10.1111/j.1365-2929.2005.02341.x
- Ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med.* (2007) 82:542–7. doi: 10.1097/ACM.0b013e31805559c7
- Englander R, Flynn T, Call S, Carraccio C, Cleary L, Fulton TB, et al. Toward defining the foundation of the MD degree: core entrustable professional activities for entering residency. *Acad Med.* (2016) 91:1352–8. doi: 10.1097/ACM.0000000000001204
- Caverzagie KJ, Cooney TG, Hemmer PA, Berkowitz L. The development of entrustable professional activities for internal medicine residency training: a report from the Education Redesign Committee of the Alliance for Academic Internal Medicine. *Acad Med.* (2015) 90:479–84. doi: 10.1097/ACM.0000000000000564
- Hanson JL, Bannister SL. To trust or not to trust? An introduction to Entrustable professional activities. *Pediatrics.* (2016) 138:e20162373. doi: 10.1542/peds.2016-2373
- Ten Cate O, Schumacher DJ. Entrustable professional activities versus competencies and skills: Exploring why different concepts are often conflated. *Adv Health Sci Educ Theory Pract.* (2022) 27:491–9. doi: 10.1007/s10459-022-10098-7
- Xin Q, Zhe J, Xiaoning H, Hong Z, Yuan L, Aimei D, et al. Establishment of the entrustable professional activities for the residents. *Chin. J. Med. Educ.* (2021) 41:104–8. doi: 10.3760/cma.j.cn115259-20201117-01599
- Ten Cate O, Schwartz A, Chen HC. Assessing trainees and making entrustment decisions: on the nature and use of entrustment-supervision scales. *Acad Med.* (2020) 95:1662–9. doi: 10.1097/ACM.0000000000003427
- Ryan MS, Khamishon R, Richards A, Perera R, Garber A, Santen SA, et al. A question of scale? Generalizability of the ottawa and chen scales to render entrustment decisions for the core EPAs in the workplace. *Acad Med.* (2022) 97:552–61. doi: 10.1097/ACM.0000000000004189
- Schumacher DJ, Turner DA. Entrustable professional activities: reflecting on where we are to define a path for the next decade. *Acad Med.* (2021) 96:S1–5. doi: 10.1097/ACM.0000000000004097
- Chen HC, Van Den Broek WE, Ten Cate O. The case for use of entrustable professional activities in undergraduate medical education. *Acad Med.* (2015) 90:431–6. doi: 10.1097/ACM.0000000000000586
- Brennan RL. *Generalizability Theory*. New York, NY: Springer (2001).
- Jiang Z, Raymond M, Shi D, DiStefano C. Using a linear mixed-effect model framework to estimate multivariate generalizability theory parameters in R. *Behav Res Methods.* (2020) 52:2383–93. doi: 10.3758/s13428-020-01399-z
- Robertson B, Schumacher L, Gosman G, Kanfer R, Kelley M, DeVita M, et al. Simulation-based crisis team training for multidisciplinary obstetric providers. *Simul Healthc.* (2009) 4:77–83. doi: 10.1097/SIH.0b013e31819171cd
- Barth J, De Boer WE, Busse JW, Hoving JL, Kedzia S, Couban R, et al. Inter-rater agreement in evaluation of disability: systematic review of reproducibility studies. *BMJ.* (2017) 356:j14. doi: 10.1136/bmj.j14
- Molleman E, Van Der Vegt GS. The performance evaluation of novices: The importance of competence in specific work activity clusters. *J Occup Organ Psychol.* (2007) 80:459–78. doi: 10.1348/096317906X154469
- Brennan RL. Generalizability theory and classical test theory. *Appl Meas Educ.* (2010) 24:1–21. doi: 10.1080/08957347.2011.532417
- Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* (2001) 357:945–9. doi: 10.1016/S0140-6736(00)04221-5
- Jiang Z, Shi D, Distefano C. A short note on optimizing cost-generalizability via a machine-learning approach. *Educ Psychol Meas.* (2021) 81:1221–33. doi: 10.1177/0013164421992112
- Von Davier M, Sinharay S. Analytics in international large-scale assessments: item response theory and population models. In L. Rutkowski, Davier M. von and D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*. Boca Raton, FL: CRC Press; (2014).
- Choi J, Wilson MR. Modeling rater effects using a combination of generalizability theory and IRT. *Psychol Test Assess Model.* (2018) 60:53–80.
- Marcoulides GA. Generalizability theory: picking up where the rasch IRT model leaves off? In: SE Embretson and SL Hershberger, editors. *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Mahwah, NJ: Erlbaum (1999). 129–52.
- Jiang Z. Using the linear mixed-effect model framework to estimate generalizability variance components in R: a lme4 package application. *Methodol Eur J Res Methods Behav Soc Sci.* (2018) 14:133.
- Jiang Z, Raymond M, DiStefano C, Shi D, Liu R, Sun J. A Monte Carlo study of confidence interval methods for generalizability coefficient. *Educ Psychol Meas.* (2022) 82:705–18. doi: 10.1177/00131644211033899
- Meyer EG, Boulet JR, Monahan PB, Durning SJ, Uijtdehaage S. A pilot study of the generalizability of preclinical entrustment assessments in undergraduate medical education. *Acad Med.* (2022) 97:562–8. doi: 10.1097/ACM.0000000000004590
- Kerth JL, Van Treel L, Bosse HM. The use of Entrustable professional activities in pediatric postgraduate medical education: a systematic review. *Acad Pediatr.* (2021) 22:21–8. doi: 10.1016/j.acap.2021.07.007