Validity of Level of Supervision Scales for Assessing Pediatric Fellows on the Common Pediatric Subspecialty Entrustable Professional Activities

Richard B. Mink, MD, MACM, Alan Schwartz, PhD, Bruce E. Herman, MD, David A. Turner, MD, Megan L. Curran, MD, Angela Myers, MD, MPH, Deborah C. Hsu, MD, MEd, Jennifer C. Kesselheim, MD, EdM, Carol L. Carraccio, MD, MA, and the Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN)

Abstract

Purpose

Entrustable professional activities (EPAs) represent the routine and essential activities that physicians perform in practice. Although some level of supervision scales have been proposed, they have not been validated. In this study, the investigators created level of supervision scales for EPAs common to the pediatric subspecialties and then examined their validity in a study conducted by the Subspecialty Pediatrics Investigator Network (SPIN).

Method

SPIN Steering Committee members used a modified Delphi process to develop unique scales for six of the seven common EPAs. The investigators sought validity evidence in a multisubspecialty study in which pediatric fellowship program directors and Clinical Competency Committees used the scales to evaluate fellows in fall 2014 and spring 2015.

Results

Separate scales for the six EPAs, each with five levels of progressive entrustment, were created. In both fall and spring, more than 300 fellows in each year of training from over 200 programs were assessed. In both periods and for each EPA, there was a progressive increase in entrustment levels, with second-year fellows rated higher than first-year fellows ($P < .001$) and third-year fellows rated higher than second-year fellows ($P < .001$). For each EPA, spring ratings were higher ($P < .001$) than those in the fall. Interrater reliability was high (Janson and Olsson's iota = 0.73).

Conclusions

The supervision scales developed for these six common pediatric subspecialty EPAs demonstrated strong validity evidence for use in EPA-based assessment of pediatric fellows. They may also inform the development of scales in other specialties.

The Accreditation Council for Graduate Medical Education (ACGME) Milestone Project[1] and the construct of entrustable professional activities (EPAs), developed by ten Cate and others,[2–4] were introduced with the hope of addressing some of the long-standing challenges of assessing clinical competence in graduate medical education (GME). The Milestone Project required every specialty and subspecialty to define trainee performance levels for each of their competencies. Consequently, milestones provide a shared mental model of trainee behaviors at a given level of performance.[5] Alternatively, EPAs represent the routine and essential activities of a practicing physician that can be observed and measured.[6] In the aggregate, EPAs define the specialty. EPAs integrate competencies into the authentic care of patients using a lens of required level of supervision in contrast to level of trainee competence.[2]

Believing that assessment across the continuum of education, training, and practice could best be achieved by a framework of competencies, milestones, and EPAs, the American Board of Pediatrics (ABP), with the help of the Council of Pediatric Subspecialties (CoPS), convened a two-day meeting in 2013 that brought together thought leaders from each of the subspecialties to develop common subspecialty EPAs.[7] Seven such EPAs were created.

Since the introduction of EPAs, scales have been proposed that define levels of supervision leading to entrustment.[3,8–13] These scales have received widespread attention as clinical faculty intuitively relate to the concept of assigning a level of supervision, which aligns with what they do when supervising trainees in the care provided to patients in clinical settings.[14,15]

The purpose of this research was to build a validity argument for the use of the supervision scales we developed to assess pediatric fellows on six of the seven common pediatric subspecialty EPAs for the 14 pediatric subspecialties with ABP certification.[16] We hypothesized that a supervision scale based on what a fellow could be trusted to perform (e.g., simple vs. complex cases) and the type of required oversight (e.g., direct supervision, indirect supervision, coaching) that aligns with the specific EPA would lead to assessments supported by validity evidence.

Method

Our validity evidence is presented in the framework advocated by Messick,[17] Downing,[18] and Cook and Beckman,[19] including the categories of content, response process, internal structure, and relations with other variables.

Content

During May–October 2014, the members of the Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN), an educational research network,[20] used a modified Delphi approach to develop the level of

Please see the end of this article for information about the authors.

Correspondence should be addressed to Richard Mink, Harbor-UCLA Medical Center, 1000 W. Carson St., Box 491, Torrance, CA 90509; telephone: (310) 222-4002; e-mail: rmink@ucla.edu.

Acad Med. XXXX;XX:00–00.
First published online
doi: 10.1097/ACM.0000000000001820

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A463.

supervision scales for six of the seven common pediatric subspecialty EPAs. (The seventh EPA, the scholarship EPA, had not yet been fully developed.) This committee is composed of up to two representatives from each of the 14 pediatric subspecialties that have ABP certification, as well as representatives from CoPS, the ABP, the Association of Pediatric Program Directors Fellowship Executive Committee, and the Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network (APPD LEARN).[21] All SPIN Steering Committee members have significant GME experience: Of the 21 individuals on the committee, 18 (86%) served as a program director and 5 (24%) had an advanced degree in medical education. The committee met via videoconference calls and in-person conferences.

Our goal was to create scales that were consistent with current approaches to fellow supervision, were intuitive to faculty so as to minimize the need for faculty development, were consistent across the six EPAs at the same level of supervision, and had progressive levels leading to entrustment. We reviewed published supervision scales related to medical education[3,8,9] but found that these were not applicable to the common pediatric subspecialty EPAs and that no single scale was suitable for all EPAs.

Study participants

Members of the SPIN Steering Committee recruited fellowship programs within their subspecialty for the study. The goal was to recruit at least 20% of programs in each ABP-certified subspecialty. IRB approval was obtained from each participating institution, as well as from the University of Utah (which served as the lead site), with most waiving trainee consent. Fellows did not participate in scale development.

Assessments of fellows in each training program were provided by both the fellowship program director (FPD) and the program's Clinical Competency Committee (CCC). Approximately one week before the CCC meeting, the FPDs were asked to use our scales to assign a level of supervision to each of their fellows for each of the six common pediatric subspecialty EPAs. Then, at the CCC meeting, the CCC members used our scales to assign a level of supervision to each of their fellows for the six EPAs, unaware of the FPD ratings. Assessments were made during the milestone collection periods in fall 2014 (November–December) and spring 2015 (May–June). All data were submitted by December 2015.

Response process

Prior to rating levels of supervision for each fellow and to preserve learner anonymity, the FPDs created a unique participant identifier number using an algorithm previously developed by APPD LEARN. Once the identifier was created, links to online data entry tools specific to the FPD and CCC were provided. For each EPA, these tools included the functions or activities needed to safely and effectively perform the EPA[7] followed by the specific level of supervision scale. There was no option to select a rating between two levels. The instructions specified that assignments should be based on what a fellow would be *trusted* to do in performing the activities, not necessarily on what had been actually observed. Copies of the data collection tools could be printed by those FPDs who preferred to upload the information later.

No centralized faculty development was provided on using the scales or explaining the EPA concept. However, FPDs with questions about data entry could send them to the study coordinating center at the Los Angeles Biomedical Research Institute (LA BioMed) at Harbor-UCLA Medical Center. Both the FPDs and CCCs were advised that there was no expectation that fellows would achieve any specific level of supervision in any particular year of training. Moreover, at the time the spring ratings were completed, neither the FPDs nor CCCs were provided access to the assignments made in the fall.

Internal structure

We measured inter-EPA reliability with Cronbach's alpha. Interrater reliability between the FPD and CCC for each of the ratings was calculated with Spearman's ρ. Multivariable interrater reliability between the assessments of the FPD and CCC for each fellow across the six EPAs was determined with Janson and Olsson's iota.[22]

Relations with other variables

For each EPA and for each data collection period, we compared the assigned levels of supervision by year of training with Kruskal–Wallis and Wilcoxon tests. In addition, using paired observations and for each EPA, we compared data from fall 2014 with data from spring 2015 using the Wilcoxon signed-rank test.

Additional data were obtained from the FPDs, including the name of the subspecialty, the number of fellows in the program, the number of years as an FPD, and whether the FPD was a member of the CCC. The FPDs were also asked to self-rate their expertise in understanding milestones and EPAs using a four-point scale (unfamiliar, basic, in-depth, expert). The contribution of each of these factors in the assignment of level of supervision for each year of training was examined using a series of linear mixed models (one for each factor) controlling for clustering within programs. Because data about the number of fellows in the program and the FPD's years of experience were skewed, these variables were analyzed both as continuous values and as quartiles. A P value $< .05$ was considered significant. Data analyses were conducted using R 3.2.2[23] and the lme4 package.[24]

Results

Content

The SPIN Steering Committee developed the EPA level of supervision scales over four videoconference calls and at two in-person conferences, involving nine different iterations. For each of the six common pediatric subspecialty EPAs, we created a separate five-point level of supervision scale, with a higher rating indicating that less supervision is required (Table 1). Three GME experts separate from SPIN reviewed early drafts, and their comments were subsequently incorporated into the assessment scales. In the final versions, three themes for level of supervision emerged across the six EPAs: (1) the complexity of the case (simple vs. complex); (2) the degree of trainee participation (contribute, mentor, lead); and (3) the amount of coaching required to perform the activity at the local or national level.

Response process

Over 200 programs (each represented by a single FPD) from approximately 80 institutions provided data about their trainees in fall 2014 and in spring 2015. The results include supervision-level

Table 1

Level of Supervision Scales for Rating Fellows on Six of the Seven Common Pediatric Subspecialty Entrustable Professional Activities (EPAs)[a]

	Level of supervision[b]				
EPA	1	2	3	4	5
Facilitate handovers to another health care provider	Trusted to *observe only*	Trusted to execute with *direct supervision and coaching*	Trusted to execute with *indirect supervision* with verification of information after the handover for selected *simple and complex* cases	Trusted to execute with *indirect supervision* with verification of information after the handover for selected *complex* cases	Trusted to execute *without supervision*
Lead an interprofessional health care team	Trusted to *participate only*	Trusted to lead with *direct supervision and coaching*	Trusted to lead with *supervisor occasionally present* to provide advice	Trusted to lead *without supervisor present* but requires *coaching* to improve *member and team performance*	Trusted to lead *without supervision* to improve *member and team performance*
Contribute to the fiscally sound and ethical management of a practice	Trusted to *observe only*	Trusted to perform with *direct supervision and coaching* with supervisor *verifying* work product for accuracy	Trusted to perform with supervisor serving as a *consultant* for all tasks	Trusted to perform with supervisor serving as a *consultant* but only for *complex* tasks	Trusted to perform *without supervision*
Provide for and obtain consultation with other health care providers caring for children	Trusted to *observe only*	Trusted to execute with *direct supervision and coaching*	Trusted to execute with *indirect supervision* and discussion of information conveyed for selected *simple and complex* cases	Trusted to execute with *indirect supervision* and may require discussion of information conveyed but only for selected *complex* cases	Trusted to execute *independently without supervision*
Lead within the subspecialty profession	Trusted to *observe only*	Trusted to *contribute* to advocacy and public education activities for the subspecialty profession with *direct supervision and coaching* at the *institutional* level	Trusted to *contribute* to advocacy and public education activities for the subspecialty profession with *indirect supervision* at the *institutional* level	Trusted to *mentor* others and *lead* advocacy and public education activities for the subspecialty profession at the *institutional* level	Trusted to *lead* advocacy and public education activities for the subspecialty profession at the *regional and/or national* level
Apply public health principles and improvement methodology to improve care for populations, communities, and systems	Trusted to *observe only*	Trusted to *contribute with direct supervision and coaching* as a member of a collaborative effort to improve care at the *institutional* level	Trusted to *contribute without direct coaching* as a member of a collaborative effort to improve care at the *institutional* level	Trusted to *lead* collaborative efforts to improve care for populations and systems at the *institutional* level	Trusted to *lead* collaborative efforts to improve care at the level of populations and systems at the *regional and/or national* level

[a]The level of supervision scales were developed by the Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN) to rate required supervision for fellows in six of the seven common pediatric subspecialty EPAs.[7]

[b]In each scale, higher supervision level ratings indicate that less supervision is required. Italics emphasize key words in the description.

assignments for over 1,000 fellows, with more than 300 fellows per year of training evaluated in each period and over 280 paired observations for trainees in each year of fellowship (Table 2). In addition, the goal of having at least 20% of programs in each subspecialty participate was met by 11 (79%) of the 14 subspecialties in both data collection periods.

The FPDs and CCCs assigned levels of supervision without specific prior faculty development. As noted above, the instructions included on the data entry form specified that assignments were to be based on the activities that a fellow would be *trusted* to do, not on what was actually observed. Calls made to the study coordinating center were associated with questions about creating the participant ID or obtaining the Web link to the data entry tools, not about the scales themselves or EPAs.

In both the fall and the spring, and within each year of training, the supervision levels assigned to fellows varied significantly by EPA ($P < .001$ in all cases), suggesting that neither the FPDs nor CCCs rated fellows uniformly at one particular level across EPAs (for CCC ratings, see Figure 1). In each period, the rating of a particular fellow by the FPD or the CCC was the same across all EPAs in less than 5% of trainees.

Internal structure

Interrater reliability between the level of supervision ratings assigned by the FPD and the CCC for each EPA varied: Spearman's ρ ranged from 0.67 to 0.79, with all values statistically significant ($P < .001$; Table 3). Jansen and Olsson's iota, a measure of multivariate interrater reliability, was 0.74 in the fall and 0.74 in the spring. Inter-EPA reliability as measured by Cronbach's alpha was 0.92 in both periods.

Table 2
Participation in the SPIN EPA Study by Rating Period, Fall 2014 (November–December) and Spring 2015 (May–June)

Study participants	Fall 2014	Spring 2015
Institutions, no.	78	81
Programs, no.	208	209
Percentage of subspecialties with program participation > 20%[a]	79%	79%
Total fellows evaluated, no.	1,011	1,036
First-year fellows	352	369
Second-year fellows	332	336
Third-year fellows	327	331
Paired observations, no.	880	
First-year fellows	308	
Second-year fellows	287	
Third-year fellows	285	

Abbreviations: SPIN indicates Subspecialty Pediatrics Investigator Network; EPA, entrustable professional activities.
[a]Fellowship programs from the 14 pediatric subspecialties with American Board of Pediatrics certification were invited to participate.[16]

Relations with other variables

In both the fall and the spring, level of supervision ratings of second-year fellows were higher than those of first-year fellows ($P < .001$), and ratings of third-year fellows were higher than those of second-year fellows ($P < .001$; CCC data are shown in Figure 2). The results were the same whether the assignments were made by the FPD or CCC. For each year of training and using the paired data, assessments made in the spring were higher than those made in the fall ($P < .001$).

We also evaluated the contribution of other factors to the assigned level of supervision, controlling for program and fellow year of training. Neither years of experience of the FPD nor whether the FPD was a member of the CCC was significantly associated with the assigned supervision level ($P > .05$). Program size, evaluated both by number of fellows in the program and by quartiles, was also not significantly associated with assigned supervision level ($P > .05$).

Only 3 (1.3%) of the 228 FPDs specified that they were unfamiliar with EPAs, whereas 117 (51.3%) indicated a basic understanding, 102 (44.7%) reported in-depth knowledge, and 6 (2.6%) considered themselves experts. Compared with the FPDs with a basic understanding, those with in-depth knowledge of EPAs gave slightly lower (−0.19) supervision ratings only on the "lead within the subspecialty profession" EPA ($P = .02$). No FPD was unfamiliar with the milestones; most indicated a basic (n = 83; 36.4%) or in-depth (n = 133; 58.3%) understanding, and a few self-reported as experts (n = 12; 5.3%). For each EPA, the FPDs with an in-depth knowledge of milestones gave slightly higher (0.19–0.25) supervision ratings than those with a basic familiarity ($P < .05$).

Discussion

We created scales to assess the level of supervision required for fellows for six of the seven common pediatric subspecialty EPAs. As recommended by Messick,[17] Downing,[18] and Cook and Beckman,[19] we obtained several sources of evidence to demonstrate the validity of these instruments.

Content

The scales were developed by medical educators with extensive experience in assessing fellows. Nearly all were or had been program directors, and several had advanced degrees in medical education.

Most published level of supervision scales[3,8,9,11–13] define the required level of supervision based on the proximity of the supervisor (i.e., physically present, available close by, or available at a distance), but none have been validated to date. Although we considered this model, the SPIN Steering Committee felt that the main factor used by subspecialty faculty to determine level of supervision of fellows was the complexity of the case. Fellows may require direct supervision for most patient encounters early in their training, but as their training progresses, supervision becomes guided by the complexity of the case. With more experience, fellows may act unsupervised for simple cases but may require more supervision for cases of greater complexity. Of note, incorporation of context complexity has been shown by others to improve entrustment scale reliability.[10] Some scales separate acting without supervision from supervising others. However, the committee thought that once fellows can act without supervision, the usual practice is for faculty to permit them to immediately provide supervision to others, making the distinction between these two levels artificial. Nonetheless, some fellows with limited supervisory experience may need to acquire additional skills before they are able to serve as supervisors.

Early in the process, we noted that a single scale could not be used for all of the EPAs. Even a model based on the complexity of the case was not applicable to all six EPAs; for example, such a scale would have no meaning in rating fellows for the "Leading within the subspecialty profession" EPA. Consequently, we created separate scales for each EPA (see Table 1). However, the three themes that emerged across these six scales will be useful when scales for the EPAs specific to each pediatric subspecialty are developed.

The scales we created are also different from those proposed in other specialties. This is related, at least in part, to variability of EPA content. Level of supervision scales have been published for assessing trainees in the operating room,[11,12] but most surgical EPAs are based on specific procedures in which the application of an entrustability scale may be easier.[15] In evaluating entrustment in internal medicine residents, Warm et al[13] created observable practice units, which are based on specific skills, because they thought that the associated internal medicine EPA was too broad to assess. Although the common pediatric subspecialty EPAs are likewise broad, including the functions needed to carry out the EPA on the data entry tool likely

provided guidance for assessors. That guidance, in addition to specificity in EPA-scale alignment, enabled us to address the broad nature of these EPAs. Hence, we believe that this provides substantial evidence to support the content of our scales.

Response process

We provided no faculty development to the FPDs or CCCs in how to use the scales or to explain the EPA concept. No questions were submitted to the coordinating center specifically about the scales, supporting the scales' clarity. Ratings across the EPAs by fellowship year differed, as would be expected based on the diversity of the EPAs and their representative functions, suggesting that raters did not choose a particular level of supervision based only on year of training.

Data for a large number of fellows were obtained in both reporting periods, and ratings were provided by a large number of programs across multiple institutions and subspecialties. Of note, the fall 2014 data collection occurred during the initial period in which the pediatric subspecialties were required to report milestones to the ACGME. It is noteworthy that a large number of the FPDs and CCCs voluntarily provided this information despite the added task and any challenges that may have been incurred in assigning a level of supervision for these six EPAs. This is consistent with the SPIN Steering Committee's goal of creating scales that were intuitive and consistent with how faculty supervise fellows in the workplace, providing very good validity evidence for the use of these scales.

Internal structure

Internal reliability was excellent in both the fall and the spring. Interrater reliability was very good for each EPA and across all of the EPAs. That this level of reliability was achieved without a program of faculty development

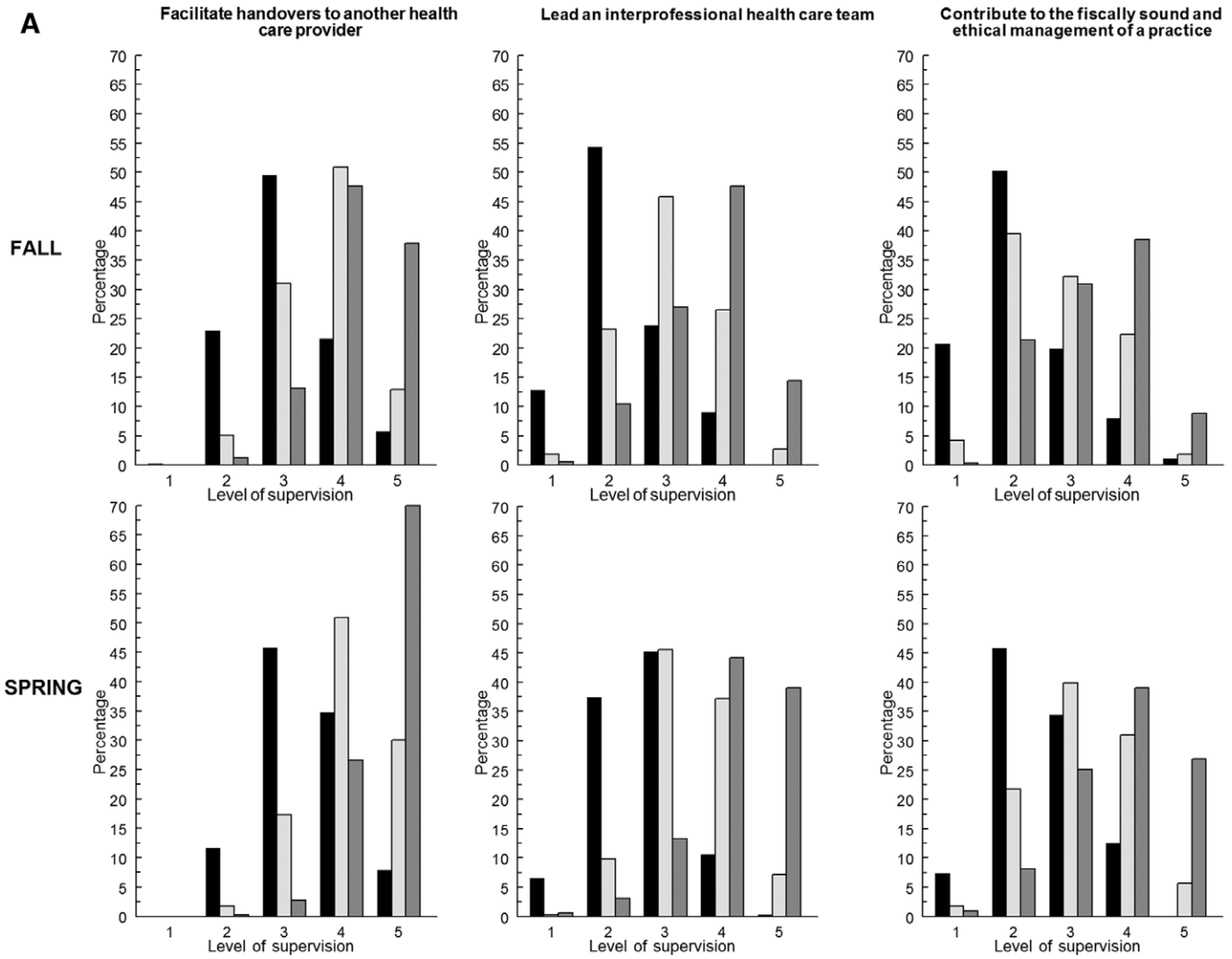

Figure 1 Distribution of the level of supervision assignments by Clinical Competency Committees for each of the six common pediatric subspecialty EPAs (three in panel A and three in panel B [next page]) using the scales (see Table 1) developed by the Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN). More than 200 programs in 14 pediatric subspecialties at approximately 80 institutions provided data for over 1,000 fellows, with more than 300 fellows per year of training assessed in each rating period (fall 2014 and spring 2015; see Table 2). In each panel, bars represent the percentage of ratings for each level of supervision by year of fellowship training (black bars = first year; light gray bars = second year; dark gray bars = third year). For each EPA, data from the spring are displayed below those from the fall. Ratings varied significantly across EPAs and in both fall and spring and within each year of training (all $P < .001$). *(Figure continues)*

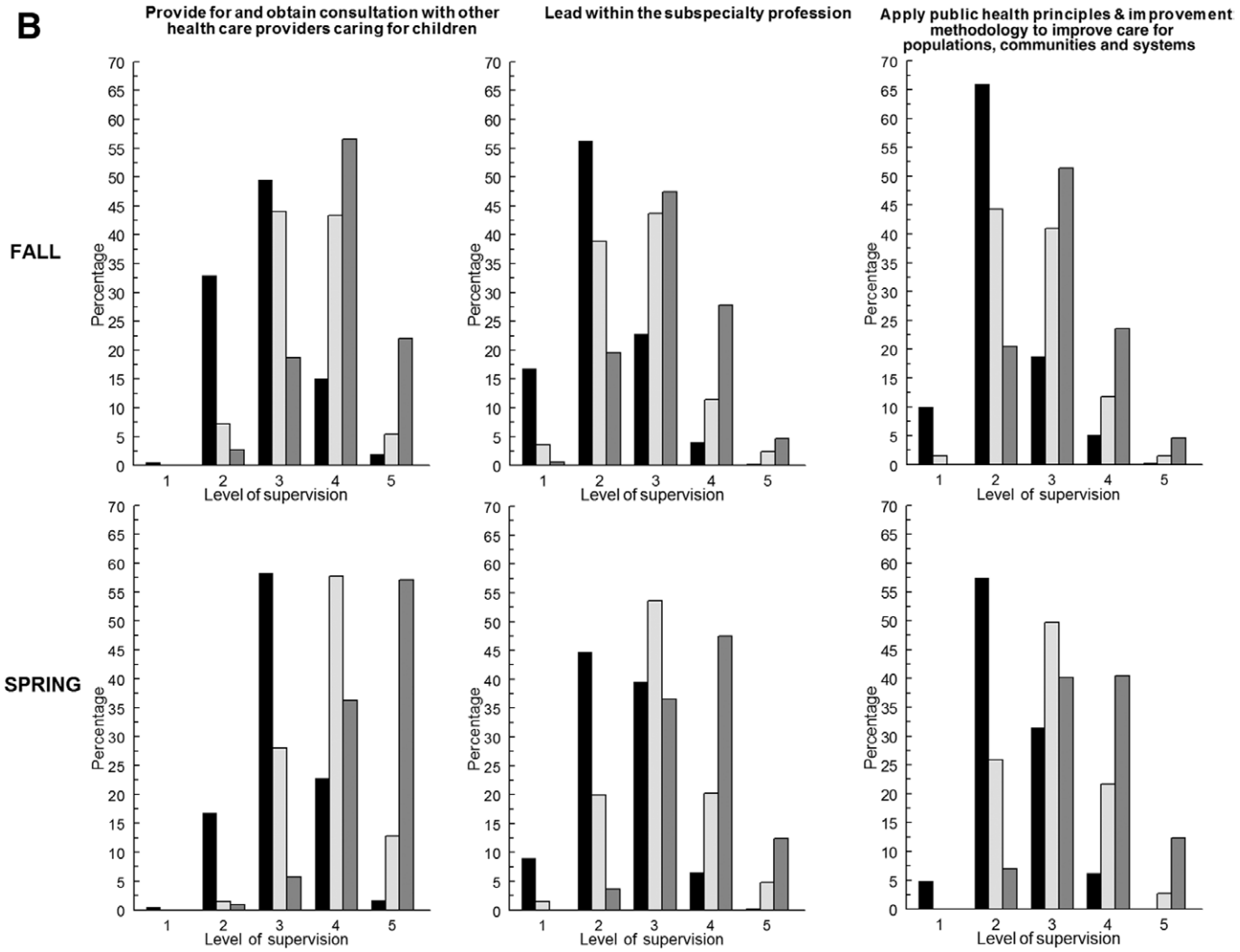

Figure 1 (Continued)

Table 3

Interrater Reliability Between FPD and CCC Ratings of Pediatric Fellows Using Level of Supervision Scales for Six Common Pediatric Subspecialty EPAs in the SPIN EPA Study, Fall 2014 (November–December) and Spring 2015 (May–June)[a]

	Spearman's ρ	
EPA	Fall 2014[b]	Spring 2015[b]
Facilitate handovers to another health care provider	0.73	0.76
Lead an interprofessional health care team	0.71	0.78
Contribute to the fiscally sound and ethical management of a practice	0.70	0.70
Provide for and obtain consultation with other health care providers caring for children	0.74	0.79
Lead within the subspecialty profession	0.68	0.67
Apply public health principles and improvement methodology to improve care for populations, communities, and systems	0.67	0.73

Abbreviations: FPD indicates fellowship program director; CCC, Clinical Competency Committee; EPAs, entrustable professional activities; SPIN, Subspecialty Pediatrics Investigator Network.

[a]The level of supervision scales (see Table 1) were developed by the SPIN Steering Committee to assess fellows on six common pediatric subspecialty EPAs.[7] More than 200 programs at approximately 80 institutions provided data for over 1,000 fellows, with more than 300 fellows per year of training assessed in each rating period (see Table 2).

[b]For all correlations, $P < .001$.

provides support for what Crossley et al[25] call "construct-aligned" scales—that is, scales aligning what faculty do in the clinical learning environment (supervise trainees providing care) with what they are asked to assess (trainees' needed level of supervision in care delivery). These reliability data add excellent support to our validity argument.

Relationship to other variables

As expected, there was a progressive decrease in the amount of supervision required, with first-year fellows requiring the most supervision and third-year fellows the least. This was observed for ratings by both the FPDs and CCCs. In addition, for each year of fellowship, ratings were higher in the spring compared with the fall, consistent with the expectation that fellows require less

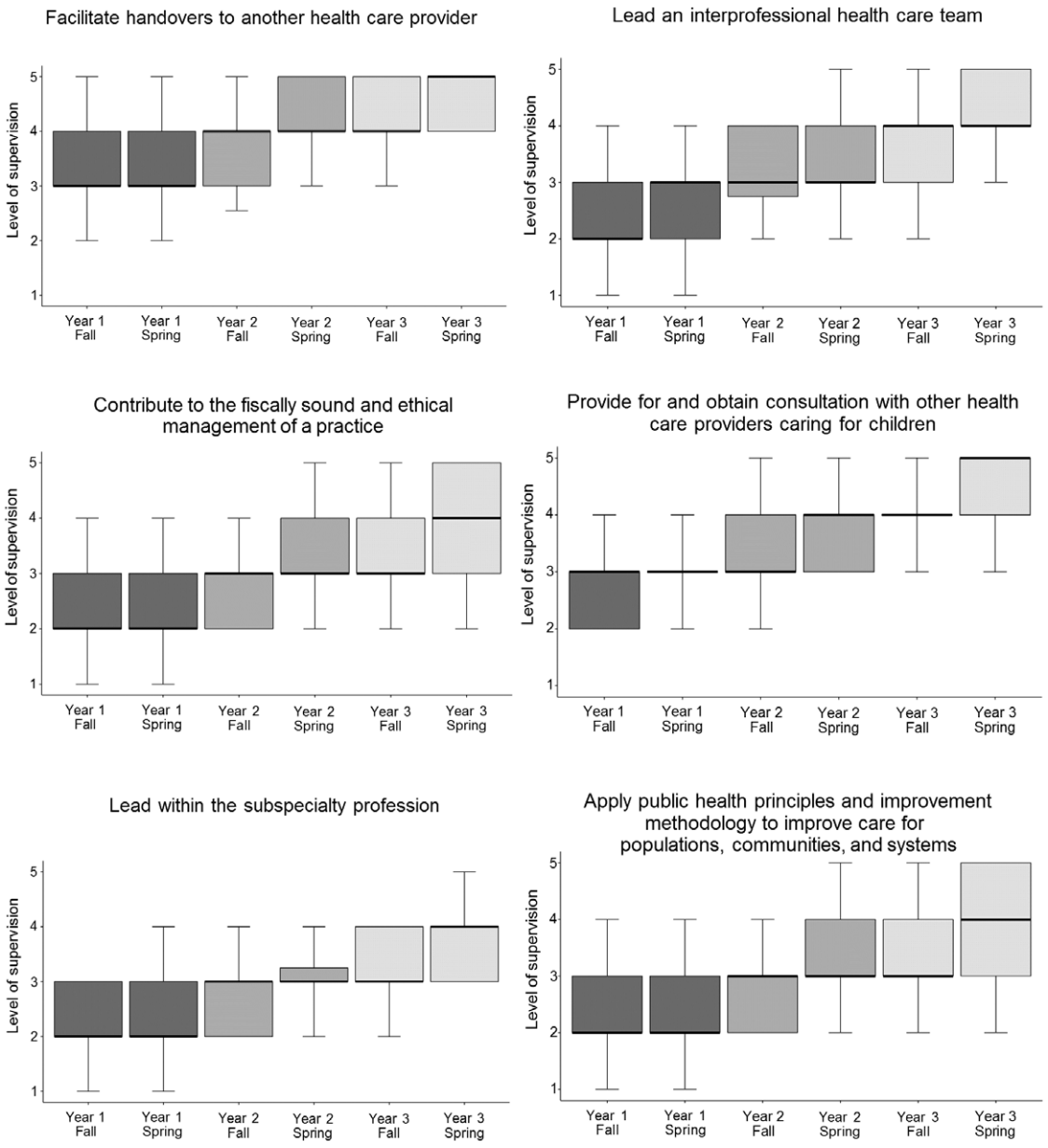

Figure 2 Progression of the level of supervision assignments by fellow year of training, as made by Clinical Competency Committees using the scales (see Table 1) developed by the Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN). More than 200 programs in 14 pediatric subspecialties at approximately 80 institutions provided data for over 1,000 fellows, with more than 300 fellows per year of training assessed in each rating period (fall 2014 and spring 2015; see Table 2). The dark line indicates the median, while the box and error bars show, respectively, the 25th and 75th and the 5th and 95th percentiles. For each period and for all EPAs, there was a progressive increase in the level of supervision rating ($P < .001$). Within each year of training, spring assignments were higher ($P < .001$) than those in the fall. Absence of the box and/or error bars indicates the same value for the percentiles.

supervision as their skills improve with more experience and teaching.

We examined other factors that could affect use of the scales. In large programs, fellows may have less interaction with many of the subspecialty faculty who serve on the CCC, but we observed no effect of program size on ratings. In addition, neither the experience of the FPD nor the participation of the FPD on the CCC influenced the assignments. Because the scales are specific to EPAs, a relatively new paradigm, we inquired about the FPDs' knowledge of EPAs. Although EPA understanding had a small effect on the mean rating, this was only for one EPA. However, those FPDs who self-reported a greater understanding of milestones rated fellows as requiring somewhat less supervision (i.e., higher ratings). Nevertheless, the importance of this finding is unclear because the scales were created to assess EPAs, not milestones. Overall, these data provide excellent evidence to support the validity of these EPA tools.

Consequences

While the primary purpose of this work was to build a case for validity evidence of the supervision scales, there are some implications that merit discussion. The added time needed to assess fellows with an additional tool did not impair the ability to attract and sustain

the participation of a large number of pediatric fellowship programs and to obtain a large data sample demonstrating the acceptability of the scales. Faculty discriminated ratings across EPAs, and the interrater reliability that was achieved without faculty development supports the intuitive nature of the scales.

Van der Vleuten's[26] assessment tool utility equation defines the utility of an assessment as a multiplicative function of reliability × validity × acceptability × cost × educational impact. This study provides evidence for three of these five variables: reliability, validity, and acceptability. We did not examine cost, although we anticipate that it would be low. The educational impact also needs to be determined, but our current findings suggest that a tool that is easily understood and completed by faculty will lead to more feedback to trainees to guide their development. In addition, the role for these assessments in making summative judgments will require additional study, particularly in determining whether performance in all of the EPAs should carry the same weight.

Limitations

There are several limitations of this effort. There is a potential for bias because the raters knew the fellows and their level of training. The FPDs and CCCs were purposely not blinded to previous ratings so as to allow the CCCs to function as they naturally would under nonstudy conditions. This may have led to anchoring bias to the subsequent level assignment. However, ratings were entered online, and unless the FPD and CCC made and kept printed copies of fellows' fall ratings, they would not have had access to them when making the level of supervision assignments in the spring. Participation in this investigation may have been influenced by factors such as obtaining academic credit and the opportunity to work with the ABP and national leaders in a medical education project. Also, we did not perform cognitive interviewing to determine how the raters made their determination, especially in relation to case complexity. In addition, for some EPAs (e.g., "Lead within the subspecialty profession"), the rating was likely not based on direct observation for many of the activities included in the EPA.

By design, the study population involved only pediatric fellows, limiting the generalizability to trainees in other disciplines in which the content of EPAs may differ. In addition, the scales we developed may be more useful in fellowships where there are more longitudinal experiences with a limited number of faculty compared with residency programs in which the number of evaluators is greater and the duration of observation is limited. Although most fellows were rated at two time points, a longer duration of assessment would provide additional validity evidence.

Conclusions

We developed level of supervision scales for six common pediatric subspecialty EPAs and presented strong validity evidence to support their use in EPA-based assessment of pediatric subspecialty fellows. The themes that emerged in the creation of these tools will be useful in the development of supervision scales for the pediatric subspecialty-specific EPAs and may inform the development of scales in other specialties and subspecialties.

Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN): All members of the SPIN Steering Committee are to be considered authors as each made substantial contributions to the design of the study and the acquisition of data, critically reviewed the manuscript, and approved its submission. *Adolescent Medicine:* Sarah Pitts; *Cardiology:* Gina Baffa; *Child Abuse:* Bruce Herman; *Critical Care Medicine:* David Turner; *Developmental and Behavioral Pediatrics:* Jill Fussell and Pam High; *Emergency Medicine:* Deborah Hsu; *Endocrinology:* Diane Stafford and Tandy Aye; *Gastroenterology:* Cary Sauer; *Hematology-Oncology:* Jennifer Kesselheim; *Infectious Diseases:* Angie Myers and Kammy McGann; *Neonatology:* Christiane Dammann and Patricia Chess; *Nephrology:* John Mahan; *Pulmonary Medicine:* Pnina Weiss; *Rheumatology:* Megan Curran; *Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network (APPD LEARN):* Alan Schwartz; *American Board of Pediatrics:* Carol Carraccio; *Association of Pediatric Program Directors Fellowship Committee:* Bruce Herman; *Council of Pediatric Subspecialties:* Richard Mink. For a full list of the SPIN collaborators, see Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/A463.

Acknowledgments: The SPIN Steering Committee sincerely thanks Alma Ramirez for her assistance with this project. The authors also appreciate the critical reviews of an early version of the scales by Robert Englander, MD, MPH, Joe Gilhooly, MD, and Chris Kennedy, MD.

Funding/Support: Financial support was provided by the American Board of Pediatrics Foundation.

Other disclosures: None reported.

Ethical approval: Institutional Review Board approval was obtained at each participating site and by the lead site, the University of Utah (#000765 on September 23, 2014).

Previous presentations: Results from this study were presented, in part, at the Association of Pediatric Program Directors 2016 Annual Spring Meeting, March 30–April 2, 2016, New Orleans, Louisiana, and at the 2016 Accreditation Council for Graduate Medical Education Annual Educational Conference, February 25–28, 2016, National Harbor, Maryland.

R.B. Mink is professor of pediatrics, David Geffen School of Medicine at UCLA, Los Angeles, California, and chief, Division of Pediatric Critical Care Medicine, and director, Pediatric Critical Care Medicine Fellowship, Harbor-UCLA Medical Center, Torrance, California.

A. Schwartz is Michael Reese Endowed Professor of Medical Education, associate head, Department of Medical Education, and research professor, Department of Pediatrics, University of Illinois at Chicago College of Medicine, Chicago, Illinois.

B.E. Herman is professor of pediatrics, vice chair for education, and residency program director, University of Utah School of Medicine, Salt Lake City, Utah.

D.A. Turner is associate professor of pediatrics, Duke University School of Medicine, and associate director of graduate medical education, Duke University Medical Center, Durham, North Carolina.

M.L. Curran is assistant professor of pediatrics and director, Pediatric Rheumatology Fellowship Program, Ann & Robert H. Lurie Children's Hospital of Chicago and Northwestern University Feinberg School of Medicine, Chicago, Illinois.

A. Myers is associate professor and director, Infectious Diseases Fellowship Program, Children's Mercy Hospital and University of Missouri–Kansas City School of Medicine, Kansas City, Missouri.

D.C. Hsu is associate professor of pediatrics, associate program director, Pediatric Residency Program, and clinical and education chief, Pediatric Emergency Medicine Section, Baylor College of Medicine/Texas Children's Hospital, Houston, Texas.

J.C. Kesselheim is assistant professor of pediatrics, Harvard Medical School, and associate fellowship program director for education, Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, Massachusetts.

C.L. Carraccio is vice president, Competency-Based Assessment, American Board of Pediatrics, Chapel Hill, North Carolina.

References

1 Swing SR, Beeson MS, Carraccio C, et al. Educational milestone development in the first 7 specialties to enter the next accreditation system. J Grad Med Educ. 2013;5:98–106.
2 ten Cate O, Scheele F. Competency-based postgraduate training: Can we bridge the gap

between theory and clinical practice? Acad Med. 2007;82:542–547.
3 ten Cate O. Nuts and bolts of entrustable professional activities. J Grad Med Educ. 2013;5:157–158.
4 ten Cate O, Hart D, Ankel F, et al. Entrustment decision making in clinical training. Acad Med. 2016;91:191–198.
5 Accreditation Council for Graduate Medical Education. Milestones. http://www.acgme.org/What-We-Do/Accreditation/Milestones/Overview. Accessed April 18, 2017.
6 ten Cate O, Snell L, Carraccio C. The interplay between individual ability and the health care environment. Med Teach. 2010;32:669–675.
7 American Board of Pediatrics. Entrustable professional activities for subspecialties. https://www.abp.org/subspecialty-epas. Accessed April 18, 2017.
8 Kennedy TJ, Lingard L, Baker GR, Kitchen L, Regehr G. Clinical oversight: Conceptualizing the relationship between supervision and safety. J Gen Intern Med. 2007;22:1080–1085.
9 Chen HC, McNamara M, Teherani A, Cate OT, O'Sullivan P. Developing entrustable professional activities for entry into clerkship. Acad Med. 2016;91:247–255.
10 Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. Br J Anaesth. 2014;112:1083–1091.
11 George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. J Surg Educ. 2014;71:e90–e96.
12 Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–1407.
13 Warm EJ, Held JD, Hellmann M, et al. Entrusting observable practice activities and milestones over the 36 months of an internal medicine residency. Acad Med. 2016;91:1398–1405.
14 Hauer KE, Ten Cate O, Boscardin C, Irby DM, Iobst W, O'Sullivan PS. Understanding trust as an essential element of trainee supervision and learning in the workplace. Adv Health Sci Educ Theory Pract. 2014;19:435–456.
15 Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: Outlining their usefulness for competency-based clinical assessment. Acad Med. 2016;91:186–190.
16 American Board of Pediatrics. Subspecialty certifications and admission requirements. https://www.abp.org/content/subspecialty-certifications-admission-requirements. Accessed April 18, 2017.
17 Messick S. Validity. In: Linn RL, ed. Educational Measurement. 3rd ed. New York, NY: American Council on Education and Macmillan; 1989:13–104.
18 Downing SM. Validity: On meaningful interpretation of assessment data. Med Educ. 2003;37:830–837.
19 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.
20 Mink RB, Carraccio CL, Schwartz A, et al; for SPIN. Creation of a pediatric subspecialty educational research network. Presented at: Association of Pediatric Program Directors 2016 Annual Spring Meeting; April 2016; New Orleans, LA. http://pedsubs.org/SPIN/PDFs/APPDNetworkdevelopment3-27-16FINAL.pdf. Accessed April 18, 2017.
21 Schwartz A, Young R, Hicks PJ; APPD LEARN. Medical education practice-based research networks: Facilitating collaborative research. Med Teach. 2016;38:64–74.
22 Janson H, Olsson U. A measure of agreement for interval or nominal multivariate observations. Educ Psychol Meas. 2001;61:277–289.
23 R Core Team. The R Project for Statistical Computing. https://www.R-project.org/. Accessed April 18, 2017.
24 Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48.
25 Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011;45:560–569.
26 Van Der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ Theory Pract. 1996;1:41–67.