Developing Validity Evidence for the Written Pediatric History and Physical Exam Evaluation Rubric

Marta A. King, MD, MEd; Carrie A. Phillipi, MD, PhD; Paula M. Buchanan, PhD; Linda O. Lewin, MD

From the Division of General Academic Pediatrics, Saint Louis University School of Medicine (Dr King), Center for Outcomes Research (Dr Buchanan), Saint Louis University, St Louis, Mo; Department of Pediatrics, Oregon Health & Science University (Dr Phillipi), Portland, OR; and Department of Pediatrics, University of Maryland School of Medicine (Dr Lewin), Bethesda, MD
The authors have no conflicts of interest to disclose.
Address correspondence to Marta A. King, MD, MEd, Division of General Academic Pediatrics, 1465 S Grand; Room 3711 North Tower, St Louis, MO 63104 (e-mail: making@slu.edu).
Received for publication January 17, 2016; accepted August 3, 2016.

ABSTRACT

OBJECTIVE: The written history and physical examination (H&P) is an underutilized source of medical trainee assessment. The authors describe development and validity evidence for the Pediatric History and Physical Exam Evaluation (P-HAPEE) rubric: a novel tool for evaluating written H&Ps.

METHODS: Using an iterative process, the authors drafted, revised, and implemented the 10-item rubric at 3 academic institutions in 2014. Eighteen attending physicians and 5 senior residents each scored 10 third-year medical student H&Ps. Inter-rater reliability (IRR) was determined using intraclass correlation coefficients. Cronbach α was used to report consistency and Spearman rank-order correlations to determine relationships between rubric items. Raters provided a global assessment, recorded time to review and score each H&P, and completed a rubric utility survey.

RESULTS: Overall intraclass correlation was 0.85, indicating adequate IRR. Global assessment IRR was 0.89. IRR for low- and high-quality H&Ps was significantly greater than for medium-quality ones but did not differ on the basis of rater category (attending physician vs. senior resident), note format (electronic health record vs nonelectronic), or student diagnostic accuracy. Cronbach α was 0.93. The highest correlation between an individual item and total score was for assessments was 0.84; the highest interitem correlation was between assessment and differential diagnosis (0.78). Mean time to review and score an H&P was 16.3 minutes; residents took significantly longer than attending physicians. All raters described rubric utility as "good" or "very good" and endorsed continued use.

CONCLUSIONS: The P-HAPEE rubric offers a novel, practical, reliable, and valid method for supervising physicians to assess pediatric written H&Ps.

KEYWORDS: assessment; clinical documentation; diagnostic reasoning; history; medical student; physical examination; undergraduate medical education

ACADEMIC PEDIATRICS 2017;17:68–73

WHAT'S NEW

The Pediatric History and Physical Exam Evaluation rubric content, internal structure, and response process validity are supported by a multicenter study. The rubric offers a novel, practical, reliable, and valid method for supervising physicians to assess medical student documentation.

OBTAINING A COMPREHENSIVE history, performing a complete physical examination, and documenting a clinical encounter are critical physician skills. The written history and physical examination (H&P) reflects the author's competency in many Accreditation Council on Graduate Medical Education domains: patient care, medical knowledge, interpersonal and communication skills, professionalism, and at times, systems-based practice and practice-based improvement.[1] Defined Pediatric Core Entrustable Professional Activities reflected in the written H&P include: gathering a history and performing a physical examination, prioritizing a differential, recommending and interpreting common diagnostic and screening tests, and documenting a clinical encounter.[2] The written H&P is a rich source of trainee assessment data.

Despite the importance placed on documenting a clinical encounter[1–6] and the large number of written notes typically required of trainees,[6,7] structured formative or summative evaluation is rare.[7] A 2010 survey showed that only 23% of internal medicine clerkships used a structured evaluation form and that only 16% incorporated written note evaluation in grade determination.[7] Although most Canadian medical students described note-writing as a valuable educational activity and reported significant effect of note feedback on subsequent documentation, they expressed a belief that written documentation assessment

did not reflect clinical abilities.[6] The authors linked this discrepancy to student and program director concerns regarding lack of standardized evaluation criteria, considerable inter-rater variability, and feedback delays. Unreliable assessment could result in poor understanding of expectations and diminished value of the written H&P as a teaching and evaluation tool.

Limited studies address clinical documentation evaluation and highlight the lack of validated scoring methods. Internal medicine clerkship H&P scoring tools were characterized by lack of specific criteria required for a particular score which contributed to poor inter-rater reliability (IRR).[8,9] Recently, Baker et al summarized a series of studies supporting instrument validity for scoring internal medicine student H&Ps, reporting a fair to moderate IRR ($\kappa = 0.02–0.56$) with 22 trained attending raters from 2 medical schools.[8] In the only pediatric study published to date, 2 raters from a single institution achieved a modest IRR ($\kappa = 0.58$) when using a 38-item tool.[9] The need for reliable, practical assessment tools for clinical documentation in undergraduate medical education is clear.

Despite the widespread adaptation of the electronic health record (EHR), previous studies did not take note format (EHR- vs non–EHR-generated) into account.[8–11] They also did not consider the role of the senior residents, rater perceptions, or the time required for scoring.[8–11] Our goal was to develop and gather validity evidence for a practical Pediatric History and Physical Exam Evaluation (P-HAPEE) rubric that could be readily used by attending physicians and senior residents to reliably assess EHR and non–EHR-generated documentation. We hypothesized that this tool, when used as described, will demonstrate strong content validity, IRR, and internal consistency.

METHODS

RUBRIC DEVELOPMENT AND CONTENT VALIDITY

We based the content of the P-HAPEE rubric on competencies outlined by the Liaison Committee on Medical Education, the Association of American Medical Colleges, the Accreditation Council on Graduate Medical Education, and core pediatric hospital medicine domains.[1,2,4,5,12] We reviewed previously published written H&P evaluation tools,[8–11,13] a consultation letter evaluation tool,[14,15] oral case presentation evaluation tool,[16] as well as written H&P instructions and sample H&Ps solicited from a number of medical schools. We drew on medical education frameworks related to diagnostic reasoning and clinical competency[17–19] in drafting behavioral anchors.

More than 50 local and national medical educators and clinicians critically reviewed the P-HAPEE rubric. Additionally, 15 undergraduate and graduate medical educators reviewed and scored a sample H&P and discussed the rubric during 1 of 2 hour-long conference call focus groups held in the fall of 2014 and facilitated by one of the investigators (C.A.P.). We asked reviewers to comment on appropriateness of rubric sections, items, order, scale, and anchors with suggestions for additions, deletions, clarification, and/or modifications.

To make the items as clear as possible and to allow some variation in note format across institutions, we chose the broadest terms for each item. As an example, we changed the heading of "problem list" to "problem identification" to allow student credit for either an enumerated list or for identifying problems in a paragraph discussion. Similarly, we chose the heading of "assessment" over "problem representation" or "summary statement" because the latter terms were unfamiliar to many clinicians. "Assessment" would also allow raters to give credit to students who did not create a concise "1-liner," but included all of the requisite critical points in a more prolonged assessment discussion. Before the study 2 of the authors (M.A.K., C.A.P.) pilot tested the rubric in our pediatric clerkships, and modified the items as well as the behavioral anchors iteratively on the basis of their experiences as well as the feedback from students and faculty members, individual expert reviewers, and conference call attendees.

P-HAPEE RUBRIC

The 10-item P-HAPEE rubric (Appendix) includes sections for information-gathering (history, physical examination/diagnostic studies), as well as information synthesis and clinical reasoning. It uses a 5-point rating scale with behavioral anchors to facilitate scoring and provide learners with specific goals for documentation improvement. The rubric is criterion-referenced with a score of "5" indicating an ideal H&P for a given patient. Additional items include global assessment of H&P quality (below, meets, exceeds expectations), and space for narrative feedback to describe what the author did well and specific suggestions for improvement.

H&P SELECTION

Investigators at the Saint Louis University, Oregon Health & Science University (OHSU), and University of Maryland (UMD) schools of medicine identified potential H&Ps for inclusion in the study from a pool of third-year medical student general pediatric and subspecialty floor H&Ps collected throughout the academic year as part of the pediatric clerkship at each institution. Saint Louis University students submitted EHR (Epic, Verona, WI)-generated H&Ps to the patients' attending of record. Although OHSU students used EHR (Epic) for medical documentation, they turned in separate H&Ps generated on a securely encrypted word processing device to a designated teaching attending. The H&Ps often contained sections initially generated in an EHR and then copied and edited in a word processing program. UMD students used a word processing program for medical documentation and submitted H&Ps to a designated teaching attending.

The selected 30 H&Ps were authored by different students from our 3 institutions and specifically chosen to reflect a variety of patient ages, diagnoses, number of hospital problems, formats (EHR and non–EHR-generated), as well as student diagnostic accuracy and documentation quality. We relied on the information included in the student H&P in formulating the list of identified problems