Adv in Health Sci Educ (2016) 21:761–773
DOI 10.1007/s10459-015-9664-3

Inter-rater reliability and generalizability of patient note scores using a scoring rubric based on the USMLE Step-2 CS format

Yoon Soo Park[1] · Abbas Hyderi[2] · Georges Bordage[1] · Kuan Xing[1] · Rachel Yudkowsky[1]

Received: 1 July 2015 / Accepted: 22 December 2015 / Published online: 12 January 2016

Abstract Recent changes to the patient note (PN) format of the United States Medical Licensing Examination have challenged medical schools to improve the instruction and assessment of students taking the Step-2 clinical skills examination. The purpose of this study was to gather validity evidence regarding response process and internal structure, focusing on inter-rater reliability and generalizability, to determine whether a locally-developed PN scoring rubric and scoring guidelines could yield reproducible PN scores. A randomly selected subsample of historical data (post-encounter PN from 55 of 177 medical students) was rescored by six trained faculty raters in November–December 2014. Inter-rater reliability (% exact agreement and kappa) was calculated for five standardized patient cases administered in a local graduation competency examination. Generalizability studies were conducted to examine the overall reliability. Qualitative data were collected through surveys and a rater-debriefing meeting. The overall inter-rater reliability (weighted kappa) was .79 (Documentation = .63, Differential Diagnosis = .90, Justification = .48, and Workup = .54). The majority of score variance was due to case specificity (13 %) and case-task specificity (31 %), indicating differences in student performance by case and by case-task interactions. Variance associated with raters and its interactions were modest (<5 %). Raters felt that justification was the most difficult task to score and that having case and level-specific scoring guidelines during training was most helpful for calibration. The overall inter-rater reliability indicates high level of confidence in the consistency of note scores. Designs for scoring notes may optimize reliability by balancing the number of raters and cases.

Keywords Patient note · USMLE Step-2 CS · Validity · Rater effects

✉ Yoon Soo Park
yspark2@uic.edu

[1] Department of Medical Education (MC 591), College of Medicine, University of Illinois at Chicago, 808 South Wood Street, 963 CMET, Chicago, IL 60612-7309, USA

[2] Department of Family Medicine (MC 785), College of Medicine, University of Illinois at Chicago, 1819 West Polk Street, 150 CMW, Chicago, IL 60612-7309, USA

Introduction

The United States Medical Licensing Examination (USMLE) is a three-step examination toward medical licensure in the United States. Step-1 is a multiple choice question (MCQ) test of basic sciences knowledge; Step-2 Clinical Knowledge (Step-2 CK) is an MCQ test of clinical knowledge; Step-2 Clinical Skills (Step-2 CS) is a test of clinical skills based on standardized patient (SP) encounters; and Step-3 is an MCQ and computer simulation-based test of applied medical knowledge and clinical science needed for unsupervised practice. Students generally take Step-1 during the second year of medical school and the Step-2 exams early in their fourth year; the Step-3 exam is typically taken after the first year of residency. All steps must be passed for licensure (United States Medical Licensing Examination 2015a, b; National Board of Medical Examiners 2015).

The Step-2 CS includes a patient note (PN) completed after each SP encounter. The PN is scored by trained physician raters who provide global ratings of (1) documented findings from the patient encounter (history and physical examination), (2) diagnostic impressions, (3) justification of potential diagnoses, and (4) initial diagnostic studies (Federation of the State Medical Boards and National Board of Medical Examiners 2015). Since 2012, the PN format has required examinees to rank order the differential diagnosis and to *justify* their differential diagnoses by listing supporting findings from their SP encounters. Prior to 2012, the PN format included diagnostic impressions, but did not require rank-ordering or justifying the differential (Haist et al. 2013). Details of the scoring rubric used in the Step-2 CS have not been disclosed, leaving institutions to locally develop and gather validity evidence for their PN rubrics (Park et al. 2013; Yudkowsky et al. 2015; Southern Illinois University 2015).

The USMLE Step-2 CS PN scores are combined with the SPs' physical examination scores to form an Integrated Clinical Encounter (ICE) subcomponent score that measures data gathering and interpretation skills. To pass the Step-2 CS, examinees must pass the ICE (Federation of the State Medical Boards and National Board of Medical Examiners 2015; Haist et al. 2013). After implementing the new PN format in 2012, the overall first-attempt ICE pass rate decreased by about 2.3 % (for 30,752 examinees) during the 2013–2014 testing cycle and decreased by about 5.0 % among non-US or Canadian school test takers, making this decline the largest drop in ICE subcomponent pass rate in the past five testing cycles (United States Medical Licensing Examination 2015a, b; National Board of Medical Examiners 2015).

The revised PN format of the USMLE Step-2 CS has prompted medical schools to re-align their curriculum and assessment methods to better prepare their students for this national examination (Gilliland et al. 2008; Park et al. 2013). Imbedded within this alignment is the challenge to develop scoring rubrics and train faculty raters to score PNs. An emerging question on the validity of PN scores is the influence that raters have on the reproducibility of scores. Prior studies by Boulet et al. (1998, 2004) found that while the selection of raters matters, the overall measurement error due to raters and rater interactions is minimal (Whelan 1999). On the other hand, Clauser et al. (2008) found that raters made a greater contribution to measurement error than case specificity (differences in learner performance by case), indicating that double scoring of notes increased the precision of PN scores, equivalent to increasing the test length by 50 %. While the two sets of studies provide conflicting implications on how raters contribute to the quality of scores, they were based on data prior to changes in the note format and do not provide clear implications on how raters score the justification of differential diagnoses. In a more recent

study by Williams et al. (2014), psychometric characteristics of the new PN format were examined to identify variability in medical students' diagnostic justifications. They found that a majority of score variance was due to case specificity; however, their generalizability study did not include raters as a facet, making variability due to raters potentially confounded in the results. Therefore, examining whether raters can be trained to score the new PN format that includes diagnostic justification is an important aspect of validity evidence that needs further investigation. Moreover, identifying practical guidelines for rater training and scoring of the note would provide meaningful information for local medical schools assessing students using the new PN format.

Following the introduction of the new PN format by the USMLE, the University of Illinois at Chicago College of Medicine (UIC COM) developed a scoring rubric to assess students' performance in data gathering and clinical reasoning as part of a graduation competency examination (GCE; Park et al. 2013). The PN scoring rubric aimed to facilitate standardization of scoring across raters and to provide feedback for students preparing for the Step-2 CS exam. Prior studies on the validity evidence of PN scores using this rubric showed moderate reliability, indicating the ability to discriminate differences in learner performance; in addition, PN scores were correlated with SP documentation scores (Park et al. 2013). Among tasks in the rubric, justification had the highest item discrimination (Yudkowsky et al. 2015). Overall, results of the original scoring rubric have shown promise when used for preparing students to write PNs using the new Step-2 CS PN format. However, ongoing feedback from faculty raters indicated that a focused analysis of rater-related issues was needed to better understand the influence of raters and their overall effect on validity.

The purpose of the present study was to gather validity evidence regarding response process (scoring patterns and consistency in raters' scoring processes) and internal structure (consistency between tasks, variability of scores, and reliability), and to examine rater-related issues that may impact the overall reproducibility of PN scores. The additional validity evidence presented in this study provides a comprehensive perspective from Messick's (1995) unified validity framework as operationalized by the *Standards for Educational and Psychological Testing* (AERA et al. 2014; Downing 2003) and highlights issues to be considered in the preparation and administration of local clinical skills examinations that include a PN section. Guidelines for rater training and calibration, including potential curricular impact are discussed.

Methods

Revised patient note scoring rubric

The UIC COM PN Scoring Rubric was developed using guidelines from the Step-2 CS manual and input from the GCE committee of UIC COM faculty to provide detailed criteria (1) to help standardize scoring of PNs and (2) to provide students and faculty with specific feedback regarding areas of deficiency. The original UIC COM PN Scoring Rubric (Park et al. 2013) was revised based on empirical evidence from a prior validity study, including qualitative feedback from faculty raters who noted difficulty in interpreting and judging specific tasks in the rubric. Table 1 contains the revised UIC COM PN Scoring Rubric used in this study, categorized into four tasks: Documentation (of findings in history and physical examination; 30 points maximum), Differential diagnosis (DDX; 30 points

Table 1 Revised Patient Note Scoring Rubric from the University of Illinois at Chicago College of Medicine Graduation Competency Examination (GCE)[a] to assess patient notes written using the USMLE Step-2 Clinical Skills patient note format

Task with description (maximum points)	Score and Anchor[b]
1. *Documentation*: Documentation of findings in history (Hx) and physical examination (PE) (30 points)	1. Key Hx and PE findings are missing or incorrect 2. Some key positive findings present but poorly documented or disorganized or missing pertinent negatives 3. Most key positive findings well documented and organized, may miss a few pertinent negatives 4. All key information present, concise and well organized with little irrelevant information
2. *DDX*: Differential diagnosis (30 points)	1. [0–1 of 3] or [0 of 2] of the correct diagnoses listed 2. [2 of 3] or [1 of 2] of the correct diagnoses listed, in any order 3. All diagnoses listed, incorrect rank order 4. All diagnoses listed and correctly rank ordered
3. *Justification*: Justification of differential diagnosis (30 points)	1. No justification provided OR many missing or incorrect links between findings and Dx 2. Some missing or incorrect links between findings and Dx 3. Only a few missing or incorrect attributions, which would not impact Dx 4. Links to diagnoses are correct and complete
4. *Workup*: Plan for immediate diagnostic workup (10 points)	1. Diagnostic workup places patient in unnecessary risk or danger 2. Ineffective plan for diagnostic workup—essential tests missed, irrelevant tests included 3. Reasonable plan for diagnostic workup, may have some unnecessary tests or missing few essential tests 4. Plan for diagnostic workup is effective and efficient, includes all essential tests, and few or no unnecessary tests

[a] Scoring rubric used in this study to conduct the inter-rater reliability analysis

[b] Each score level is worth 25 % of the maximum points for each task: Documentation: "1" = 7 points, "2" = 15 points, "3" = 23 points, "4" = 30 points; DDX: "1" = 7 points, "2" = 15 points, "3" = 23 points, "4" = 30 points; Justification: "1" = 7 points, "2" = 15 points, "3" = 23 points, "4" = 30 points; Workup: "1" = 2 points, "2" = 5 points, "3" = 8 points, "4" = 10 points

maximum), Justification (of differential diagnosis; 30 points maximum), and Workup (plan for immediate diagnostic workup; 10 points maximum). Each task is scored on a 4-point scale, each anchored to a descriptor on the quality of the note. Each score level is worth 25 % of the total possible points for each task; for example, a student with a score of "4" for documentation (30 points), "3" for DDX (23 points), "2" for Justification (15 points), and "2" for Workup (5 points) will get a total of 73 points (= 30 + 23 + 15 + 5 points). The scores ranged from 23 (i.e., getting "1" for all four tasks) to 100 (i.e., getting "4" for all four tasks) points. The rationale for having the lowest possible score as 23 points rather than 0 points was to signal students that their PNs had been submitted, reviewed, and scored; this decision was based on faculty committee discussion and consensus. Both students and faculty members were informed of this range when interpreting their scores.

The main change in the revised scoring rubric was the expansion of the single "justification of differential diagnosis" task in the previous rubric to the current (1) DDX and (2) Justification tasks to clearly distinguish the two tasks and to better identify gaps in students' clinical reasoning (Cianciolo et al. 2013; Williams et al. 2014). In addition, descriptors in the anchors for Documentation were revised based on rater feedback.

Data collection

Graduation competency examination

Primary data were collected from 177 graduating UIC COM medical students in May 2014, as part of the UIC COM GCE. Medical students encountered five SP cases (dizziness, shortness of breath, abdominal pain, weight loss, and chest pain), each lasting 15 min. Following each encounter, students had up to 10 min to complete the USMLE Step-2 CS PN template at a computer terminal; this is the same time allotted by the USMLE. A total of 11 trained faculty raters scored the PNs using the revised UIC COM PN Scoring Rubric (see Table 1); the notes of each case were divided among 2 or 3 raters, each of whom scored notes for only that case. Faculty participation in scoring was voluntary and no incentives were provided. Each note was scored independently by only one rater. Scores were submitted through an online scoring portal, and raters scored the PNs from separate locations on their own schedule. Scoring was completed between June and July 2014.

Rescoring of patient notes

Calculating inter-rater reliability requires each PN to be scored by two or more raters. Fifty-five medical students were randomly selected from the May 2014 GCE participants so as to maintain the original score distribution of the 177 students. *T*-tests were conducted to check that the mean and variance of the original distribution was maintained; tests for equality of distributions were conducted using the two-sample Kolmogorov–Smirnov test. Six raters from the initial scoring in June-July 2014 participated in rescoring; all five PN cases of the subsample of 55 students were rescored, for a total of 275 PNs.

Each rater was assigned specific notes to score, ensuring that a rater who scored the original PN did not rescore the same PN. A spiral design was used, where each rater was sequentially assigned a PN to score, after randomly ordering the sequence of PNs; the spiral design ensures that each rater has a balanced number of notes to score from different students and cases (Hombo and Donoghue 2001). The subsample size of 55 was selected to optimize the spiral design, considering the number of available raters, number of cases, and time available, in addition to having sufficient power to calculate inter-rater reliability indices and conduct other psychometric analyses. Raters were blinded to students and their prior performance in the June-July 2014 scoring. Similar to scoring during the GCE, raters rescored the notes independently online from separate locations at their own time. Rescoring was conducted between November and December 2014. Each rater was assigned about 46 PNs.

Rater training

Each rater was provided the following case materials prior to attending rater training: (1) detailed case information (patient information, summary of case, history and physical

examination checklists used by the SP to score the encounter), (2) "gold standard" exemplar note with key positive and negative findings highlighted and case-specific scoring guidelines on factors discriminating between score levels, (3) three sample notes (actual student notes) of varying PN quality (low, moderate, and high) for raters to practice scoring prior to training, (4) ten extra sample notes as optional cases to review, for raters who wanted to practice reviewing more notes, and (5) the revised UIC COM PN Scoring Rubric (Table 1).

During training, a moderator (one of the faculty raters) presented a summary of the case and facilitated a discussion of the three sample notes per case for calibration. Discrepancies in scoring were discussed until consensus was reached. The discussion of case- and level-specific guidelines was recorded in audio format and transcribed for distribution; discussions to clarify scoring guidelines and "gold standard" exemplar notes were made during the meeting. Two sessions totaling 3 h of training were provided to train all raters on all five cases.

Survey and debriefing meeting

A debriefing meeting was conducted in December 2014 to gather rater feedback on the scoring experience. An online survey was administered between December 2014 and January 2015 to identify specific tasks or cases that were more difficult to score, an estimate of the time to score each note, and number of PNs scored in one sitting.

Analysis

Descriptive statistics were used to examine distributional differences in PN scores between the primary (June–July 2014) and rescored data (November–December 2014). Measures of inter-rater reliability were calculated to compare ratings between the primary and rescored data: (1) % exact agreement (proportion of exact agreement between the primary and rescored data), (2) kappa (agreement between raters, taking into account chance agreement), and (3) quadratically-weighted kappa (extension of kappa using weights, where larger differences between raters would result in lower agreement). Measures of inter-rater agreement were calculated for pass-fail decisions, applying a 50 % passing standard based on a prior Angoff standard setting exercise.

A fully-crossed generalizability study (G-study) was conducted, with person (p) × cases (c) × raters (r) × task (t), using unweighted scores from each task (Brennan 2001). Scores were obtained from both primary (GCE) and rescored data for the subset of 55 students. Cases and raters were assumed to be random, sampled from a population (universe) of potential cases and raters. Tasks were assumed to be fixed as the finite set of items measured. A decision study (D-study) was conducted to project reliability when altering conditions in the number of raters and cases. Qualitative data from the survey and debriefing meeting were transcribed and analyzed for themes; member checking was used to confirm themes identified. The institutional review board of the University of Illinois at Chicago approved this study.

Results

Descriptive statistics

The overall composite score mean for the primary and rescored data were 64.76 (SD = 9.11) and 61.51 (SD = 10.46). Correlations between primary and rescored

Table 2 Percentage Scores by Task (Row %) and Composite Score[a] (Mean, SD) across cases ($n = 55$ notes scored for each case, for a total of 275 notes)

Task	*Primary*: row % (June–July 2014)				*Rescored*: row % (November–December 2014)				p value[b]
	"1"	"2"	"3"	"4"	"1"	"2"	"3"	"4"	
Documentation	25	19	41	14	33	25	29	13	.260
DDX	18	45	13	25	19	46	14	21	.937
Justification	4	31	43	21	9	32	42	18	.540
Workup	9	30	48	13	9	33	43	14	.930
Overall	14	31	36	18	17	34	32	17	.876
Composite Score	Mean = 64.76 (SD = 9.11)				Mean = 61.51 (SD = 10.46)				.085

[a] Composite score is the sum of the scores from each task using weights associated with each score; "1", "2", "3", and "4" correspond to the score levels in the scoring rubric (see Table 1)

[b] *T*-tests conducted for comparisons between composite scores; χ^2 tests conducted for comparisons between proportions of score levels

composite scores were .87 for the overall scores, $p < .001$ (correlations ranged between .70 and .87 by case). Table 2 contains descriptive statistics comparing the primary and rescored data, stratified by score levels by task and composite score. Comparisons between primary and rescored data indicate similar proportions of score categories assigned between the two scoring occasions, both overall and at the task levels; there were no significant differences in mean composite scores between primary and rescored data.

Inter-rater reliability

The overall % exact agreement for the unweighted score levels was 60 % (Kappa = .44, Weighted Kappa = .67). For the composite score, weighted kappa was .79. Weighted kappa can be interpreted as an intraclass correlation as they are mathematically equivalent (Fleiss and Cohen 1973). Justification had the lowest inter-rater reliability with % exact agreement of 46 % (Kappa = .20, Weighted Kappa = .48). See Table 3 for specific task-level details on % exact agreement, kappa, and weighted kappa. Across cases, the weighted kappa for task-level overall agreement ranged between .55 and .76; the weighted kappa for the composite score ranged between .62 and .87.

For pass-fail decisions, the overall % exact agreement was 90 % (Kappa = .61). Because pass-fail decisions are dichotomous, weighted kappa was not calculated. Although Justification had low inter-rater reliability for pass-fail decisions, the % exact agreement ranged between 78 and 100 % (Kappa = .14–1.00). The lowest agreement for pass-fail decision was Documentation, which had an overall 78 % exact agreement, resulting from the low agreement in two cases.

Generalizability study

Variance components

Using the fully-crossed G-study design, person variance accounted for 6.7 % of the total variance. Rater variance accounted for less than 1 %; variance due to rater interactions was

Table 3 Inter-Rater Reliability: % Exact, Kappa, and Quadratic-Weighted Kappa by tasks

Task	All four levels			Pass–fail decision	
	% Exact[a]	Kappa[b]	Wtd. Kappa[c]	% Exact[a]	Kappa[b]
Documentation	51	.33 (.04)	.63 (.06)	78	.48 (.06)
DDX	88	.83 (.04)	.90 (.06)	96	.85 (.06)
Justification	46	.20 (.04)	.48 (.06)	93	.42 (.06)
Workup	53	.30 (.04)	.54 (.06)	91	.48 (.06)
Overall	60	.44 (.02)	.67 (.03)	90	.61 (.03)
Composite Score			.79 (.01)		

[a] " % Exact" indicates proportion of exact agreement between raters

[b] "Kappa (unweighted)" measures agreement between raters, taking into account chance agreement. Values in parenthesis represent standard errors

[c] "Weighted Kappa" uses quadratic weights to penalize larger differences between raters and is equivalent to intraclass correlation. Values in parenthesis represent standard errors

also less than 5 %, indicating that variability due to raters was modest. However, the person-by-case interaction accounted for 13 %, and person-by-case-by-task interaction accounted for 31 % of the total variance, indicating that students performed differently between cases and between tasks from different cases. See Table 4 for details. The fully-crossed G-study design makes the assumption that the two raters scoring the note are interchangeable; however, this may not fully capture the unbalanced nature of the rater assignment during primary scoring (some raters scored more notes than other raters). A reanalysis of the data using an unbalanced random-effects G-study design (Brennan 2001) for comparison indicated similar results, further confirming our findings (rater variance <1 %, person-by-case interaction 12 %, and person-by-case-task interaction 29 %).

Reliability

The generalizability coefficient (G-coefficient) and Φ-coefficient based on the five cases were .59 and .50, respectively; the G-coefficient is used as a reliability index for making normative decisions, while the Φ-coefficient is used for making criterion-based decisions. For pass-fail decisions, the G-coefficient and the Φ-coefficient were .50 and .37, respectively; the lower reliability indices for pass-fail decisions may be explained by the lower variability in score distribution, relative to the score distribution for four score levels.

Projections in reliability and measurement precision

Projections in reliability (D-studies) indicate that double scoring increases reliability by .05 points. However, beyond two raters, the marginal gain was minimal; see Fig. 1. Using the G-coefficient, a reliability of .70 can be reached with 1 rater when 10 cases are used; when PNs are double scored, only 8 cases are needed. Projections in reliability can be translated into the amount of score precision gained using the standard error of measurement (SEM). For example, if a test administrator wants to reduce measurement error by over ±1.75 % points from 6.00 % using a single rater (resulting in increased precision of 3.5 % overall), the number of cases will need to be doubled from 5 to 10 cases. However, using two raters, the same increase in precision can be obtained by adding 3 more cases. Overall, double-scoring all PNs can increase measurement precision equivalent to adding two cases.

Table 4 Variance components of the patient note scoring rubric: generalizability study[a]

Effect[b]	*df*	Variance components	% Variance components
p	54	.063 (.023)	6.7
c	4	.079 (.057)	8.4
r	1	.003 (.005)	.3
t	3	.028 (.024)	3.0
$p \times c$	216	.120 (.023)	12.7
$p \times r$	54	.000 (.004)	.0
$p \times t$	162	.020 (.013)	2.1
$c \times r$	4	.008 (.011)	.8
$c \times t$	12	.031 (.022)	3.3
$r \times t$	3	.000 (.004)	.0
$p \times c \times r$	216	.023 (.008)	2.5
$p \times c \times t$	648	.290 (.023)	30.9
$p \times r \times t$	162	.008 (.007)	.9
$c \times r \times t$	12	.035 (.015)	3.7
$p \times c \times r \times t$, error	648	.231 (.013)	24.6

[a] All facets were considered random; only tasks were fixed. For five cases, the G-coefficient = .59 and Φ-coefficient = .50. For pass-fail decisions based on the five cases, the G-coefficient = .50 and the Φ-coefficient = .37

[b] Values in parenthesis represent standard errors. G-study design used person (*p*) × case (*c*) × rater (*r*) × task (*t*)

Survey and debriefing results

Faculty raters felt they were able to clearly identify excellent and failing students; however, they had difficulty discriminating between level 2 and level 3 within a given task. Within the Documentation section, raters wanted more specific guidance on key findings that warrant higher priority and what frequency determined few, some, and most.

Raters unanimously felt the Justification task was the most difficult to score given the range of nuances possible. Ambiguous situations included notes in which findings listed in the Justification section were not documented in the H&P, and notes in which incorrect diagnoses were listed, but the Justification correctly linked findings to these diagnoses. Generally, raters felt the DDX was the easiest to score.

Raters suggested possible changes to the rubric. Suggestions were made to clarify the frequency of key findings and the frequency of incorrect links (e.g., replacing "some" with "about half") to facilitate differentiation between score levels. The lowest scoring level for the Workup task was expanded to include placing the patient in unnecessary risk due to either ordering or *omitting* tests. Raters agreed that case- and level-specific scoring guidelines were especially helpful in achieving calibration. They recommended that raters review all sample notes as a group and develop scoring guidelines for each case prior to rater training. Despite granular case and level-specific scoring guidelines, raters still felt the need to exercise clinical judgment. Raters took about 5–7 min to score each note and typically scored about 10–15 notes in one sitting.

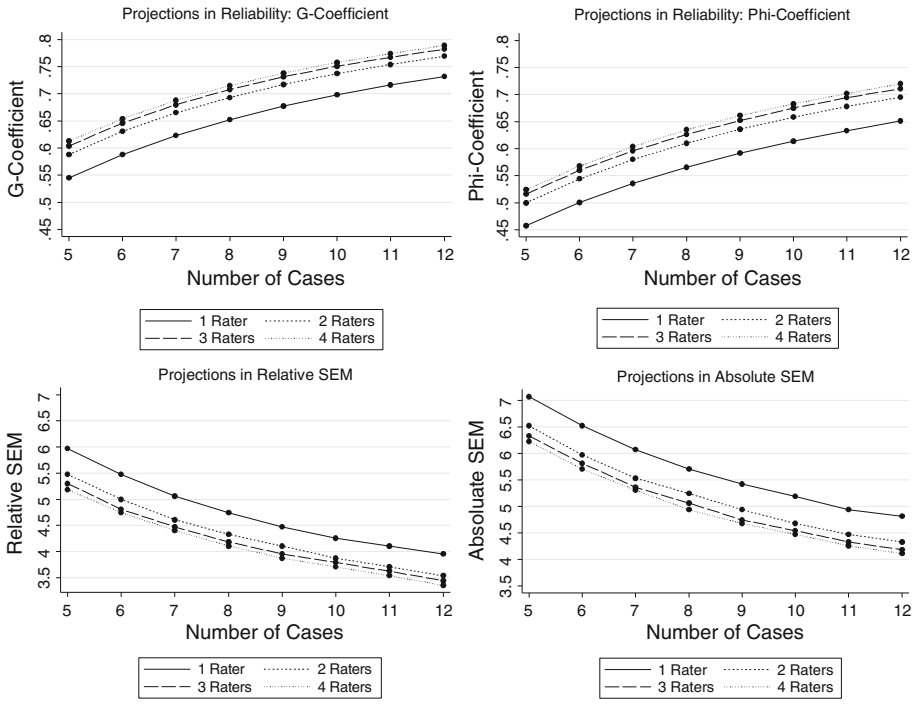

Fig. 1 Projections in reliability and Standard Error of Measurement (SEM) for the revised patient note scoring rubric. *Note* Absolute SEM, $\sigma(\Delta)$, based on Φ-coefficient, and the relative SEM, $\sigma(\delta)$, based on the G-coefficient. The SEMs correspond to percent scores. For example, using the absolute SEM, a student with a 70 % note score, based on 6 cases and 1 rater has a confidence interval of 70 $\pm$ 6.5 %; for 6 cases and 2 raters, the confidence interval becomes narrower, 70 $\pm$ 6.0 %

Discussion

This study is based on a revised PN scoring rubric to address previously noted limitations in Park et al. (2013) and examines rater-related issues when scoring the new PN format with the added diagnostic justification section. We examined whether raters contribute a substantial degree of variability to PN scores and whether raters can be effectively trained to generate reproducible PN scores. Prior studies that examined the impact of raters predate changes to the USMLE PN format or have not specifically designed their study to investigate how raters contribute to the validity of PN scores.

The overall composite score weighted kappa of .79 indicates a high level of confidence in the consistency of scores assigned by raters (Landis and Koch 1977). At the task-level, %-exact agreement ranged between 46 % (Justification) and 88 % (DDX), and the corresponding weighted kappa ranged between .48 and .90, indicating moderate to high task-level rater agreement. While the inter-rater reliability for the overall PN score indicates good rater agreement, partly due to high agreement from DDX, further efforts to train and achieve consensus for Justification may be indicated; the difficulty in scoring the Justification task was also noted by the raters in the qualitative data gathered. For pass-fail decisions, the %-exact agreement increased from 60 to 90 % (Kappa = .61). PN scores across cases were more reliable than case-specific scores, which were more subject to

variability due to raters. As expected for a local examination consisting of only five cases, the G-and Φ-coefficients were moderate at .59 and .50, respectively. Double scoring the PNs can increase reliability and measurement precision; however, having more than two raters score a single PN has a diminishing return.

The variance component analyses based on the G-study and associated reliability projections from this study provide new insights, as previous studies were unable to combine both rater and case facets into a single design. Rater-related variance was modest, when compared to case and task specificity, with the person-by-case-by-task interaction accounting for 31 % of the total variance; person-by-case interaction accounted for 13 %.

While increasing the number of cases results in the most rapid gains in reliability and measurement precision, a combination of adjustments in raters and the number of cases may provide an optimal condition in some settings. Double-scoring the PNs was generally equivalent to adding two more cases. Considering that each PN takes about 5–7 min to score, double scoring may represent a worthwhile investment compared to resources needed for preparing additional case materials, training SPs, and administering new cases. However, the precision gained from double scoring is a reduction in measurement error of only 1 %.

The literature on assessor cognition indicates that raters are trainable, yet have substantial idiosyncrasies that are often unrelated to clinical expertise; raters may use these unique performance models or samples of learner behavior to shape their frame of reference when assessing students (Gingerich et al. 2014; Williams et al., 2003). However, raters agreed that training was effective to overcome difficulties in scoring. Based on feedback from the raters, this study proposes the following guidelines to improve inter-rater reliability and facilitate future training efforts:

- Review the exemplar note and case- and level-specific scoring guidelines at the start of each scoring session;
- Keep a running list of difficult scoring decisions and rationales, to help maintain consistency; share these frequently across raters for a given case;
- Identify a few notes exemplifying different levels of performance, to help raters calibrate and discriminate differences between score levels;
- Note any concerns about case portrayal, case content, or the exemplar note for subsequent follow up.

This study was conducted at a single institution using a single class of medical students; however it builds upon findings from earlier administrations of the exam and addresses limitations noted in a prior study. Additional discussions are underway to further refine the rubric, including efforts to improve training for the Justification task. Collaborations are underway to replicate findings from this study with other medical schools to examine the generalizability of our study. In conclusion, results from this study indicate a high level of confidence in the consistency of note scores, supported by response process and internal structure validity evidence that could aid the assessment of and feedback for medical students preparing for the new USMLE Step-2 CS PN format. This study also underscores threats to validity that may be overcome through effective scoring guidelines, rater training, and scoring designs that optimize reliability, by balancing the number of raters and cases.

Acknowledgments The authors thank the following faculty raters who participated in rescoring the patient notes for this study: Ananya Gangopadhyaya, MD, Nimmi Rajagopal, MD, Olga Garcia-Bedoya, MD, Alexandra Van Meter, MD, and Asra Khan, MD. The authors also thank Robert Kiser for creating an online scoring system to compile rater scores.

Compliance with ethical standards

Conflict of interest None.

Ethical standards This study was approved by the institutional review board of the University of Illinois at Chicago.

References

AERA, APA, & NCME. (2014). *The standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Boulet, J. R., Ben-David, M. F., et al. (1998). An investigation of the sources of measurement error in the post-encounter written scores from standardized patient examinations. *Advances in Health Sciences Education: Theory and Practice, 3*, 89–100.

Boulet, J. R., Rebbecchi, T. A., et al. (2004). Assessing the written communication skills of medical school graduates. *Advances in Health Sciences Education: Theory and Practice, 9*, 47–60.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Cianciolo, A. T., Williams, R. G., et al. (2013). Biomedical knowledge, clinical cognition and diagnostic justification: A structural equation model. *Medical Education, 47*, 309–316.

Clauser, B. E., Harik, P., et al. (2008). The generalizability of documentation scores from the USMLE step 2 clinical skills examination. *Academic Medicine, 83*, S41–S44.

Downing, S. M. (2003). Validity: On meaningful interpretation of assessment data. *Medical Education, 37*, 830–837.

Federation of the State Medical Boards & National Board of Medical Examiners. (2015). Step 2 clinical skills (CS) content description and general information. Philadelphia, PA: United States Medical Licensing Examination. Retrieved June 25, 2015 from http://www.usmle.org/pdfs/step-2-cs/cs-info-manual.pdf

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Gilliland, W. R., La Rochelle, J., et al. (2008). Changes in clinical skills education resulting from the introduction of the USMLE step 2 clinical skills (CS) examination. *Medical Teacher, 30*, 325–327.

Gingerich, A., Kogan, J., et al. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education, 48*, 1055–1068.

Haist, S. A., Katsufrakis, P. J., et al. (2013). The evolution of the United States Medical Licensing Examination (USMLE): Enhancing assessment of practice-related competencies. *Journal of the American Medical Association, 310*, 2245–2246.

Hombo, C. M., Donoghue, J. R., et al. (2001). A simulation study of the effect of rater designs on ability estimation (ETS Research Report No. RR-01-05). Princeton, NJ: ETS. Retrieved June 25, 2015 from http://www.ets.org/research/policy_research_reports/publications/report/2001/hseq

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5–8.

National Board of Medical Examiners. (2015). NBME 2013 Annual Report. Philadelphia, PA: NBME. Retrieved June 25, 2015 from http://www.nbme.org/PDF/Publications/2014Annual-Report.pdf

Park, Y. S., Lineberry, M., et al. (2013). Validity evidence for a patient note scoring rubric based on the new patient note format of the United States Medical Licensing Examination. *Academic Medicine, 88*, 1552–1557.

Southern Illinois University. (2015). DX Justification Scoring Form. Carbondale, IL: SIU School of Medicine. Retrieved June 25, 2015 from http://www.siumed.edu/oec/CCX_ASSESSMENTS/2015/DX%20Justification_scoring%20form.pdf

United States Medical Licensing Examination. (2015). 2014 Performance Data: Step 2 CS. Philadelphia, PA: NBME. Retrieved June 25, 2015, from http://www.usmle.org/performance-data/default.aspx#2014_step-2-cs

Whelan, G. P. (1999). Educational commission for foreign medical graduates: Clinical skills assessment prototype. *Medical Teacher, 21*, 156–160.

Williams, R. G., Klamen, D. A., et al. (2003). Special Article: Social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*, 270–292.

Williams, R. G., Klamen, D. L., et al. (2014). Variations in senior medical student diagnostic justification ability. *Academic Medicine, 89*, 790–798.

Yudkowsky, R., Park, Y. S., et al. (2015). Characteristics and implications of diagnostic justification scores based on the new patient note format of the USMLE Step 2 CS exam. *Academic Medicine, 90*, S56–S62.