Title

Developing an OSTE to address lapses in learners' professional behavior and an instrument to code educators' responses

Journal
Academic Medicine, 79(9)

ISSN
1040-2446

Authors
Srinivasan, M
Litzelman, D
Seshadri, R
et al.

Publication Date
2004-09-01

Peer reviewed

Developing an OSTE to Address Lapses in Learners' Professional Behavior and an Instrument to Code Educators' Responses

Malathi Srinivasan, MD, Debra Litzelman, MD, Roopa Seshadri, PhD, Kathleen Lane, MS, Wei Zhou, Stephen Bogdewic, PhD, Margaret Gaffney, MD, Matt Galvin, MD, Gary Mitchell, MD, Patricia Treadwell, MD, and Lynn Willis, PhD

ABSTRACT

Purpose. To develop an instrument for measuring medical educators' responses to learners' lapses in professional behavior.

Method. In 1999, at the Indiana University School of Medicine, a 22-item checklist of behaviors was developed to describe common responses used by educators responding to learners' lapses in professional behavior. Four medical students were trained to portray lapses in professional behaviors. These students and seven clinical observers trained to categorize behaviors as present or absent. Interrater reliability was assessed during 18 objective structured teaching evaluations (OSTEs). Videotaped OSTEs were coded twice at a one-month interval for test–retest reliability. Items were classified as low, moderate, or high inference behaviors. Script realism and educator effectiveness were assessed.

Results. Educators rated OSTE scripts as realistic. Raters observed an average of 6 ± 2 educator behaviors in reaction to learners' lapses in professional behavior. Educators' responses were rated as moderately effective. More experienced educators attempted more interventions and were more effective. Agreement was high among raters (86% ± 7%), while intraclass correlation coefficients decreased with increasing inference level. From videotaped OSTEs, raters scored each behavior identically 86% of the time.

Conclusions. Accurate feedback on educators' interactions in addressing learners' professionalism is essential for faculty development. Traditionally, educators have felt that faculty's responses to learners' lapses in professional behavior were difficult to observe and categorize. These data suggest that educators' responses to learners' lapses in professional behavior can be defined and reliably coded. This work will help provide objective feedback to faculty when engaging learners about lapses in professional behavior.

Acad Med. 2004;79:888–896.

Assessing a learner's professional behavior has recently become a significant focus in medical education, as patients, educators, and accreditation bodies have defined competencies that together promote better medical practice.[1–3] In 1999, the Accreditation Council for Graduate Medical Education (ACGME) endorsed six general

***Dr. Srinivasan** is assistant professor, Department of Medicine, University of California, Davis, School of Medicine, Sacramento, California. At the time of this study, Dr. Srinivasan was a National Research Service Award Fellow at the Regenstrief Institute for Health Care, which is affiliated with the Indiana University School of Medicine (IUSM), Indianapolis, Indiana. **Dr. Litzelman** is professor and associate dean of medical education, Department of Medicine, IUSM. **Dr. Seshadri** is assistant professor, Departments of Pediatrics and Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois. **Ms. Lane** is research statistician, Department of Medicine, IUSM. **Mr. Zhou** is research statistician, Department of Medicine, University of California, Davis, School of Medicine. **Dr. Bogdewic** is associate dean for faculty affairs and professional development, IUSM. **Dr. Gaffney** is clinical associate professor, Department of Medicine and Department of Dermatology, IUSM. **Dr. Galvin** is clinical associate professor, Department of Psychiatry, IUSM. **Dr. Mitchell** is professor of clinical medicine, Department of Medicine, IUSM. **Dr. Treadwell** is assistant dean of cultural diversity and professor of pediatrics, Department of Pediatrics, IUSM. **Dr. Willis** is professor, Department of Pharmacology and Toxicology, IUSM. Drs. Srinivasan, Litzelman, Bogdewic, Gaffney, Galvin, Mitchell, Treadwell, and Willis are members of the Health Ethics Leadership Program, IUSM.*

Correspondence and requests for reprints should be addressed to Dr. Srinivasan, Department of Medicine, University of California, Davis, School of Medicine, 4150 V. Street, Sacramento, CA 95817; telephone: (916) 734-7005; e-mail: ⟨malathi@ucdavis.edu⟩.

competencies that define core areas of medical training.[4] As a consequence, over the next several years residency programs must develop measures to demonstrate that residents are competent in these areas for continued program accreditation. Professionalism, one of the six ACGME competencies, is a content area that has been difficult to develop and evaluate objectively.[5,6] For instance, cognitive educational theory focuses on the attitudes and thought processes of the person whose behavior manifests as either professional or unprofessional. However, internal thought processes cannot be observed in the workplace. Use of behavior educational theory, conversely, allows us to define professionalism as an observable set of behaviors—such as caring for the patient, use of language that promotes respect for the patient, and responsible followthrough on commitments. Identifying critical issues that may cause lapses in professional behavior is important in medical education if educators are to improve the learning environment and help learners better interact with their patients and colleagues.[5]

Lapses in learners' professional behavior occur frequently at all levels of training in medicine. In one study of inpatient medical teams, physicians and students displayed lapses in professional behavior or attitudes up to once a hour.[7] In another study, over one-quarter of anesthesia residents were cited for lapses in professional behavior at end-of-rotation clinical evaluations. Despite the prevalence of lapses, clinician educators usually ignore learners' hostility, uncaring attitudes, and disrespectful behaviors when they occur.[8] Educators might be reluctant to engage lapses in professional behaviors because of discomfort in addressing these behaviors, the belief that the learner's behavior or attitudes are intrinsic to the person and cannot be changed, the fear that engaging a learner who is behaving unprofessionally might reflect negatively on their own evaluations by residents, the belief that other medically oriented issues take precedence during their interactions, or lack of experience in confronting lapses in professional behaviors.

There is a need to develop reliable measures of professional behavior,[9] to help educators intervene during lapses in their learners' professional behavior,[10,11] and to provide educators with feedback about their ability to address these lapses. These measures will allow faculty to identify, and later change, root causes of learners' lapses in professional behaviors. To help educators address such lapses, the Health Ethics Leadership Program at Indiana University School of Medicine created a faculty development course that taught educators to recognize, evaluate, and intervene during learners' lapses in professional behaviors.[11] Educators who participated in the course reported that it was effective. Educators' skills improved immediately after the course, and new skills emerged at six months. However, self-report of performance might over- or underestimate an educator's actual ability and performance. Objective methods to determine an educator's response to learners' lapses in professional behavior are needed.

Developing objective measures to describe how educators address their learners' lapses in professional behavior will allow reliable feedback to educators about their abilities and help them better engage their learners. Since there is no "gold standard" (criterion validity) for determining educators' responses to lapses in professional behavior, or for the effectiveness of their responses, our first step was to describe and quantify faculty's behaviors in approaching lapses in professional behavior. Specifically, in this study we sought to develop a simulated educational encounter in which learners demonstrate lapses in professional behavior to educators, develop a checklist of operationally-defined educator responses in confronting learners' lapses in professional behavior, and determine the interrater and test-retest reliability of this instrument during simulated encounters with learners. In this report we will describe how educators performed when confronted with lapses in professional behaviors, and we present data on the reliability of a checklist used to quantify their responses.

Method

An Overview of OSTEs

Because learners' lapses in professional behaviors occur sporadically, we created a simulation during which educators interacted with actors who portrayed obvious unprofessional behaviors. Educational simulations in which educators interact with students-actors are called objective structured teaching evaluations, or OSTEs. These OSTEs allow us to carefully observe the educator's responses in a controlled setting in which the stimuli provided by the student-actors are carefully standardized.

The educators are aware that the situation is artificial, and that they are being observed. These OSTEs are meant to assess *best* educator performance in a testing situation, rather than *usual* educator performance in more natural educational settings, such as in clinic or on wards. Studying best educator performance addresses the question: "Can the educator address these behaviors?" Conversely, studying usual educator behavior addresses the question: "Will the educator address learners' lapses in professional behavior if observed in natural educational settings?" Before assessing how educators would address lapses in professional behaviors in a natural educational setting, we wanted to assess whether they could address the behaviors at all. Thus, we also chose not to use unannounced standardized students in clinics, without first gauging the "best performance" ability of our educators.

During the OSTEs, medical student actors portray uncaring, disrespectful, or hostile behaviors. These standardized students provoke the educators during a clinical presentation. The interactions are stopped after five minutes. Raters observe and rate interactions using a checklist.

Below we describe the OSTE script development, rater training, and checklist psychometric properties. Our study was approved by the Institutional Review Board of the Indiana University School of Medicine.

Script Development

In 1999, we developed two OSTE scripts from observed clinical interactions in which a learner, who is behaving unprofessionally, presents a stable patient to an attending physician. The patient has no pressing medical issues. In the first script, a resident becomes extremely frustrated with a noncompliant, hypertensive, obese clinic patient and uses derogatory language to describe the patient. In the second script, a tired medical student presents an intoxicated, stable emergency department patient, who wants admission for food and shelter. The student again uses provocative behaviors. Four of us (MS, DKL, GM, PT) tested and revised the scripts. Scripts delineated key elements, such as the learner's background, stresses and perspective, and the patient's history and physical exam findings. Since each OSTE lasted only five minutes, the script also included a "30-second presentation," after which the educator would be given time to respond.

Student Training

We recruited four fourth-year medical students at the Indiana University School of Medicine as standardized students, since they could address clinical questions more realistically than could actors without a medical background. Students trained for six hours over three weeks. Learners' lapses in professional behavior were described as a composite of unprofessional words, body postures, or tones of voice. To help train the students, we generated over 90 unprofessional phrases from comments heard in clinic or inpatient wards. These unprofessional phrases were grouped by our Health Ethics Leadership Program members as mildly, moderately, or extremely unprofessional. Learners' unprofessional body postures included rolling eyes, slouching, or not making eye contact. Unprofessional tones of voice included speaking very loudly, punctuating words harshly for emphasis, and using tones indicating annoyance with the patient or educator. Students practiced these behaviors in role-plays with each other and with the trainer (MS). They learned to modulate the level of their lapses in professional behavior from mildly disrespectful to extremely antagonistic and confrontational.

The students' goal was to have the educator engage their lapses in professional behavior and not spend time on the patient's stable medical issues. Thus, although the scripts varied in patient presentation, the students' behaviors were consistently provocative. Students began the OSTE with moderate lapses in professional behaviors, escalating their behaviors if the educator focused on medical issues. Students often "frontloaded" the interaction to immediately attract the educator's attention. For instance, they made statements such as, "You won't believe what just dragged in off the street last night," or used terms such as "dirtball" and "waste of space." Determinations of "moderate" or "severe" lapses in professional behavior were made after raters and MS watched trigger videotapes of prerecorded interactions from prior faculty development courses, and through group consensus.

Only students with good clinical and professional evaluations were recruited into our study, so that we could confidently inform educators that the students were truly acting, and did not need professional remediation.

Instrument Development

Since identifying medical educators' responses to provocative stimuli is a relatively new area of study, we sought to determine a reasonable (but not exhaustive) set of responses that educators might exhibit during these encounters. Those of us who are Health Ethics Leadership Program members drew 22 common educator responses to learners' lapses in professional behavior from interactions observed during our faculty development seminar,[9,11] from the literature,[6,8,10,12] from review of faculty development videotapes, and from other observed clinical interactions. These responses were defined as concretely and explicitly as possible for our checklist (see Table 1). For instance, "listening actively" was coded if educators paraphrased and asked relevant follow-up questions, or paraphrased soon after the standardized student's comments. "Clarification of specific choice of word or action" was coded if the educator used the unprofessional phrase in a sentence. "Redirecting the learner towards medical care" was coded if educators redirected the learner to discuss medical care.

To cluster the checklist items, four Health Ethics Leadership Program members ranked how easily the raters could correctly code the checklist items. This ease of observability, or "inference score," for items was ranked on a five-point scale. Items were categorized as low inference (≤2, easiest to observe), moderate inference (2–4), and high inference (≥4, hardest to observe) based on mean scores. For example, "interrupting" or discussing the legal implications of the student's behavior ("You'll be sued!") were considered low inference (easily observable), while "attempting to stimulate learners' self-reflection" was considered high inference.

A summative effectiveness question asked raters to assess the educator's effectiveness ("Overall, did the educator's interventions seem effective?"), where 1 was ineffective, 4 was moderately effec-

Table 1

Educators' Responses to a Learner Who States "This diabetic patient came in smelling like alcohol and threw up on my shoes. I just walked out and let the other student pick up his care. What a dirtball!" from an Objective Structured Teaching Evaluation on Lapses in Professional Behavior, Indiana University School of Medicine, 1999

Educators' Response	Description	Response Coded by Rater as "Present"	Response Coded by Rater as "Absent"
Low inference*			
Redirecting the learner toward medical care	Educator ignores the lapse in professional behavior (trigger word, attitude or action) and discusses medical issues.	"OK, so what was his blood sugar?"	"So, you walked out of the room, and transferred his care to someone else?" *(restating learner's statement)*
Clarified specific choice of words or action	Educator must use the unprofessional trigger word or action in a sentence that inquires about intent of the learner.	"What did you mean when you called the patient a 'dirtball'?"	"What did you just say?!" *(did not use the disrespectful term in a sentence)* "A 'dirtball' huh?" *(example of repeating disrespectful word and waiting)*
Moderate inference			
Listening actively	Educator paraphrases (restates what the learner has said in the educator's own words) and asks relevant follow-up questions within a short time frame.	"What I hear you saying is that . . . *(summary, restating key points)* . . . I understand that you have had several very difficult call nights. Are you getting enough sleep?"	"I'm surprised that you are so angry about this patient. I expected more from you. What is going on?" *(example of judgmental statement)*
Assessing the learner's motivating behavior	Educator tries to determine root cause of the learner's lapse in professional behavior. These can be statements or questions.	"You seem to be angry. Tell me about how you're feeling." "What about this patient makes it difficult for you to take care of him?" "Why did you leave the room so quickly?"	"Are you tired?" *(naming learner's potential state, but not directly assessing the behavior)* "That smell really pushes your buttons." *(summative, observational statement, but indirect)*
High inference			
Stimulating self-reflection in the learner	The educator attempts to elicit the learner's internal value system by having the learner reflect on her/his desires, motivations or goals.	"Tell me, thinking back when you applied for medical school, are you now the type of doctor that you envisioned at that time?" "If you put yourself in this patient's shoes, what kind of care would you want from your doctor?"	"Medicine is a tough field. I've been challenged by a lot of these patients as well." *(empathic statement)*

*Inference is the ease of observability. Inference scores for 22 checklist behaviors were ranked by Health Ethics Leadership Program members on a five-point scale (1 = low inference; 5 = high inference). The authors used the mean scores per item to define three categories: low inference (≤2, easiest to observe), moderate inference (2–4), and high inference (≥4, hardest to observe).

tive, and 7 was exceptionally effective. The scale was anchored by the raters watching videotapes of ineffective interventions (an educator ignores the lapse in professional behavior = 1) and exceedingly effective interventions (an educator helped the learner to self-reflect and gain insight = 7).

Rater Training

Seven health care professionals (two physicians, two registered nurses, and three individuals with a Masters in Social Work), each with over five years of direct teaching experience, were recruited as additional clinical observers. Our four standardized students and seven clinical observers rated educators' performances in each OSTE. All 11 raters trained together for eight hours over three weeks. During training sessions, raters practiced using the checklist to identify educators' responses during role-plays and by watching videotaped OSTEs. They resolved areas of disagreement about coding educators' responses through discussion.

We originally attempted to determine how often educators exhibited each checklist behavior during their five-minute interactions. However, our raters reported difficulty counting the number of times each checklist behavior occurred. Therefore, instrument items were coded as "absent" or "present." We originally wanted to assess the effectiveness of each educator behavior on the standardized student.

Raters also reported difficulty in assessing the effectiveness of individual behaviors on the learner; therefore, a global item on educator effectiveness was included as a final checklist item.

Educator Recruitment

We recruited educators with different levels of experience in clinical teaching to help increase the likely responses during the OSTE. Nine volunteer educators were recruited: three members of the Health Ethics Leadership Program, three faculty (two residency program directors and one senior nurse faculty member), and three relatively inexperienced medicine residents (who teach students and other residents). Educators were asked to address the learner's lapses in professional behavior during the interaction, and signed an informed consent form, which allowed us to videotape their encounters.

Educators' Responses during OSTEs and Checklist Interrater Reliability

To assess interrater reliability, 18 OSTEs were performed in one afternoon. During these OSTEs, educators were seated at the front of a large conference room with a standardized student sitting next to them in a chair. Ten raters sat at tables around the room watching the educator and standardized student interaction (the 11th rater was the standardized student acting in the OSTE). The interactions were videotaped, and the educator and standardized student had microphones attached to their lapels for improved audio quality. Each of the nine educators had the opportunity to engage each script once. After each five-minute OSTE, the educator and raters were given time to complete their evaluation/rating forms. Educators rated the realism of the interactions on a five-point scale (1 = not realistic; 5 = very realistic), described their reaction to the learners' lapses in professional behavior, and in an open-ended format described the interventions they attempted.

The students and clinical observers rated each interaction. Thus, for each of the 18 OSTEs, 11 checklists were completed. Because of scheduling conflicts, three raters were absent for a few of the OSTEs, bringing the total number of completed checklists to 189 of a potential 198.

Average raw rater agreement was determined for each item. However, raw agreement scores do not correct for agreement that occurs by chance. Readers may be familiar with the kappa statistic, which is used to correct for chance agreement when two raters evaluate a behavior. Less familiar may be the intraclass correlation coefficient (ICC) statistic, which is used to correct for chance agreement when more than two raters rate an item (as in our data set).[13,14] Like an analysis of variance (ANOVA), ICCs measure interrater variability as a proportion of total variability due to the model, and are reported as 0 (no agreement after chance) to 1.0 (perfect agreement after chance). ICCs were calculated using SAS version 8.0 (SAS Institute, Inc., Cary, NC) and were nested for educator and script using fixed and random effects. Generally, items with ICCs $> .8$ are considered to have very good agreement, $>.6$ good agreement, $>.4$ moderate or reasonable agreement, and $< .4$ poor agreement.

Using the live 18 OSTEs, overall educator effectiveness was averaged among all raters, and by three rater types (standardized student acting in the OSTE, all standardized students, and clinical observers). Using repeated measures ANOVA, effectiveness scores were compared for the level of educator experience (Health Ethics Leadership Program members, faculty, and residents), as well as the rater type (clinical observers, standardized students, and the standardized student participating in the OSTE). Exploratory factor analysis was conducted to help cluster the instrument items.

Test–Retest Reliability

To assess test–retest reliability, we asked raters to score ten OSTEs pretaped from prior faculty development workshops at two time points—short enough to assume that the underlying constructs had not changed, but long enough so that raters were unlikely to remember their responses. Thus raters watched the prerecorded videotapes a few days after the live OSTE, and then one month later.

Of the total rater group, two standardized students and six clinical observers were able to complete rating at both time points, while the others had conflicting out-of-town rotations or other time conflicts. Agreement ratings per item per behavior per OSTE over time were calculated for each rater, then averaged over rater groups. Finally, the six Health Ethics Leadership Program members who helped develop the instrument also watched the videotaped OSTEs at one point in time, so we could compare the responses of raters with more experience with the behaviors of other raters.

Results

Rater Reliability

For each of the 18 live OSTEs, ratings for the 22 items had high raw agreement. A behavior was considered to have occurred if more than half of the raters coded the behavior as "present." Using this definition, an average of 5.8 ± 1.9 responses were attempted by each educator during each OSTE. Five responses were noted in over half of the OSTEs (see Table 2). Educators actively listened during all 18 live OSTEs, and in 17 OSTEs, redirected the discussion to the medical care of the patient. In 12 OSTEs, educators explored the motivating behavior of the patient, and in 11 OSTEs, educators both verbally acknowledged the learner's emotions and

Table 2

Inter- and Intrarater Reliability of Objective Structured Teaching Evaluations (OSTEs) for Evaluating Educators' Potential Responses to Learners' Lapses in Professional Behavior, Indiana University School of Medicine, 1999

		18 Live OSTEs						10 Pretaped OSTEs Evaluated at 0 and 1 Month		
		Interrater Reliability†						Intrarater Reliability‡		
		All Raters (no. = 11)		Clinical Observers (no. = 7)		Standardized Students (no. = 4)				
Educators' Response	Response Frequency*	Agreement	Intraclass Correlation Coefficient	Agreement	Intraclass Correlation Coefficient	Agreement	Intraclass Correlation Coefficient	Raters % Identical (no. = 8)	Clinical Observers % Identical (no. = 6)	Standardized Students % Identical (no. = 2)
Average over all items	5.8	.87	.30	.85	.30	.88	.30	.85	.84	.88
Low inference										
Redirected toward medical care provided	17	.91	.27	.90	.27	.92	.27	.92	.86	.83
Acknowledged learner's emotions verbally	9	.89	.69	.89	.70	.88	.67	.80	.80	.80
Clarified specific choice of words or actions	8	.91	.69	.92	.69	.89	.70	.87	.87	.87
Interrupted learner	8	.86	.58	.85	.60	.86	.55	.73	.77	.83
Repeated disrespectful words	2	.90	.48	.89	.48	.92	.48	.93	.93	.93
Rescheduled discussion time explicitly	0	.96	.43	.95	.43	.97	.43	.97	.96	.93
Pondered the impact of learner's behavior lawsuits/legal implications	0	1.00	—	.71	—	1.00	—	.97	.71	1.00
Moderate inference										
Listened actively	18	.83	.01	.72	.01	1.00	.02	.75	.82	.97
Assessed patient's motivating behavior	12	.78	.34	.77	.35	.81	.32	.78	.78	.77
Empathized with learner	11	.79	.36	.82	.40	.74	.30	.77	.76	.73
Assessed learner's motivating behavior	3	.87	.50	.83	.50	.93	.51	.75	.77	.80
Educated learner about gaps in knowledge/skills/attitudes	2	.71	.07	.65	.07	.82	.08	.75	.78	.83
Pondered impact of lapse in professional behavior on patient or family	3	.83	.40	.86	.40	.78	.41	.95	.86	.93
Pondered impact of lapse in professional behavior on medical care	2	.81	.22	.82	.21	.85	.24	.70	.71	.73
Discussed behavior in third person	0	.82	.02	.81	.02	.82	.03	.88	.88	.87
Explicitly stopped or prevented future behavior	0	.96	.06	.95	.06	.97	.06	.93	.96	1.00
Used humor	0	.94	.15	.97	.15	.92	.15	.98	.94	.87
Raised voice or appeared impatient	0	.90	.24	.89	.24	.93	.23	.93	.94	.97
Gave nonverbal disapproval	0	.91	.13	.92	.14	.90	.13	.87	.94	.93
High inference										
Stimulated self-reflection in learner	6	.76	.28	.75	.28	.79	.30	.68	.72	.80
Judged or moralized	2	.84	.37	.83	.37	.86	.38	.87	.89	.93
Pondered impact of lapse in professional behavior on learner	0	.92	0	.94	0	.89	0	.93	.93	.93

*Positive responses per OSTE, where over half of raters (supramajority) coded a response "present."

†Average agreement of the individual rater with supramajority response and intraclass correlation coefficient.

‡Percentage identical responses at 0 and 1 month, per rater, for every item per OSTE, then averaged over raters.

empathized with the learner. Three other behaviors occurred in six or eight OSTES: clarification of specific choice of words or actions, interrupting the learner, and attempting to stimulate learners' self-reflection.

Raw agreement for all 22 checklist items was high among raters (.71–1.00) (see Table 2). Average uncorrected agreement was .85 for clinical observers, and .88 for standardized students. However, once corrected with ICC, only six of the 22 checklist items had psychometrically useful properties. Two of the checklist items had ICCs slightly great than .6: clarified specific choice of words or actions (.69), and acknowledged the learner's emotions verbally (.69). Five responses had ICC between .4 and .6: interrupted the learner (.58), assessed the learner's motivating behavior (.50), repeated disrespectful words (.48), rescheduled discussion time explicitly (.43), and pondered impact of learner's behavior on patients or family (.40). Ratings of educators between the first and second live OSTE did not significantly differ in either overall educator effectiveness (repeated measures ANOVA) or specific item performance (chi-square with fixed and random effects).

ICCs varied with inference level. Our development team considered low-inference items easier to observe than higher-inference items. Five of the seven low-inference items had better ICCs, while only two of the 12 moderate-inference items, and none of the three high-inference items had ICC > .4. From the ten pretaped OSTEs, baseline raw agreement among the raters from the Health Ethics Leadership Program (.86, range .701.00) was similar to other raters (.84, range .65–.99). The ICCs were not correlated with the frequency of responses. No individual rater's scores significantly affected the overall agreement measures. Factor analysis did not detect any significant clustering within the instrument.

Test–retest reliability to evaluate educators' responses was determined by comparison of ratings at a one-month interval using the ten videotaped OSTEs (see Table 2). Overall agreement was 85% for each item, per OSTE, over one month. Two standardized students rated interventions identically for 88% of the items while the six clinical observers rated identically for 84% of items.

Educator Responses to Learners' Lapses in Professional Behavior

After the OSTEs, educators reported that the scripts were realistic, with a mean realism score of 4.2 ± .7, out of 5. When facing lapses in professional behavior, the educators reported feeling frustrated (no. = 6), irritated (no. = 3), déjà vu (no. = 2), empathetic (no. = 2), surprised (no. = 1), angered (no. = 1), concerned (no. = 1), and disappointed (no. = 1). In four OSTEs, educators reported no real emotional response. Three educators described unrelated issues.

On average, raters identified more behaviors than the educators reported using during the OSTEs (5.8 versus 3.1, $p < .05$, t test). Some educators used techniques not on our checklist, such as historical personification,[15] while none discussed the legal implications of the student's behavior.

Educators who were more experienced (Health Ethics Leadership Program members and faculty) attempted a greater variety of responses to the lapses in professional behavior compared to resident educators during the OSTEs ($p < .05$, repeated measures ANOVA). Health Ethics Leadership Program members used an average of 7.5 ± 1.5 different responses per OSTE, while the other faculty tried 6.0 ± 2.3 responses and residents tried 4.5 ± 1.2 responses.

Effectiveness

The effectiveness score captures overall educator performance. During the 18 live OSTEs, observers rated educators as less than adequate in their effectiveness, with a mean score of 3.2 ± 1.5. However, the Health Ethics Leadership Program members were rated as more effective than were residents (4.1 ± 1.4 versus 2.5 ± 1.2, $p < .05$, repeated measures ANOVA with fixed and random effects). Faculty's effectiveness fell between these two groups (3.1 ± 1.4). There was no significant difference in effectiveness scores among different rater groups, although the students acting in the OSTE tended to rate the educators more favorably (4.0 ± 1.4) than did nonacting students (3.4 ± 1.3, $p = .02$). The agreement after chance (ICC) for this global item was .40, and did not vary by rater type.

Discussion

The ability to provide reliable feedback to educators about their interventions in addressing learners' lapses in professional behavior is a cornerstone of faculty development efforts to teach professionalism. In this study we evaluated educators' behaviors in response to learners' lapses in professional behavior in a testing situation, and examined the reliability of an instrument to measure those behaviors. The artificial OSTE environment provides the educator a chance to develop, practice, and demonstrate interventions for learners' lapses in professional behavior, in advance of their actual occurrence in clinical settings. Our data lend insight into the responses of educators to learners' lapses in professional behavior.

Educators, even after being instructed that they were to address learners' lapses in professional behavior, directed their discussions with standardized students toward the medical care of a stable patient. When they did address the unprofessional attitude or behavior of the learner, they used six techniques predominantly: active listening, assessing the learner's motivating behavior, acknowledging the learner's emotion, empathizing with the learner, clarifying learner's choice or words or actions, and attempting to stimulate self-reflection in the learner. Members of the Health Ethics Leadership Program used significantly more interventions than did residents, suggesting that training and experience can increase the range of responses used by educators. There were no significant differences between ratings of educators using the two scripts, suggesting that the two OSTEs may test the same underlying behavioral construct of educators' responses to learners' professional lapses.

While it is important to provide educators with specific feedback about the techniques that they use during their interactions with standardized patients, the use of a global rating score on educator effectiveness provides key summative information about their overall performance. Global performance scores have been shown to have better correlation coefficients than specific item scores, and may be a better indicator of an educator's true skills.[16] Despite using several sets of responses during the OSTEs, the overall effect of the educators' interventions were generally rated as less than adequate. We noted an experience effect, with the more experienced educators obtaining higher effectiveness ratings than did the less experienced residents. It is important to note, however, that the students who were acting in the OSTE rated the educator as more effective than did the other raters, which suggests that there may be a stronger impact for the learner than is obvious to the outside observer. In a clinical setting, our goal is to ensure that educators' interventions have a positive effect on learners displaying lapses in professional behavior. Thus, slightly higher ratings by the standardized student displaying lapses in professional behavior suggest that the impact of these interpersonal interactions may not be gauged adequately by outside observers. Additionally, in a natural clinical setting, the educator would not be constrained by the five-minute time limit imposed during our OSTE, and could spend more time interacting with the learner. Thus, even an adequate level of effectiveness demonstrated during our OSTE is reassuring that short educator interventions in clinical settings may affect learners' lapses in professional behavior.

The instrument, however, did not perform as well as anticipated. Interrater raw agreement was high for both standardized students and clinical observers in rating the presence or absence of 22 educator responses to learners' lapses in professional behavior during live OSTEs. Further, the global effectiveness score had low-moderate agreement after chance, similar to that reported by Regehr et al.[16] Yet the variability in the corrected agreement scores demonstrates how complex this area of study can be, even with careful training and evaluation methods. Our data do provide some evidence that raters, given low inference measures of educators' interventions, can be trained to recognize and categorize an array of behavioral interventions. The low-to-moderate agreement scores in the global effectiveness item point to the difficulty of gauging effectiveness of interventions even with rigorous training of raters.

Test–retest agreement was high in coding educators' behaviors on videotaped OSTEs at a one-month interval. Members of the Health Ethics Leadership Program also coded the educator responses from the videotaped OSTEs. The ratings of standardized students and clinical observers did not significantly vary from ratings of the Health Ethics Leadership Program members.

Since conducting OSTEs can be expensive, we attempted to determine whether it was necessary to have multiple raters observing the educators' interventions. We did not find meaningful differences between our three rater groups (standardized students acting in the OSTE, and standardized students and clinical observers watching the interactions). This finding confirms work done by others who demonstrated that rater reliability is influenced more by training than by professional background.[17] Future OSTEs on professionalism using these instruments may use only the participating standardized students as raters.

Limitations

Our data have several limitations. First, educators reported their behaviors in free text and did not use a checklist. Second, we found educationally insignificant differences in ratings between rater groups of about 3%. Because of the complexity of the number of repeated measures in our data (educators, over raters, over items, and over scenarios), and because we used current generalizability analysis, we could not statistically state "no difference" between groups. Third, although educators' effectiveness was judged to be adequate or less than adequate, these results might be inflated by the volunteer bias of educators already interested in this area.

Fourth, for our data set, we feel that ICCs underestimate the true agreement beyond chance. ICCs may be influenced by prevalence of a behavior. Behaviors with very high or low prevalence produce data near the extremes of 0 and 1, and may produce low ICCs regardless of the agreement levels.[18] In observing undirected educator responses during a five-minute OSTE, the prevalence of any of the 22 interventions of interest is low. For instance, for one item, agreement was greater than 96%, yet the ICC was only .05. This striking lack of correlation between our ICC and agreement scores emphasizes the need to develop better statistical measures to describe relationships near data extremes.

However, for standard comparison, ICCs were used to explore agreement trends in our data. Using this method, we found that only six of the 22 behaviors had ICC $> .40$. Items characterized as low inference behaviors (that is, easier to observe and categorize) had ICC of $> .40$, while only one of the moderate inference, and none of the low inference responses had ICC of $> .40$. These trends suggest that describing specific observable faculty responses may be helpful in an even more robust model. Despite limitations of the ICC statistics, our data suggest that lower inference (easier to observe) behaviors had better agreement after chance than higher inference behaviors.

Finally, our study was designed to test instrument reliability, which must be shown before issues of validity can be addressed.[17] We took pains to develop an instrument that was likely to be valid by using direct observations of educators' behavior and expert consensus as the basis for our item generation. Our instrument also showed some ability to discriminate between levels of educators' experience in addressing lapses in learners' professionalism. Additionally, the OSTE scenes were drawn from real cases, and the participating educators felt that the scenes were realistic. In the absence of a criterion standard, our study provides a first step in developing valid models of educators' responses to learners' lapses in professional behavior.

Education Policy Implications

We hypothesize that both personal and system-wide issues contribute to lapses in professional behavior. Development of a systematic approach to addressing learners' professionalism will allow us to identify personal contributors to lapses in professional behavior (such as poor communication or intrapersonal skills in need of remediation) and system-wide contributors (such as chronic sleep deprivation or learners' excessive workload). As medical educators begin to quantify all aspects of physician behavior, the ability to remediate learners who demonstrate lapses in professional behavior must occur during both formative and summative evaluations. We hope that this description and initial categorization of educators' responses to learners' lapses in professional behavior will provide a nidus in faculty development seminars. We hope to sensitize educators to the range of responses that they might decide to use in these encounters. Educators should have data to show they are able to intervene in learners' lapses in professional behavior, that their interventions can be quantified, and finally that their interventions are effective.

Addressing lapses in professional behavior directly is a first step toward improving patient care and the learning environment. Stimulating learners to self-reflect and compare their behavior to their personal or professional value system may help them successfully address stressful or difficult situations. Yet to fully approach lapses in professional behavior, both the individual learner and the system conditions that provide the settings for lapses in professional behavior must be addressed.[19] As the ACGME begins to require an outcomes based curriculum for residency education, medical educators must be prepared to address the challenges in both the individual learner and the systems of care that promote undesirable professional behaviors.

References

1. Cruess SR, Cruess RL. Professionalism must be taught. BMJ. 1997;315:1674–7.
2. Irvine D. The performance of doctors. I: professionalism and self-regulation in a changing world. BMJ. 1997;314:1540–2.
3. ABIM Foundation, ACP-ASIM Foundation, European Federation of Internal Medicine. Medical professionalism in the new millennium: a physician charter. Obstet Gynecol. 2002;100:170–2.
4. Outcome Initiative. Chicago: Accreditation Council for Graduate Medical Education, Outcome Initiative Advisory Council, 1999.
5. Ginsburg S, Regehr G, Stern D, Lingard L. The anatomy of the professional lapse: bridging the gap between traditional framewords and students' perceptions. Acad Med. 2002; 77:516–22.
6. Arnold L. Assessing professional behavior: yesterday, today and tomorrow. Acad Med. 2002;77:505–5.
7. Stern DT. Values on call: a method for assessing the teaching of professionalism. Acad Med. 1996;71:S37–S39.
8. Burack JH, Irby DM, Carline JD, Larson EB, Root RK. Teaching compassion: attendings' responses to problematic behavior. J Gen Intern Med. 1996;11:113.
9. Swick H. Toward a normative definition of medical professionalism. Acad Med. 2000;75: 612–6.
10. Cottingham A, Marriott D, Litzelman D. Teaching caring attitudes. Acad Med. 1998; 73:571.
11. Srinivasan M, Bogdewic S, Gaffney M, et al. Effectiveness of a faculty development course on "Teaching Caring Attitudes". J Gen Intern Med. 1999;14:156.
12. Abbott LC. A study of humanism in family physicians. J Fam Pract. 1983;16:1141–6.
13. Fleiss JL. The Measurement of Inter-rater Agreement. Statistical Methods for Raters and Proportions. 2nd ed. New York: John Wiley, 1981:212–36.
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420–8.
15. Charon R. To render the lives of patients. Lit Med. 1986;5:58–74.
16. Regehr G, Freeman R, Hodges B, Russell L. Assessing the generalizability of OSCE measures across content domains. Acad Med. 1999;74:1320–2.
17. Stillman P, Swanson D, Regan MB, et al. Assessment of clinical skills of residents utilizing standardized patients. A follow-up study and recommendations for application. Ann Intern Med. 1991;115:158–9.
18. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43:543–9.
19. Srinivasan M. Medical professionalism: more than simply a job. JAMA. 1999;282:814.