

Using the Integrated Standardized Patient Examination to Assess Clinical Competence in Physical Therapist Students

Karen J. Panzarella, PT, PhD, and Andrea T. Manyon, MD

Background and Purpose. To test the success of integrated curricula and “clinical readiness” of students in Doctor of Physical Therapy (DPT) programs, meaningful measurements of student competence are necessary. Competence testing must include an authentic clinical encounter. Standardized patients (SP) have long shown to be reliable teaching and assessment tools in health professional education and should be explored with DPT students.

Method/Model Description and Evaluation. This paper analyzes a new performance assessment tool, the Integrated Standardized Patient Examination (ISPE), specifically designed for assessing clinical competence in DPT students. Students performed 2 ISPE cases; one portraying a herniated lumbar disc (HD) and the other a status post-cerebral vascular accident (CVA). A maximum of 30 minutes was allotted to conduct a history, to respond to 4 “integration” questions asked by the SP, and to perform a targeted examination. Expert physical therapist (PT) evaluators, the investigator, and students scored each section and an overall-encounter quality section. The SP also scored the overall-encounter section. All participants filled out a feedback survey about the experience.

Outcomes. Acceptable reliability was demonstrated via interrater agreement on items that used a dichotomous scale, whereas the items requiring the use of the 4-point rubric were somewhat less

reliable. For the entire scale, both cases had a significant correlation between the expert–investigator pair of raters; for the CVA case ($r = .547, P < .05$); and for the HD case ($r = .700, P < .01$). On the overall encounter section, the SPs scored students higher than the other raters. Students’ self-assessments were most closely aligned with the investigator. Content validity was gathered in the process of developing cases and patient scenarios.

Discussion and Conclusion. Future studies should examine the effect of rater training upon reliability. Criterion or predictive validity could be further studied by comparing students’ performances on the ISPE with other estimates of students’ competence. The unique integration questions of the ISPE were judged by experts and students to have good content validity; confirming that integration—a most crucial element of clinical competence—can be practiced, assessed, and used to determine student clinical competence.

Key Words: Standardized patients, Objective structured clinical examination (OSCE)

BACKGROUND AND PURPOSE

Physical therapist students are educated to be autonomous practitioners, with direct access to patient care in 43 states, and practice privileges in all settings.¹ For PTs, *competence* is the ability to perform an examination that includes taking a history, reviewing systems, selecting and administering tests, and effectively communicating to the patient.²

Competence is not demonstrated by skills that are present in isolation, but in situations requiring integration of knowledge, attitudes, and skills.³ In education settings, the goal of competence is for students to develop a systematic approach that considers pathophysiologic mechanisms, while simultaneously taking a history, formulating initial hypotheses, performing the physical exam, selecting clinical tests, reprioritizing the hypotheses, making the differential diagnosis, prescribing treatment, and educating the patient.⁴

The Commission on Accreditation of Physical Therapy Education (CAPTE) requires that program faculty utilize a variety of effective methods to assess student competence, safety, and clinical readiness.⁵ Assessment methods should be meaningful, applied consistently to all students, and linked to the intended learning outcomes.⁶ While curricular reform and integrated curricula are well described and implemented in the health professions, meaningful measures of clinical competence have yet to be described, implemented, or validated.⁷ In particular, integration is challenging for educators to measure because it is an internal process occurring in the mind of the student when solving clinical problems.⁸

Over the past 30 years, standardized patient examinations have become accepted as valuable teaching and assessment tools in medical education.⁹ *Standardized patients* (SP) are lay people trained to represent a given diagnosis or problem. They can be trained to represent conditions of varying complexity that are tailored to expectations for student performance.

Since the early 1980s, medical schools have used SPs in objective structured clinical examinations (OSCEs) to assess student performance.¹⁰ An OSCE uses a circuit of brief stations (5-10 minutes each) that focus on recall and application of knowledge to examine a range of clinical skills. SP responses are reproducible, ensuring all students experience the same patient portrayal, thus allowing equitable evaluation of student performance.^{11,12} The use of SPs in OSCEs helps approximate an authentic clinical encounter.¹⁰ SP-based OSCEs have demonstrated reliability and validity, leading to the 2004 decision by the National Board of Medical Examiners (NBME) to begin including an SP examination in the licensure process for physicians.^{13,14}

Although OSCEs are an accepted method to evaluate students’ clinical skills, they are not very effective in assessing overall clinical competence. The OSCE format does not afford students the opportunity to demonstrate their integration of scientific knowledge and

Karen Panzarella is professor and director of Clinical Education for the Doctor of Physical Therapy Program at the University at Buffalo, 515 Kimball Tower, Buffalo, NY 14214 (kjp1@buffalo.edu). Please address all correspondence to Karen Panzarella.

Andrea Manyon is professor and chair of the Department of Family Medicine at Upstate Medical University, Syracuse, NY.

Received January 17, 2008, and accepted August 4, 2008.

patient-communication skills.⁷ Its dichotomous scoring scale cannot document the progression of competence over time, adding to its limitations.¹⁵

The use of a SP in physical therapist education is being advocated as a cost-effective learning tool for transition from classroom to clinic. Students report that an SP encounter is very realistic.¹⁶ Several studies report use of SPs in physical therapist education to support teaching experiences. Learning activities in which faculty and classmates role play “mock” patients have shown increased student confidence and satisfaction over traditional teaching methods.¹⁷ One model used SP interaction as part of a teaching module on diabetes, demonstrating that students were able to integrate and communicate newly acquired information more effectively than those who participated in a typical lecture-only experience.¹⁸ Another reported use of standardized patients is in teaching students professional behaviors. In this usage, students are required to integrate prior learning and allowing faculty to assess core value acquisition.¹⁹

The only reported study in PT education using SPs to evaluate clinical performance obtained acceptable reliability and validity from student and physical therapist subjects who performed a history and physical examination of an SP with shoulder pain.¹¹ The authors of this study stated that, “evaluation of clinical performance using standardized techniques is imperative for physical therapy educators if we are concerned about best practice in health sciences education.”¹¹ To date, there have been no reports of an SP-based examination as a high-stakes assessment of clinical readiness in PT education.

The Integrated Standardized Patient Examination (ISPE), initially piloted with medical students, was developed as a model for integrated assessment of clinical competence.²⁰ The ISPE builds on the strengths of the OSCE format, adding a method for evaluating students’ integration of the curriculum and communication of their knowledge to the patient within an authentic clinical encounter. The ISPE model was applied to Doctor of Physical Therapy students and evaluated for its utility and psychometric qualities.

METHOD/MODEL DESCRIPTION AND EVALUATION

Thirty-four DPT students, in their second year of the 3-year program at the University at Buffalo, participated in this study. Twenty of the DPT students were female and 14 were male. The age range of the subjects was 23-40 years of age with a mean age of 27. The

overall grade point averages (GPAs) of the 34 students, for their academic year, ranged from 3.14 to 3.90, with a mean of 3.7.

Twelve expert evaluators were recruited from the PT faculty and the practicing community; 6 with PhDs, 4 with master’s degrees, and 5 with APTA specialization. The average length of clinical experience was 15 years. The expert evaluators scored the encounter while sitting as non-obtrusively as possible in the examination room with the student and SP. The expert evaluators did not intervene during the procedure or provide any feedback to the student at the time of the encounter. The first author evaluated all videotaped encounters, serving as the criterion evaluator.

Twelve standardized patients (SPs) were recruited from a pool of SPs used in medical education and trained to portray either a patient with an acute herniated lumbar disc (HD) or a patient who suffered a cerebral vascular accident (CVA). All SPs were trained to score the students on the overall encounter quality section and were encouraged to provide anecdotal feedback on the score sheet.

The case scenarios for training the SPs were developed by the authors with input from faculty. Designed patient cases had to be common in physical therapy practice, covered in the newly integrated systems-based curriculum, and compatible with the skill level of the second-year DPT student.²¹ Training the SP consisted of reviewing the scenario in detail, a demonstration of the 30-minute encounter (with the investigator playing a physical therapy student and each SP practicing their role with feedback). With the prior experience of the SP, training only took 1 hour for the group.

This standardized patient examination experience was built into the newly integrated DPT curriculum as an assessment exercise in PT 605: Case Management. The syllabus stated students must pass the ISPE to pass the course. The passing score was set at 80% of the total score for each case. Students who failed the first attempt were given one more opportunity, following remediation, to pass the ISPE. Students participated in an informational lecture and a question-and-answer session to help prepare them. Students were told that they would perform 2 patient cases (encounters), one involving the neuromuscular system and one involving the musculoskeletal system. They were informed that the cases chosen were covered in corresponding courses that semester. The concept of the standardized patient was explained and videotapes of encounters with medical students were reviewed. Students were informed

of the time constraints, the procedure that would be followed on the day of the exam, and the format for scoring the exam. Particular discussion was held about the integration section and the rubric for scoring. They all agreed to sign informed consent.

The unique feature of the ISPE (ie, the component that tests the students’ integration) is the questions the SP asks the students to gain more information about their condition. The questions are matched to an attribute of integration (Table 1). These attributes of integration were positioned into a rubric used to score the student’s response. The rubric uses a 4-point grading scale in which 0 is detrimental, 1 is below expectations, 2 is meets expectations, and 3 is exceeds expectations.

A medical school suite established for the purpose of SP encounters was used. The medical suite consists of 12 patient exam rooms, each equipped with a ceiling-mounted video camera, a sink, a patient examination table, and chairs. To supplement scenarios, arm slings and quad canes were used by the SP to portray the CVA case. To aid the students in responding to patient questions, lumbar disc models were present in the examining rooms for the herniated disc case and models of the brain were present in the rooms for the CVA case.

Each of the 34 students performed the 2 cases, with a short break in between. Students did not receive feedback until a few days after the encounters. The order of the cases was counterbalanced across the sample. The expert evaluator, SP, and the student were in the examination room. For each case, the students had 30 minutes to conduct a problem-focused history, respond to the 4 questions asked by the SP (integration), and then perform a targeted physical examination.

The expert evaluators and criterion evaluator scored the ISPE in 4 sections: history, integration, physical exam, and overall encounter quality. The core faculty agreed upon the items to include for each section of the score sheets (individualized to each case). Faculty who taught courses in neuromuscular and musculoskeletal finalized the decisions. The first section, history, is a binary checklist of either below expectations or meets expectations. For the HD disc case, a total of 16 points on the history portion were possible and for the CVA case, a total of 15 points were possible. Examples of items on the history section are: determining chief complaint, specifics of pain, functional ability, past medical history, medications, associated symptoms, and social/employment history. The second section, integration, provides a rubric of 4 choices to score the re-

sponse to the SP question. The rubric uses a 3-point grading scale in which 0 is detrimental, 1 is below expectations, 2 is meets expectations, and 3 is exceeds expectations for a total of 12 points. The third section, physical examination (PE), is a binary checklist of either below expectations or meets expectations. For the HD disc case, a total of 9 points were possible, and for the CVA case, a total of 11 points were possible. Items on the PE section are case specific and include such items as: posture inspection, repeated movements, functional mobility, special tests, range of motion, strength, and sensory testing.

The fourth section, overall encounter quality, challenges the rater to give an overall global score for 4 dimensions of communication and knowledge based upon the entire encounter. It is a rubric of 4 questions, using a 4-point grading scale in which 0 is detrimental, 1 is below expectations, 2 is meets expectations, and 3 is exceeds expectations. A total of 12 points are possible for each case. Following the completion of cases, students and the expert evaluators filled out a feedback survey (Table 9).

All data cleaning, recoding, and manipulation were carried out using SPSS Version 11.0²² and Microsoft Excel. Descriptive statistics were used to examine the internal consistency among exam items through calculation of mean scores, variance or spread of scores, and scores in each of the 4 sections of the exam. Internal consistency/factor analysis was calculated using Eigen values and percent variance between cases.

Data were triangulated with scores from the criterion evaluator, the expert evaluators, and the students. The criterion evaluator scored the encounters via videotape, blinded to the other scores. Students self-scored their videotaped encounter using the same score sheets as the other evaluators. The students' self-evaluation was compared to the criterion standard and the expert evaluator's score. Cohen's κ was used to examine reliability across all raters, by student, by rater, and by case. In addition, agreement by item was examined for each case. Correlations and differences between total scores and the integration item totals were calculated between raters.

The effect of the rater was examined to determine the significant sources of variance contributing to scores on 68 cases graded by 2 evaluators. Each student had a different set of raters (r), but all students (s) took both cases (c). This is a partially nested design, where the rater is nested in the interaction between student and case (r:s) x c. Raters are nested within students because who the raters were depended on the student and different students had different set of raters. In this

Table 1. Measurable Attributes of Integration

<p>The student:</p> <ul style="list-style-type: none"> • Explains the relationship between structure and function. • Explains the role of injury in contributing to dysfunction. • Explains the scientific basis of patients' symptoms. • Explains normal physiologic or biochemical mechanisms. • Explains abnormal physiologic or biochemical mechanisms. • Explains the role of specific risk factors in the pathophysiology or prevention of a disease. • Explains the rationale of a therapeutic intervention, and expected side effects. • Discusses prognosis—the role of intervention, prevention, and/or secondary complications.

Table 2. Mean Total Scores

Case	Expert Evaluator	Criterion Evaluator	Student Self-Assessment
CVA Max Score=50	— X=36.3 SD=6.03 Range=21-45	— X=37.78 SD=3.34 Range=33-44	— X=37.75 SD=4.33 Range=26-44
HD Max Score=49	— X=38.15 SD=5.38 Range=28-49	— X=39.52 SD=4.62 Range=27-46	— X=39.71 SD=3.90 Range=33-46

Abbreviations: CVA, cerebral vascular accident; X, mean; SD, standard deviation; HD, herniated lumbar disc.

design, the variance components estimates for s, c, r:s, sc, and residual error were obtained using ANOVA (analysis of variance). Traditional parametric statistics were used to examine differences in student scores as a function of case or rater.

The feedback survey was analyzed using descriptive statistics and chi-square (χ^2). Pearson correlations were used to test for the relationship between students' ISPE results and their GPA in the DPT program (convergent/concurrent validity). To examine this notion, a series of Pearson correlations and cross tabulations with statistics were carried out using SPSS Version 11.0.²²

A descriptive report of the students' self-evaluations was documented by each student completing a 2-3-page paper about their strengths and weaknesses, which was included in their academic portfolio.

OUTCOMES

All students were able to complete the CVA-case encounter within the 30 minute session. Approximately 25% of the students did not finish the HD encounter or felt they needed more time. Scoring was based upon items completed and therefore students did not fail the exam due to time constraints. Using de-

scriptive statistics, no significant differences occurred among the mean total score of the 3 groups of raters for either the CVA case ($F_{2,62} = .71$) or for the HD case ($F_{2,68} = .81$) (Table 2). The range and standard deviation of the total scores was largest for the expert group of raters for both cases, with the mean total exam score for both cases most congruent between the criterion evaluator and students (Table 2).

For the history section, students could score a maximum of 15 points on the CVA case and 16 on the HD case. On the CVA case, students scored an average of 12.19 points from the experts, 12.61 points from the criterion, and 11.89 from their self-evaluation. On the HD case, students scored an average of 14.36 points from the experts, 13.67 points from the criterion, and 15 points from their self-evaluation. There was no significant difference among the mean scores of the 3 groups of raters (the experts, criterion, and students) for the history section of the CVA case ($F_{2,87} = 1.189, P = .310$). For the HD case, the students rated themselves significantly higher than did the investigator ($F_{2,78} = 3.907, P = .024$). The most congruent mean score on the history section for the CVA case was between the experts and

Table 3. HD Case: Integration Section

"I've been reading about disc problems on line—why do most discs pop out toward the back or side?"						
	3 Exceeds Expectations	2 Meets Expectations	1 Below Expectations	0 Detrimental	Mean	SD
Expert	35%	53%	12%	0%	2.24	.65
Student	61%	36%	3%	0%	2.58	.56
Criterion	35%	65%	0%	0%	2.35	.49
"Why does it feel better when I tilt to the left side?"						
Experts	46%	39%	12%	3%	2.27	.80
Student	39%	54%	7%	0%	2.32	.60
Criterion	48%	49%	3%	0%	2.45	.56
"I hurt my back, but why do I have this numb and tingling sensation in my left leg?"						
Expert	33%	61%	6%	0%	2.27	.57
Student	42%	48%	10%	0%	2.32	.65
Criterion	42%	55%	3%	0%	2.39	.56
"When I raise my leg when I'm lying down, why is it painful?"						
Expert	49%	39%	12%	0%	2.36	.70
Student	39%	58%	3%	0%	2.36	.55
Criterion	52%	48%	0%	0%	2.52	.51
Integration Totals						
Rater	Minimum	Maximum	Range	Mean	SD	
Expert	4	12	8	9.12	1.98	
Student	6	12	6	9.57	1.48	
Criterion	7	12	5	9.73	1.33	

criterion. For the HD, it was between the experts and the students. The history section, being a binary checklist, resulted in fairly close percent agreement on most items, with the exception of 2 items, where the criterion evaluator's score was different than the experts or students by approximately 30%.

Overall, students performed better when answering the integration questions on the HD case than on the CVA case (Tables 3, 4). The means for the integration section are approximately 2 points higher for the HD case (experts: 9.12, students: 9.57, criterion: 9.73) than for the CVA case (experts: 7.73, students: 7.62, criterion: 7.97). However, the means are very similar amongst the raters within each case. There were no significant differences among the mean scores of the 3 groups of raters for the integration section for the CVA case ($F_{2,85} = .772, P = .465$) or for the HD case ($F_{2,96} = 1.145, P = .320$).

There is a significant difference among the mean scores of the 3 groups of raters for the physical examination section of each case. For the CVA case ($F_{2,89} = 8.344, P = .001$) the criterion evaluator rated the students significantly higher than the experts. For the HD case ($F_{2,94} = 8.742, P = .001$), the criterion evaluator rated the students significantly higher than both the students and experts. For the CVA case, the closest means

are between the students and criterion (.97-point difference), with experts and students also very close (1.01 difference). The range of scores was highest from the experts. For the HD case, the means are very close between the experts and students (.61-point difference), while the mean for the criterion is almost 2 points greater than either the experts or students.

The overall-encounter quality section was the only section completed by the SP (Table 5). For this section of the CVA case, there was no significant difference among the mean scores ($F_{3,128} = 2.172, P = .10$) for the 4 groups of raters (experts, criterion, students, and SPs). However, there was a significant difference in the scores from the SP on the HD case ($F_{3,130} = 7.639, P = .001$), with the SP rating significantly higher than the experts, students, or criterion. Using 80% as the standard for passing the exam, 7 students (or 20.5%) failed the CVA case and 8 students (or 23.5%) failed the HD case. Two students failed both cases.

Interrater reliability was evaluated in 2 complementary ways: the percent agreement for each pair of rating dyads, and standard Pearson correlations between raters within students for all items for each case. For the history, integration, and physical exam sections of the ISPE, interrater reliability

was calculated for 3 dyads of raters: the expert–criterion, the expert–student, and the criterion–student. For the overall-interview/encounter-quality section, the SP also rated the students; therefore, 6 dyads of raters were used.

Pearson correlations were calculated across all students, for each rater, and for each section of the 2 cases (Table 6). Scores were correlated from each of the 3 raters (experts, criterions, and SPs) for both cases, with rating domains within each case (history, integration, PE, and overall). The students' self-ratings were compared to the criterion ratings. The overall correlations were slightly higher for the HD case than the CVA case. The expert–criterion pair produced the highest correlations for all 6 rater dyads. Significant correlations occurred between the expert–criterion on the total score for both cases with $P < .01$ on the HD case and $P < .05$ on the CVA case. The expert–student dyad has the lowest correlations out of the 3 pairs of raters. The SP raters are not significantly correlated with the 3 types of raters.

To examine the major source(s) of variance in student scores on the ISPE, the effect of the rater was analyzed. To determine the effect the 12 different expert raters had upon the reliability of the ISPE, mean total scores were examined by rater (Table 7). A

Table 4. CVA Case: Integration Section

“My left arm isn’t much use to me—will the movement ever come back?”						
	3 Exceeds Expectations	2 Meets Expectations	1 Below Expectations	0 Detrimental	Mean	SD
Expert	6%	73%	21%	0%	1.85	.50
Student	13%	72%	15%	0%	1.97	.54
Criterion	6%	84%	10%	0%	1.97	.40
“Why am I always banging into things on my left side when my stroke was on the right side of my brain”						
Expert	19%	68%	13%	0%	2.06	.57
Student	21%	59%	17%	3%	1.97	.73
Criterion	15%	73%	12%	0%	2.03	.53
“When I walk my left foot drags, why is this happening?”						
Expert	13%	56%	25%	6%	1.75	.76
Student	14%	64%	22%	0%	1.93	.60
Criterion	32%	55%	10%	3%	2.16	.74
“I have muscle spasms in my leg and my Doctor told me there is some medication for that. What is the medication and will it help?”						
Expert	6%	64%	27%	3%	1.73	.63
Student	10%	59%	24%	7%	1.72	.75
Criterion	6%	76%	18%	0%	1.88	.49
Integration Totals						
Rater	Minimum	Maximum	Range	Mean	SD	
Expert	2	12	10	7.73	1.79	
Student	3	11	8	7.62	1.81	
Criterion	4	11	7	7.97	1.45	

Table 5. Overall-Encounter Quality Section

Student appeared comfortable conducting the interview; non-verbal communication and body language were appropriate, professional.			
3 exceeds expectations	2 meets expectations	1 below expectations	0 detrimental
Student conducted the interview in a well-organized manner.			
3 exceeds expectations	2 meets expectations	1 below expectations	0 detrimental
Student’s speech was clear and easily heard.			
3 exceeds expectations	2 meets expectations	1 below expectations	0 detrimental
Overall interface between knowledge and communication.			
3 The interface is fluent. Knowledge level and communication skills are consistently excellent.	2 The interface is reasonable. Knowledge level and communication skills are good.	1 Interface is awkward. Knowledge and communication skills are marginal.	0 The interface leads to confusion. Inadequate knowledge or communication skills.

12 x 2 (12 raters by 2 cases) factorial ANOVA was performed finding a significant rater effect ($F_{11,42} = 6.3, P < .001$). There was no significant difference in mean total scores between cases, and no interaction between case and rater. Clearly there are some lenient raters such as Raters 4 and 11 and more harsh raters such as Rater 5.

Internal consistency was measured for the integration section and the overall/interview quality section for each case via Cronbach’s

. The internal consistency measures how well each item correlates with each other and the whole test. Table 8 reports the internal consistency estimates for each case for the different raters, ranging from .47 to .91. Overall, 13 of the 20 correlations were above .70, which is acceptable but worth exploring in further research.

Validity was obtained through the construction of cases and SP scenarios, along with feedback from the experts and students

who participated in this study. Content validity of the ISPE was achieved by the prudent construction of the cases, with particular attention paid to the development of the questions asked by the SPs to prompt integration. The construct of competence was measured through the development of real-life case scenarios of commonly seen patient conditions in the field of physical therapy. Each case was developed according to the objectives outlined in the respective course

Table 6. Correlations Between Section and Overall Totals Among Raters

Section	Expert–Criterion	Expert–Student	Criterion–Student	SP–Criterion	SP–Student	SP–Expert
CVA Case						
History	.442 ^a	.533 ^b	.388	—	—	—
Integration	.360	.055	.247	—	—	—
Phys Exam	.167	-.064	.324	—	—	—
Overall	.313	.230	.024	.199	.090	.550
Total	.547^a	.076	.121	—	—	—
HD Case						
History	.348	.046	.444 ^a	—	—	—
Integration	.286	.237	.391	—	—	—
Phys Exam	.562 ^b	.254	.709 ^b	—	—	—
Overall	.315	.196	.120	.241	-.071	.138
Total	.700^b	.313	.457	—	—	—

^a*P* < .05^b*P* < .01.

syllabus and DPT program accreditation report.²³

Construct validity was also achieved from feedback about the ISPE and its ability to measure the intended construct of competence through integration. The feedback forms contained 4 questions about the ISPE experience (Table 9). A chi-square comparison was used to compare the responses from the 2 groups on each of the 4 questions. Questions 2 and 3 did not have a significant difference in the responses between the 2 groups. However, Questions 1 and 4 resulted in a significant difference [Question 1: $X^2(3) = 8.61, P = .035$, Question 4: $X^2(3) = 10.27, P = .016$].

The experts perceived more relevance to real-life cases than did the students. An overwhelming majority of the experts (91%) felt the ISPE emulated an actual patient visit well or extremely well, adding to the content and construct validity of the tool. Question 2, asking whether the exercise demonstrates an effective integration skill, was pivotal in inferring validity of the tool. Experts and students combined responded favorably (76.9%) that the ISPE demonstrated an effective integration skill well or extremely well.

Following the 4 Likert responses on the feedback form, 5 open-ended questions were asked about integration and the clinical setting. Experts' responses included statements of integration being extremely important to clinical decision making. They also commented that performing a patient assessment requires integration with analysis of didactic information and, as demonstrated from the ISPE, it can be practiced and improved upon. Some experts responded that integration may have an inherent component but that it can certainly be improved with practice and feedback—it can be discussed and/or modeled to explore students' ability to relate to patients.

Table 7. Mean Total Score Ratings by Rater

Rater	N	CVA Mean	N	HD Mean	Overall N	Overall Mean
1	0 ^a	—	2	37.5	2	37.5
2	3	39.0	0 ^a	—	3	39.0
3	3	39.3	3	40.7	6	40.0
4	3	41.3	3	42.0	6	41.7
5	3	25.0	3	30.7	6	27.8
6	3	31.0	3	34.3	6	32.7
7	3	38.3	2	40.5	5	39.2
8	2	39.5	3	41.3	5	40.6
9	0 ^a	—	2	34.0	2	34.0
10	2	33.5	1	32.0	3	33.0
11	2	37.5	3	44.7	5	41.8
12	3	39.0	2	36.5	5	38.0
Overall	27	36.3	27	38.2	54	37.2

^aMissing data precluded the calculation of a total.

Student responses on the open-ended questions focused on the importance of practice prior to entering the real world of patient care. One student responded, "Integration is so important because this is what we are going to do for our careers." When asked how important integration is in a clinical setting, one student responded, "Extremely, hey, this is real life, nothing gets closer." When asked about the preparation for this exercise, students responded with, "We need more simulated PT clinics like this," "We need more practice with feedback," and "I would like the evaluator to give me feedback when I am finished." The inherent component of integration was evident in the student responses as well. However, students responded that integration cannot be taught: "It is something you learn with practice and not taught."

Criterion-related validity was also examined through the relationship between students' overall GPA and their mean total score on the ISPE. Students' GPAs ranged from 3.14 to 3.90, with a mean GPA of 3.48. As shown in Table 10, there is low

association between students' scores on the ISPE and their overall GPA. This low correlation between these 2 constructs may suggest discriminant validity, inferring that the ISPE measures different phenomena from those measured by GPA.²⁴ Further studies will have to explore the convergent and discriminate relationships of this construct before definitive conclusions can be reached.

DISCUSSION

This model examined and critiqued a new performance assessment tool, the Integrated Standardized Patient Examination, for assessing clinical competence. The ISPE is distinct from the OSCE, which has been the cornerstone of competency testing in medical education since the early 1980s.²⁵ The OSCE, using brief stations, tests students' clinical skills in isolation.

The ISPE was designed to be more aligned with real-life clinical settings, where a PT interacts with a patient for 30 minutes and is prompted to integrate through a prede-

Table 8. Internal Consistency for Both Cases

Rater	HD Case (****)	CVA Case (*****)
Expert		
Integration Items	.69	.67
Overall quality	.91	.84
Combined	.87	.84
Student		
Integration Items	.47	.60
Overall quality	.83	.70
Combined	.73	.72
Criterion		
Integration Items	.49	.54
Overall quality	.79	.74
Combined	.79	.77
Overall Quality	.64	.87

Table 9. Results From Feedback Survey

1. During the patient encounter, how well do you think you/students were able to integrate basic science knowledge with clinical skills?						
Rater	1 Extremely Well	2 Well	3 Fairly Well	4 Poorly	Mean	SD
Experts	0%	91%	0%	9%	2.18	.60
Students	25%	57%	18%	0%	1.92	.66
Combined	18%	66%	13%	3%	2.00	.64
2. How well did this demonstrate an effective integration ability or skill?						
Expert	18%	64%	18%	0%	2.00	.63
Student	18%	57%	25%	0%	2.07	.66
Combined	18%	59%	23%	0%	2.05	.64
3. How prepared were you/the students for this integration exercise?						
Expert	36%	45%	9%	9%	1.90	.94
Student	21%	71%	7%	0%	1.85	.52
Combined	26%	63%	8%	3%	1.87	.65
4. How well did the setting emulate an actual patient visit?						
Expert	73%	18%	0%	9%	1.45	.93
Student	21%	25%	36%	18%	2.50	1.1
Combined	36%	23%	26%	15%	2.20	1.1

terminated series of questions asked by the SP. An evaluation of the students' ability to integrate, therefore, is indicator of their emerging competence and clinical readiness.

The outcomes of this model are promising for the ISPE to be a valid and reliable tool. Interrater reliability was high among the 3 groups of raters for the entire scale. Interrater reliability was very good for the history and PE sections of the ISPE. The integration and overall sections show fair agreement. The expert-criterion evaluator pair produced the highest correlations for all rater pairs, which is pivotal when the ISPE is used as a high-stakes exam. Standardized patients seem to be more lenient raters, scoring students consistently higher than the other raters and perhaps should be used for feedback but not for pass/fail decisions. Students' self-assessments were most closely aligned with the criterion, which could be attributed to scoring from a video versus the experts scoring in real time. Effects of the raters were apparent as well as

the effects of the specific case.

These outcomes concur with a similar study by Ladyshevsky et al, which used an SP to portray a rotator cuff injury, with PT students and clinicians producing acceptable reliability.¹¹ Differences in this current study include the integration component, the number of subjects (3 times more), a time limit of 30 minutes for the encounter, and a neurological case in addition to the musculoskeletal case used as a high-stakes assessment. The clinical-like environment of the ISPE also enhances clinical decision-making skills similar to what was demonstrated in a recent study of teaching screening and referral skills using mock patients.¹⁷ The use of the ISPE in PT academic programs is supported by students finding the SP to be realistic and worthwhile in improving their patient care skills as also similarly demonstrated by a pilot study of SPs by Black and Marcoux.¹⁶

Eighteen percent of students said that

this exercise poorly emulated a real clinical situation. This may have been due to the CVA case occurring in the same room design (small medical room) as the HD case. Students expressed that they would typically have a lower, larger mat table to assess patient mobility when evaluating a patient with a CVA in the clinical setting.

Follow-up studies would be useful to further explore issues of validity. For this model, content validity was the primary consideration via the careful construction of cases, symptom portrayal, and SP questions to sample central instructional objectives of the DPT curriculum at various levels of thinking in Bloom's taxonomy.²⁶ Criterion-related or predictive validity was assessed superficially by correlating total scores on the ISPE with GPA—a notoriously weak criterion variable for such purposes with non-interpretible results. This suggests that GPA and the ISPE are measuring discrete constructs (a suggestion of discriminate validity). The ISPE attempts to emulate clinical performance and, therefore, it would be worthy to correlate with other independent estimates of students' clinical competence such as performance on clinical experiences using the Clinical Performance Instrument. Criterion or predictive validity could be further studied by comparing students' performances on the ISPE with scores on the national physical therapy licensing exam. Construct validity may prove more difficult yet, since there seem to be few similar viable measures of the type of competencies the ISPE is designed to measure.

The entire process of participating in the ISPE has been described by the students as an extremely valuable experience that provoked anxiety and challenged them to think on their feet like no other experience ever had. Anecdotal accounts from students who have participated in the ISPE were perceived as closely reflective of real-life clinical encounters. The ISPE created a cognitive conflict for students that may have challenged them to apply their knowledge in a different way. Authentic learning experiences and reflective practice are supported in the PT literature for development of clinical reasoning.¹⁹

Self-assessment (ie, viewing their videotaped encounters) may have challenged students to analyze their performance, distinguish/discriminate their behaviors, and organize their SP encounter. The self-assessment paper that students wrote following the viewing of their videotapes was intended to stimulate thinking at the highest levels of the cognitive domains, synthesis, and evaluation of their performance.²⁶ The integration sections, the unique component of the ISPE, were judged by the experts and students

Table 10. Correlations Between Total Exam Score and Student GPA

Rater	CVA	HD
Experts	-.027	.288
Student	.117	-.267
Criterion	.200	.041

(through questions directly relating to the ability to integrate) to have a great deal of content validity.

The expense of administering an SP examination is always a concern. SPs usually receive between \$10-14 per hour for their training and participation.¹⁶ Costs per student can range from \$35-100, depending upon fees for technical support or reimbursement to the SP center.^{16,26} Costs for this project were grant funded and included payment of \$25 per hour to the expert evaluators, \$10 per hour for the SPs, and an average cost of approximately \$50 per student. Typical cost per student averages around \$75-100 per hour.

Since this study, the ISPE has been adopted into the DPT curriculum at the University at Buffalo. Several changes were made to the procedure, including increasing the time of each encounter from 30 to 45 minutes and using only one case (musculoskeletal). The medical examination suite was not conducive to evaluating neurological patients that require more room for mobility. Students are now sequestered in separate rooms for the entire length of the exam to prevent sharing of information about the cases following an encounter. Over the past 2 years of using the ISPE, approximately 20% of students fail on their first attempt and all students (following remediation), except one, passed on their second attempt. The student who failed after a second attempt was dismissed from the program as the student was on probation.

In order to improve upon the reliability of ratings on items using the rubric, the ratings were collapsed to 3 choices rather than 4. Formal and frequent training of raters is occurring to help reduce rater bias, and to ensure scorers are interpreting the rubrics in the same way. In hindsight, the raters were not given sufficient training prior to scoring. Several raters left items blank and some even scored between points on the rubric scale. This may have been a result of missing something in real time. Since this study, all encounters are scored via DVD. Future studies should examine the effect of bringing all raters to a demonstrably acceptable level of agreement (ie, equal to or greater than 80% agreement with each other or a predetermined criterion). Rater training could be accomplished through a CD-ROM that contained video clips of students performing parts of the ISPE that raters would score.

Perhaps the most important implications of the ISPE resulted from cooperative efforts amongst faculty. This project brought PT faculty (clinical and tenured; on-campus and clinic environments) together to talk about assessment that begins with the end in mind.²⁷ It prompted faculty to view assessment in terms of the integrated curriculum objectives and real-life situations. The entire faculty being involved in the ISPE leads to greater cooperation, validation, and integration of curricular content. It forced the faculty to look at what really matters when we send our graduates out to practice in our health care community.

CONCLUSION

The Integrated Standardized Patient Examination (ISPE) is an important and innovative performance evaluation instrument. This instrument is unique in its ability to measure students' ability to integrate scientific knowledge and clinical skills and to demonstrate that ability through interaction and communication with a patient.

DPT students performed a real-life clinical encounter with an SP portraying either a herniated lumbar disc or a cerebral vascular accident. The 30-minute encounter of taking a history, answering patient questions, and performing a physical examination of tests and measures was evaluated by expert physical therapists, a criterion evaluator, and the students themselves. Reliability was acceptable on items that used a dichotomous scale and there was less reliability for items that used a 4-point rubric. The highest correlation was between the expert PTs and the criterion evaluator.

The outcomes of this model show promise for the ISPE to become a valid and reliable tool, to assess clinical competence in physical therapist students, and to determine clinical readiness prior to sending students on clinical experiences. We hope the tool will prove valuable to other programs in the health sciences seeking to assess student competence. The unique portion of the ISPE, the integration questions, displayed good content validity. The ISPE experience confirms that integration (a most crucial element of clinical competence), when done in the mind of the student, can be practiced, learned, and assessed.

REFERENCES

1. American Physical Therapy Association. 2006 annual report—advancing the profession: enhancing your ability to practice. <http://www.apta.org/AM/Template.cfm?Section=Home&TEMPLATE=/CM/ContentDisplay.cfm&CONTENTID=43142>. Published

- January, 2007. Accessed September 2007.
2. American Physical Therapy Association. *Guide to Physical Therapist Practice*. 2nd ed. Alexandria, VA: American Physical Therapy Association; 2001.
3. Caney D. Competence—can it be assessed? *Physiotherapy*. 1983;69:302-304.
4. Manyon A, Panzarella K, Feeley T, Servoss T. Development of an assessment tool measuring medical students' integration of scientific knowledge and clinical communication skills. *Assessment Update*. 2003;15(1):1-14.
5. Commission on Accreditation in Physical Therapy Education. Evaluative criteria for accreditation of education programs for the preparation of physical therapists. <http://www.apta.org/AM/Template.cfm?Section=CAPTE3&Template=/CM/ContentDisplay.cfm&ContentID=19980>. Effective January 1, 2006. Updated October 2007.
6. Vendrely A. Student assessment methods in physical therapy education: an overview and literature review. *J Phys Ther Educ*. 2002;16(2):64-69.
7. Mavis B, Henry R, Ogle K, Hoppe R. The emperor's new clothes: the OSCE reassessed. *Academic Med*. 1996;71(5):447-453.
8. Harden RM. The integration ladder: a tool for curriculum planning and evaluation. *Med Educ*. 2000;34:551-557.
9. Colliver J, Swartz M, Robbs R, Cohen D. Relationship between clinical competence and interpersonal and communication skills in standardized-patient assessment. *Academic Med*. 1999;74(3):271-274.
10. Doig C, Harasym P, Fick G, Baumber J. An objective look at OSCE. *Academic Med*. 2000;75(10):96-98.
11. Ladyshevsky R, Baker R, Jones M, Nelson L. Evaluating clinical performance in physical therapy with simulated patients. *J Phys Ther Educ*. 2000;14(1):31-37.
12. Edelstein R, Reid H, Usatine R, Wilkes M. A comparative study of measures to evaluate medical students' performances. *Academic Med*. 2000;75:825-833.
13. United States Medical Licensing Examination. Comprehensive review of USMLE. <http://www.usmle.org>. Accessed October 2006.
14. Rose M, Wilkerson L. Widening the lens on standardized patient assessment: what the encounter can reveal about the development of clinical competence. *Academic Med*. 2001;76(8):856-859.
15. Fowell SL, Bligh JG. Recent developments in assessing medical students. *Postgrad Med J*. 1998;74:18-24.
16. Black B, Marcoux BC. Feasibility of using standardized patients in a physical therapist education program: a pilot study. *J Phys Ther Educ*. 2002;16(2):49-56.
17. Boissonnault W, Morgan B, Buelow J. A comparison of two strategies for teaching medical screening and patient referral in a physical therapist professional degree program. *J Phys*

- Ther Educ.* 2006;20(1):28-35.
18. Hale L, Lewis D, Eckert R, Wilson C, Smith B. Standardized patients and multidisciplinary classroom instruction for physical therapist students to improve interviewing skills and attitudes about diabetes. *J Phys Ther Educ.* 2006;20(1):22-27.
 19. Hayward L, Blackmer B, Markowski A. Standardized patients and communities of practice: a realistic strategy for integrating the core values in a physical therapist education program. *J Phys Ther Educ.* 2006;20(2):29-37.
 20. Panzarella KJ, Manyon AT. A model for integrated assessment of clinical competence. *J Allied Health.* 2007;36(3):157-164.
 21. Collins JP, Gamble GD. A multi-format interdisciplinary final examination. *Med Educ.* 1996;30:259-265.
 22. SPSS [computer program]. Version 11.0. Chicago, IL: SPSS Inc; 2008.
 23. Meier ST. *The Chronic Crisis in Psychological Measurement: A Historical Survey.* New York, NY: Academic Press; 1994.
 24. Harden RM. What is an OSCE? *Med Teacher.* 1988;10(1):19-22.
 25. Bloom BS. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain.* New York: McKay; 1956.
 26. Cusimano M, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE. *Academic Med.* 1994;69(7):571-576.
 27. Covey SR. *The Seven Habits of Highly Effective People.* New York, NY: Simon and Schuster; 1989.