Generalizability of a Progress Clinical Skills Examination and the Assessment of the Growth in Clinical Skills in a Medical School Curriculum with Early Clinical Experiences

Heather S Laird-Fick

Michigan State University

Chi Chang

Michigan State University

Ling Wang

Michigan State University

Carol Parker

Michigan State University

Robert Malinowski

Michigan State University

Matthew Emery

Michigan State University

David Solomon (✉ dsolomon@msu.edu)

Michigan State University <https://orcid.org/0000-0002-3130-5240>

Research article

Keywords: “progress assessment”, “standardized patients “clinical skills”, “OSCE”, “program evaluation”

Posted Date: July 31st, 2019

DOI: <https://doi.org/10.21203/rs.2.12217/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background This study evaluates the generalizability of an eight-station progress clinical skills examination and assesses the growth in performance for six clinical skills domains among first- and second-year medical students over four time points during the academic year. **Methods** We conducted a generalizability study for longitudinal and cross-sectional comparisons and assessed growth in six clinical skill domains via repeated measures ANOVA over the first and second year of medical school. **Results** The generalizability of the examination domain scores was low but consistent with previous studies of data gathering and communication skills. Variations in case difficulty across administrations of the examination made it difficult to assess longitudinal growth. It was possible to compare students at different training levels and the interaction of level of training and growth. Second-year students outperformed first-year students, but first-year students' clinical skills performance grew faster than second-year students narrowing the gap in clinical skills over the students' first year of medical school. **Conclusions** Case specificity limits the ability to assess longitudinal growth in clinical skills through progress testing. Providing students with early clinical skills training and authentic clinical experiences appears to result in the rapid growth of clinical skills during the first year of medical school.

Background

Progress testing uses broad-based examinations that are designed to assess end of curriculum objectives and are given at regular intervals over a course of study. Progress testing was first implemented in the 1970's and has grown to be widely used in medical schools and to some extent residency programs.¹ There is a considerable body of research that suggests progress testing is particularly well suited for less structured problem focused curricula and encourages learners to study in ways that promote understanding rather than rote memorization.² Progress testing also provides a rich and systematic source of data for student feedback, program evaluation and identifying students that need remediation.

Progress testing has almost exclusively been implemented in the form of written examinations and rarely as clinical skills examinations. We could only identify one systematic implementation of clinical skills progress testing, and it was in an internal medicine residency program.³ Many medical schools are now placing first-year students into authentic clinical settings, increasing the potential value of Progress Clinical Skills Examinations (PCSEs) for providing systematic feedback on the growth of the students' clinical skills. Furthermore, in the USA, the expectations for medical school graduates entering residency training are being operationalized through Core Entrustable Professional Activities (EPAs) for Entering Residencies⁴ that do not necessarily lend themselves to being assessed via written examinations.

The changes in both the structure of undergraduate medical education and the expectations for graduates has increased the value of PCSE as an integral part of medical school assessment and evaluation programs. Although there has been ample research on the psychometrics, acceptance and impact of written progress testing¹, there has been relatively little research done on PCSEs.

This study focuses on the psychometric and practical challenges of implementing a progress clinical skills program and our early findings about the growth in clinical skills over the first two years of medical school in a curriculum that includes early clinical experiences and extensive clinical skills training. Specifically, this study estimates the generalizability of an eight-station PCSE for assessing longitudinal growth in clinical skills over the course of the curriculum as well as for assessing cross-sectional differences in student performance at different levels of training. Secondly, this study assesses the impact of authentic clinical experiences and weekly clinical skills training in a simulation laboratory with faculty feedback on the growth of clinical skills over the first two years of medical school.

Methods

Subjects and Setting - In the fall of 2016, Michigan State University's College of Human Medicine (CHM) implemented a new curriculum called the Shared Discovery Curriculum (SDC). The SDC is organized around patient chief complaints and concerns. Students start their medical training by learning basic data gathering and patient communication skills through simulation-based educational experiences that includes standardized patient encounters and direct observation and feedback from clinicians. After eight weeks of training, students begin working two half-days a week in clinic settings with medical assistants and nurses. As they gain more experience, their clinical responsibilities grow. At the same time, they work to master clinical applications of basic science knowledge independently, in small groups, and in a weekly large group session. They also continue to receive four hours per week of clinical skills training in the medical school's simulation centers.

Measures - The students are evaluated for both formative and summative purposes via progress testing twice each semester using both written and clinical skills examinations. Descriptions of the development and piloting of the PCSE used in the SDC has been published elsewhere.^{5,6} The examination uses an Objective Structured Clinical Examination (OSCE) format⁷ and consists of eight 15-minute Standardized Patient (SP) encounters with 10-minute post encounter stations. Each encounter assesses some combination of patient interaction skills, hypothesis-driven history gathering, physical examination, counseling and safety behaviors using checklist and rating items completed by the SP. Post-encounter tasks, which are graded by faculty members, assess the application of medical knowledge, clinical reasoning, and clinical documentation. The PCSE stations are designed to assess EPAs not easily assessed by written examinations. The performance of an examinee is reported as the percentage of possible points students achieved across all eight cases in each of the six domains.

The standardized patients used in the PCSE are trained to the specific PCSE cases they simulate. Both their portrayal of the case script and the accuracy of their completion of the checklist/rating forms are assessed by the staff in the simulation center before each PCSE is given and include measurements of inter-rater reliability. Adjustments to either the case or how the SP is trained are made when these quality assurance efforts identify a problem.

Design - The PCSE is given as part of a broad-based progress assessment that also includes written examinations. These progress assessments occur twice each semester for a total of 20 assessments over the course of the medical school curriculum. Third- and fourth-year students are assessed using the PCSE once each semester. Depending on their rotational schedule each third- and fourth-year student is assessed either in the first or second PCSE given that semester. To pass in a semester, students must pass at least one of the two PCSE given that semester with scores deemed appropriate for their level of training. Third- and fourth-year students who do not meet course-specific expectations for all skill areas on the PCSE take a make-up exam to demonstrate their competency.

Since students in all four years of training take the same PCSE examination at roughly the same time, we can potentially observe the growth in clinical skills both longitudinally over the course of each students' medical training as well as cross-sectionally between students with different levels of training taking the same PCSE. The SP cases for each PCSE are drawn from a pool of cases that are continually being developed. The SP cases will eventually be reused but only after the students who were originally assessed via the case have graduated. As a result, students will not encounter cases from a previous PCSE in which they were evaluated.

As noted, third- and fourth-year students take a single PSCE each semester with a portion of the students taking the first administration of the PCSE in a given semester and others, the second. Given this complication, we chose to focus on first- and second-year student performance for this study. During fall semester 2017 and spring semester 2018, four PCSEs were conducted as part of the SDC progress assessment. Second-year students from the first matriculation class in the SDC and first-year students from the second matriculation class completed the assessments. The scores in these four administrations of the PCSE for the two classes of students were used to assess growth in the students' clinical skills during the first two years of the curriculum and the psychometric characteristics of the PCSE.

Generalizability Study - We conducted a generalizability analysis⁸ of the PCSE domain scores separately for both first- and second-year students. We considered standardized patient cases as the only facet in the universe of admissible observations. This resulted in a student by SP case ANOVA design for estimating the variance components used in the generalizability study.

As noted above, we are interested in both cross-sectional comparisons of the first- and second-year students' performance as well as the longitudinal growth of the students' performance across multiple administrations of the PCSE. These two types of comparisons have different generalizability coefficients, and standard errors of measurement.⁹ In the cross-sectional comparisons, the students at each level of training are assessed on the same eight SP cases. The error variance is equivalent to the error variance as defined in classical test theory.¹⁰ When making longitudinal comparisons of the same students over multiple examinations, the comparisons are based on different SP cases that are not perfectly parallel. As such, longitudinal comparisons include an additional source of error from the variation among cases, have lower generalizability and larger standard errors of measurement than cross-sectional comparisons.

The difference between these two types of measurement is also often referred to as “norm-referenced” and “domain-referenced” measurement.¹¹

We used GENOVA for conducting the generalizability analyses.¹² As noted, PCSE scores are reported as the percentage of possible points a student achieves in the domain across all eight cases. Since the generalizability analysis is based on case-level data, we conducted the generalizability analysis on the number of points achieved for each case. Since the generalizability coefficients are ratios of the expected values of variance components, the difference in metric did not impact the generalizability coefficients. It did, however, impact the standard error of measurement provided by GENOVA. To avoid this problem, we calculated standard errors of measurement from the observed standard deviation of the domain scores and the generalizability coefficients using a formula provided by Magnusson.¹⁰

We conducted the analysis separately for first- and second-year students. Since there was no easy way to combine the estimated variance components from multiple administrations of the PCSE, we conducted the generalizability analysis on a single administration of the PCSE and used the data from the first administration of the PCSE given in spring semester 2018 for conducting the generalizability analysis.

Repeated Measures Analysis - To assess both cross-sectional growth, longitudinal growth and their interaction, the data from both classes and the four administrations of the PCSE given over fall 2017 and spring 2018 semesters were analyzed using repeated measures ANOVA. The two classes of students formed the design over subjects and the four administrations of the PCSE formed the design over measures. Orthogonal polynomial contrasts were used to assess the shape of the growth curve over the four administrations of the PCSE. The repeated measures analysis and the generation of summary statistics was done using SPSS Version 25. We considered ($p < 0.01$) as statistically significant in the repeated measures analysis.

Human Subject Protection - The student performance data and matriculation class were provided to the researchers in a deidentified format by the Office of Medical Education Research and Development (OMERAD) honest broker. Given the PCSE was administered as a normal part of the SDC student evaluation program and the student data were deidentified by the recognized honest broker within the medical school, the data used in the study are not considered to be human subjects data by the Michigan State University Human Research Protection Program¹³

Results

There was complete data over the four administrations of the PCSE for 183 first-year and 170 second-year students. Table 1 presents the means and standard deviations for each of the six domains in each class over the four administrations of the PCSE. Figure 1 presents the mean performance for each class across the four administrations as graphs. Table 2 provides a summary of the generalizability coefficients and estimated standard errors of measurement for cross-sectional and longitudinal comparisons.

In the repeated measures analysis, the main effect for medical school year was significant ($p < 0.001$). As can be seen in Table 1 and Figure 1, the second-year students outperformed first-year students in all six domains. The main effects for administration (linear, quadratic, and cubic) were also significant ($p < 0.001$). It appears that case difficulty significantly impacted on the change in scores from administration to administration. There was a statistically significant ($p < 0.001$) interaction between year and linear change over administrations for all six domains. The quadratic and cubic components also were significant for the history domain ($p < 0.001$). As can be seen in Figure 1, the gap in performance between the first- and second-year students narrowed across the four administrations for all six domains.

Table 1, Table 2 and Figure 1 about here

Discussion

PCSEs offers many of the benefits of written progress testing while assessing skills that cannot be measured via written examinations. Unfortunately, measurement error is a significant challenge when implementing a PCSE program, particularly for assessing longitudinal growth. We found the generalizability of the PCSE domain scores as our examination is currently configured to be lower than is generally acceptable for high stakes examinations. Physical examination and the post-encounter stations had the highest generalizability coefficients at around 0.50 for cross-sectional comparisons. The generalizability coefficients were the lowest for second-year students in the patient interaction domain and in the safety domain for both first- and second-year students.

The low generalizability coefficients for the patient interaction domain for the second-year students appears to be due mainly to a ceiling effect. Students in the SDC have largely mastered these skills by the end of the first year of the curriculum. At that point, students on average achieve over 90% of the possible points in this domain. The variability in the scores that is left appears to be mostly error variance. This, in of itself, is not necessarily a problem. It means we cannot easily differentiate among second-year students in their ability to communicate with patients because, for the most part, they have mastered this skill domain and what little difference there is in the scores does not replicate from case to case.

There is general agreement that ensuring patient safety and avoiding potential medical errors is a very important focus in medical training. These skills often cannot be adequately assessed through written examinations. We believe our PCSE is the first clinical skills examination to break these skills out as a specific domain that is scored separately from other clinical skills domains. Examples of these skills include proper identification of patient and visitors; infection prevention/control; medication safety; handovers; conflict resolution; team communication; informed consent; risk identification; open disclosure of adverse events; patient and family engagement.

From a curricular standpoint focusing on safety as a clinical skill domain make sense. From a psychometric perspective, it appears that safety, at least as we have conceptualized it in the PCSE, is not a unidimensional construct. This does not mean that safety is any less important. It just brings into

question whether it is possible to treat safety as a unified clinical skill domain. Clearly, more research is needed on how to conceptualize safety. It may turn out that we have defined safety too broadly, and there are multiple unidimensional clusters of skills that we are currently conceptualizing under the broad rubric of safety.

Measurement error in classical test theory is assumed to be random with an expectation of zero.¹⁰ While this source of measurement error remains a significant problem for assessing the performance of individual students using PCSE, it is less of a problem for assessing groups of students for in research and evaluation since the error in individual student scores tends to cancel out when averaged over multiple students in a research or evaluation study. This is not the case for the error associated with measuring longitudinal growth over multiple PCSE administrations using different sets of cases. The difficulty of the cases in different administrations of the PCSE is confounded with student growth in performance, making it difficult to assess longitudinal growth in clinical skills. Unfortunately, assessing longitudinal growth one of the important benefits of progress testing.

The interaction between longitudinal growth and level of training is not directly subject to this type of measurement error. The repeated measures analysis demonstrated that there was a statistically significant interaction between level of training and linear growth in all six domains. As can be seen in Table 1 and Figure 1, the difference in the scores between first- and second-year students narrowed in successive administrations of the PCSE. In other words, first-year students' growth in all six clinical skills domains exceeded the second-year students' growth. It is not clear if this was due to participation in authentic clinical experiences early in their training, other extensive clinical skills training the students received, or the combination of both. Clearly a focus on early clinic training results in rapid growth in clinical skills over the first year of the curriculum. We expect the second-year students' clinical skills are also improving, but due to variability in difficulty among the different cases used in each administration of the examination, it is difficult to measure.

While the generalizability of the PCSE domains is low, the results are consistent with earlier research on clinical skills examinations. Prior studies found similar generalizability coefficients for data gathering skills^{14,15} and communication skills.¹⁶ As noted by Swanson and Norcini¹⁴, clinical skills examinations given in medical schools, like our PCSE, are typically about two hours in length. Extrapolating from the generalizability coefficients observed in these studies and ours, the examination time needed to obtain generalizability coefficients considered appropriate for high stakes decisions would be about eight hours. Given logistic and resources constraints, conducting eight-hour clinical skills examinations in medical schools is not feasible in most situations. Swanson and Norcini¹⁴ suggest focusing on pass/fail decisions and using sequential testing as a means of making more accurate decisions about the competency of students when the reproducibility of clinical skills examinations is limited. Our approach to this problem is similar to what they propose. We have used a standard setting process to set the minimum pass requirements for students for each semester of the curriculum. First- and second-year students can receive a passing grade if they meet these requirements on either or both PCSEs given in a semester. Third- and fourth-year students, who only take the examination once a semester, can take a

makeup examination if they perform below the requirement and pass the semester if they achieve the passing requirements on the makeup examination.

Conclusions

PCSEs provide a standardized methodology for assessing important clinical skills that cannot be evaluated by written examinations. As in previous research on clinical skills examinations, we found that case specificity and random measurement error are major impediments for using PCSE scores to make high stakes decisions about student competency. While the situation is not ideal, this limitation can be addressed by focusing on pass/fail decisions based on rigorous standard setting and using sequential testing to help ensure the accuracy of decisions about student competency.

One of the advantages of progress testing is the ability to assess longitudinal growth in knowledge and skills over a course of training. Unfortunately, the variability in case difficulty among different administrations of the PCSE limited our ability to assess longitudinal growth. Over time, as we develop a pool of cases that have been used previously in the PCSE. Hopefully we will be able to use the data from previous administrations to balance out case difficulty in each administration of the PCSE and be in a better position to assess longitudinal growth in clinical skills.

While measurement error associated with using different cases makes it difficult to directly assess longitudinal growth in clinical skills, we did find that first-year students in the SDC gain basic clinical skills rapidly, narrowing the gap in skill level with their second-year counterparts. This study suggests implementing regular clinical skills training and early authentic clinical experiences results in the rapid growth of clinical skills over the first year of medical school.

Ensuring patient safety and reducing the chances of serious medical errors are important curricular goals in both undergraduate and graduate medical training. Clinical skills examinations offer a standardized means for assessing actions that help ensure patient safety. We defined safety broadly in our examination, and it appears that the safety domain as currently designed is not a unidimensional construct. Further research is needed on how to conceptualize safety and create reliable measures of this important aspect of medical training.

Abbreviations

College of Human Medicine (CHM)

Core Entrustable Professional Activities (EPAs) for Entering Residencies

Office of Medical Education Research and Development (OMERAD)

Objective Structured Clinical Examination (OSCE)

Declarations

Ethics approval - Per request from the Michigan State University's Human Research Protection Program (MSU-HRPP), a designated honest broker is used to deidentify curricular and student evaluation data collected as a normal part of the medical school's educational programs. By the MSU-HRPP's determination, these data are not considered human subject data. Documentation concerning the honest broker program can be found at <https://omerad.msu.edu/research/honest-broker-for-educational-scholarship>.

Consent for publication—Not relevant for this paper.

Availability of data and material –The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests - The authors declare that they have no competing interests.

Funding - The authors received no financial support for the research, authorship, and/or publication of this paper.

Authors' contribution - HF and ME oversee the administration of the PCSE, case development and scoring. DS performed the analyses and wrote the original draft. CC and LW provided advice on the psychometric and statistical approaches. All authors provided advice and support to the project, edited and approved the manuscript.

Acknowledgements—Ann Taft maintains the student evaluation database used in the study and acted as the honest broker, creating a deidentified research dataset for use in the study. The authors greatly appreciate her help.

References

- 1.Albanese MC, Case S M. Progress testing: critical analysis and suggested practices. *Advances in Health Sci Educ Theory Practice* 2016;21(1): 221–234
- 2.Blake JM, Norman GR, Keane DR, Mueller B, Cunningham J, Didyk N. Introducing progress testing in McMaster University problem-based medical curriculum: Psychometric properties and effect on learning. *Acad Med* 1996;71(9):1002–7
- 3.Pugh, DT, Humphrey-Murto, Wood, TJ. The OSCE progress test—Measuring clinical skill development over residency training. *Med Teach* 2016;38(2):168–173.

- 4.Association of American Medical Colleges. (2014) Core Entrustable Professional Activities for Entering Residency. <http://members.aamc.org/eweb/upload/Core%20EPA%20Curriculum%20Dev%20Guide.pdf> Accessed 2017–06–04
- 5.Gold JG, DeMuth RH, Mavis B, Wagner D. Progress testing 2.0: clinical skills meet necessary science. Med Educ Online, 2015;20, 27769 - <http://dx.doi.org/10.3402/meo.v20.27769>
- 6.DeMuth, RH, Gold JG, Mavis BE, Wagner DP. (2017) Progress on a New Kind of Progress Test: Assessing Medical Students' Clinical Skills. Acad Med. 2017;93(5):724–728, May 2018 doi: <http://dx.doi.org/10.1097/ACM.0000000000001982>
- 7.Harden RM, Stevenson M, Downie WW, Wilson GM. (1975) Assessment of clinical competence using objective structured examination. Br Med J 1975;1:447 doi: <https://doi.org/10.1136/bmj.1.5955.447>
- 8.Cronbach LJ, Gleser GC, Nanda, H, Rajaratnam J. The dependability of behavioral measurement: Theory of generalizability for scores and profiles. Wiley, New York 1972.
- 9.Brennan RL. Generalizability Theory. Springer-Verlag, New York 2001.
- 10.Magnusson D. Test Theory. Addison-Westly, Reading, MA 1967.
- 11.Brennan RL, Kane MT. Signal/Noise ratios for domain-referenced tests. Psychometrika 1977;42(4):610- 25.
- 12.Crick J.E. & Brennan R. L. GENOVA Suite Programs. <https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs#GENOVA> Accessed 2019–02–15
- 13.Honest Broker for Educational Scholarship. Office of Medical Education Research and Development. <https://omerad.msu.edu/research/honest-broker-for-educational-scholarship> Accessed 2019–02–07
- 14.Swanson DB, Norcini JJ. Factors influencing the reproducibility of tests using standardized patients. Teaching and Learning in Medicine 1989;1(3):158–166.
- 15.Clauser BE, Balog K, Harik P, Mee J, Kahraman N. A Multivariate Generalizability Analysis of History-Taking and Physical Examination Scores from the USMLE Step 2 Clinical Skills Examination. Acad Med 2009 Oct;84(10 Suppl):S86–9. doi: 10.1097/ACM.0b013e3181b36fda.
- 16.Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the objective structured clinical examination format: reliability and generalizability. Med Educ. 1996 Jan;30(1):38–43.

Tables

Table 1
Percentage of Ratings/Checklist Items Correct by Domain and Administration

Matriculation Year		Fall 1 2017	Fall2 2017	Spring 1 2018	Spring 2 2018
Interactive Skills					
First Year	Mean	85.48	84.93	91.01	92.32
	SD	10.62	7.51	6.08	4.81
Second Year	Mean	91.52	91.08	91.96	92.88
	SD	7.70	5.80	4.67	4.65
History					
First Year	Mean	31.28	55.50	68.14	64.52
	SD	11.76	10.75	8.13	8.13
Second Year	Mean	63.76	74.98	73.85	70.44
	SD	10.47	5.78	8.22	7.03
Physical Examination					
First Year	Mean	24.01	38.16	40.26	51.32
	SD	7.94	11.71	11.07	10.80
Second Year	Mean	52.01	61.06	58.79	60.81
	SD	10.37	9.15	9.74	9.49
Counseling					
First Year	Mean	57.23	62.02	62.22	61.39
	SD	11.37	11.38	12.24	10.21
Second Year	Mean	73.26	71.48	68.15	64.28
	SD	10.75	9.13	8.67	8.77
Safety					
First Year	Mean	72.41	45.90	72.30	83.66
	SD	9.72	8.42	9.47	6.75
Second Year	Mean	86.96	54.94	77.81	85.68
	SD	5.53	6.49	9.02	6.83
Post Encounter					
First Year	Mean	37.26	29.64	56.10	46.04
	SD	8.20	5.79	7.20	6.91
Second Year	Mean	57.49	47.28	68.21	55.59
	SD	7.23	8.63	6.97	6.60

Table 2**Generalizability Coefficients and Standard Errors of Measurement¹**

Skill Domain	Level	Observed St. Dev.	Generalizability		Standard Error	
			Norm	Domain	Norm	Domain
			Referenced		Referenced	
Patient Interaction	MS-1	6.08	0.39	0.35	4.75	4.90
	MS-2	4.67	0.08	0.07	4.48	4.50
Hypothesis Driven History	MS-1	8.13	0.27	0.17	6.95	7.41
	MS-2	7.03	0.26	0.16	6.05	6.45
Physical Examination	MS-1	11.07	0.54	0.43	7.51	8.36
	MS-2	9.74	0.45	0.36	7.22	7.79
Counseling	MS-1	12.24	0.44	0.20	9.16	10.95
	MS-2	8.67	0.27	0.08	7.41	8.32
Safety	MS-1	9.47	0.08	0.04	9.08	9.28
	MS-2	9.02	0.04	0.02	8.84	8.93
Post Encounter	MS-1	7.20	0.51	0.23	5.04	6.32
	MS-2	6.97	0.43	0.19	5.26	6.28

¹The generalizability coefficients, observed standard deviations and the estimated standard errors are based on first administration of the PCSE Spring Term 2018.

Figures

Figure 1

Growth in Six Clinical Skills Domains Among First- and Second-Year Medical Students