A report on the piloting of a novel computer-based medical case simulation for teaching and formative assessment of diagnostic laboratory testing

Clarence D. Kreiter¹*, Thomas Haugen¹, Timothy Leaven¹, Christopher Goerdt¹, Nancy Rosenthal¹, William C. McGaghie² and Fred Dee¹

¹Department of Pathology, University of Iowa Carver College of Medicine, Iowa City, IA, USA; ²Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Objectives: Insufficient attention has been given to how information from computer-based clinical case simulations is presented, collected, and scored. Research is needed on how best to design such simulations to acquire valid performance assessment data that can act as useful feedback for educational applications. This report describes a study of a new simulation format with design features aimed at improving both its formative assessment feedback and educational function.

Methods: Case simulation software (LabCAPS) was developed to target a highly focused and well-defined measurement goal with a response format that allowed objective scoring. Data from an eight-case computer-based performance assessment administered in a pilot study to 13 second-year medical students was analyzed using classical test theory and generalizability analysis. In addition, a similar analysis was conducted on an administration in a less controlled setting, but to a much large sample (n = 143), within a clinical course that utilized two random case subsets from a library of 18 cases.

Results: Classical test theory case-level item analysis of the pilot assessment yielded an average case discrimination of 0.37, and all eight cases were positively discriminating (range = 0.11–0.56). Classical test theory coefficient alpha and the decision study showed the eight-case performance assessment to have an observed reliability of $\sigma = G = 0.70$. The decision study further demonstrated that a $G = 0.80$ could be attained with approximately 3 h and 15 min of testing. The less-controlled educational application within a large medical class produced a somewhat lower reliability for eight cases ($G = 0.53$). Students gave high ratings to the logic of the simulation interface, its educational value, and to the fidelity of the tasks.

Conclusions: LabCAPS software shows the potential to provide formative assessment of medical students’ skill at diagnostic test ordering and to provide valid feedback to learners. The perceived fidelity of the performance tasks and the statistical reliability findings support the validity of using the automated scores for formative assessment and learning. LabCAPS cases appear well designed for use as a scored assignment, for stimulating discussions in small group educational settings, for self-assessment, and for independent learning. Extension of the more highly controlled pilot assessment study with a larger sample will be needed to confirm its reliability in other assessment applications.

Keywords: computer-based simulation; clinical skills assessment; formative assessment; laboratory medicine; performance assessment

Received: 17 September 2010; Revised: 17 September 2010; Accepted: 17 September 2010; Published: 14 January 2011

Recent technological and computational advances have allowed the development of realistic computerized case management teaching instru-
cases, little research has been conducted to help understand the most efficient design features and how best to provide valid formative feedback to learners. This lack of knowledge was noted in a recent critical literature review of computer-based virtual patients by Cook and Triola, who point out that few if any studies have rigorously explored design features of these computer-based simulations (4). For computer-based case management simulations to achieve their full potential, research is needed on how presentation and formative scoring methods impact learning and assessment outcomes (5, 6). Although case fidelity is an important feature in medical case simulation, other design features may also play a more important role in learning and assessment.

A new simulation format (LabCAPS) was developed that incorporates three important design features. First, the cases address well-defined and highly focused learning and measurement objectives that include four subsets (phases) of the clinical skills needed to deliver competent patient care (Fig. 1). Although all four phases outlined in Fig. 1 are engaged by the student and can provide useful feedback within these case simulations, the second phase was the primary focus of this study and involved the reasoning skills that underlie diagnostic laboratory test ordering to manage a clinical case. The second design feature simulated an actual test ordering checklist to elicit student test ordering responses. The level of cueing used in this response format conformed to cueing in the actual practice of ordering diagnostic tests. The checklist, closely resembling that used by clinicians (Fig. 2), presents 315 available tests with a *find function* to allow selection of tests from all the available options. Given that this response format is consistent with how tests are actually ordered in clinical practice, the semi-cued checklist format was viewed to have a positive impact on case fidelity. The third modification, made possible by the focused nature of the cued response environment, was the use of an objective automated ‘correct vs. incorrect’ method of scoring student responses. Unlike other simulations that record and attempt to score by rating a much larger collections of actions taken within branching clinical case encounters, this new performance assessment relies on more structured indicators of students’ decisions with little branching within a scored phase.

This study reports on the statistical analyses of data collected from the diagnostic test ordering portion of a pilot administration of eight newly developed LabCAPS case simulations. It also examines how the simulations perform as an educational tool when score results provide formative feedback to promote learning in an educational environment that is less structured than the piloted performance assessment context.

**Materials and methods**

**Description of simulation**

LabCAPS is created in a Perl-scripted MySQL database structure as diagrammed in Fig. 1. Upon engaging the simulation and being presented with a clinical vignette, students prioritize and finalize their diagnostic hypotheses from an extended checklist and are then given an expert’s prioritization of diagnostic hypotheses. Hence, before ordering tests, students are provided with expert prioritization of differential diagnoses to consider and the test ordering phase of the simulation involves attempting to confirm or refute differential diagnoses. Although the examinees’ ability to generate correct differential diagnosis from the vignette can be scored using this software, the cases investigated here are at a very basic level and the main goal of this pilot was to measure diagnostic test

---

**Fig. 1.** Flow chart of simulated patient workup.
ordering. Asking examinees to generate case differential diagnoses was done to assure that they were engaged in cognitive problem-solving pursuant to case workup.

Next, the student is asked to select tests from a menu of 315 possible options available in the diagnostic laboratory (Fig. 2). All cases used the same 315-item diagnostic test checklist, which includes an extensive list of clinical laboratory tests and a smaller number of diagnostic clinical and radiological procedures. During case development, all 315 results are initially set at default normal within the database and the case developer changed case-specific tests to an abnormal result as appropriate. As further outlined in Fig. 1, after students submitted orders and viewed the results, they were allowed to return to the menu and order the follow-up tests necessary to confirm or refute diagnoses. They could use as many follow-up test ordering encounters as needed to reach a final diagnosis and treatment plan. After completion of the test ordering phase, students are presented with the diagnostic tests recommended by experts before progressing to the final phases of the simulation (diagnosis and treatment). After diagnosis and treatment are finished, a summary page appears that compares each student’s workup with that of an expert. This is followed by an interpretation page that provides information to the student including: an overview of the patient’s disease process, a test ordering strategy for evaluating the diagnostic hypotheses, and a discussion of the treatment of the patient’s disease.

Although the component of LabCAPS evaluated here assessed only diagnostic test ordering, an important feature of this simulation software is that it offers the potential to acquire information on four independent aspects of performance within each case (diagnostic hypotheses, diagnostic test ordering, diagnosis, and treatment). As shown in Fig. 1, this is possible because after generating differential diagnoses, each student is given expert recommended differential diagnoses. This allows the students to overcome poor performance in the first phase of the simulation. For example, initially considering an incorrect differential diagnoses list will not penalize performance during the second phase (test ordering). Similarly, after ordering tests, students are given the expert recommended list of tests before entering the third phase (diagnosis). Finally, the student is given the expert’s diagnosis before entering the treatment phase (phase 4). This allows each phase of the simulation to assess an independent aspect of performance, which may yield a more generalizable result. By removing dependence between sections of the simulation, both the

![Fig. 2. Example of screen used to order tests.](image-url)
measurement and the educational function of the simulation may be enhanced.

**Student populations**

For the pilot test administration, the examinee sample was 13 second-year medical students who had recently completed an academic unit on hematology at the University of Kansas as part of their medical school training. Student participation was voluntary and each student/examinee was paid US$200 for studying the preparatory information and completing the eight cases. The examinees’ scores on the simulation did not influence their medical schools grades. An evaluation of the software as an educational tool was also performed for 143 second-year medical student learners at the University of Iowa as part of a class assignment. The pilot study was approved by the Iowa Institutional Review Board and the evaluation of the student learners component was part of an ongoing quality control effort aimed at improving teaching and assessment within the college.

**Pilot test administration and scoring**

All examinees were first given a practice case and then subsequently administered eight LabCAPS cases that covered basic anemia-related topics. Cases required a median time of approximately 14 min to complete. Scoring of the pilot administration was accomplished by comparing examinee responses to a key generated by expert consensus. In this case, two hematopathology experts at the University of Iowa Carver College of Medicine were asked to engage the case and indicate the ‘correct’ responses. Consensus between the two experts was high and the few disagreements between experts were resolved through discussion. Experts initially agreed on 112 (82%) (56 identical tests ordered) of the 137 total tests independently ordered between them across the eight cases. After discussion of the remaining 25 tests, consensus was reached to score 8 of the 25 as expert-recommended tests or ‘correct’ responses. This resulted in a key with 64 (56 + 8) correct tests, for an average of 8 (range 6–10) orders per case. Potential responses for each case were selected from a standard checklist of 315 diagnostic tests that was identical for all cases. As displayed in Fig. 2, the checklist format used in the simulation closely resembles those that clinicians employ in actual clinical practice.

Table 1 displays the point assignment strategy used to score each item (ordered test) within a case. A score of 1 was awarded for examinees selecting a test indicated by the experts. A score of −0.25 was assigned for ordering a test that was not keyed as correct and for not ordering a test that was keyed as correct. The total case score was simply the sum of all items scored within a case. As Table 1 indicates, the scoring incorporated a penalty for ordering too many tests. All recorded item responses represented either ordering or not ordering a particular test. The number of scored tests for the eight cases ranged from 13 to 44 (Table 2) with an average of 26.1 scored items (tests) per case and a total of 209 items (tests) across the eight cases. Phases 1, 3, and 4 (diagnostic hypothesis, diagnosis, and treatment, respectively) were not scored in the pilot study.

**Evaluation of LabCAPS as an educational tool**

In a second-year medical pathology course at the University of Iowa, 143 student learners were assigned a subset of 2–3 LabCAPS cases (from a library of 18 cases) to engage before attending small group sessions and then presenting and discussing their workup to the group. Case content included autoimmune disease and hematologic disorders, and the cases were assigned pursuant to course lectures on these subjects. The scores

| Case                              | Number of items | Correlation with total discrimination |
|-----------------------------------|-----------------|--------------------------------------|
| Case 1: Anemia of chronic disease | 31              | 0.11                                 |
| Case 2: Pernicious anemia         | 22              | 0.42                                 |
| Case 3: Hereditary spherocytosis   | 22              | 0.33                                 |
| Case 4: Aplastic anemia           | 44              | 0.48*                                |
| Case 5: Iron deficiency anemia    | 35              | 0.49*                                |
| Case 6: Hemoglobin S-C disease    | 19              | 0.56*                                |
| Case 7: Anemia of renal failure   | 23              | 0.15                                 |
| Case 8: Beta-thalassemia minor    | 13              | 0.44                                 |
| Average                           | 26.1            | 0.37                                 |
| Classical test theory reliability | $\alpha=0.76$  | $\alpha=0.70$                        |
| Cronbach’s alpha                  | (n = 209)       | (n = 8)                              |

*p < 0.10.
generated by this assignment were used to evaluate the across case reliability of the formative feedback.

**Classical test theory analysis**

After initial scoring of the diagnostic test ordering for the pilot test using the rules displayed in Table 1, classical test theory case level item analysis was performed using the across case total score to calculate each case’s discrimination index (case-total score correlation with case score removed from the total). A Cronbach’s alpha coefficient of reliability was calculated first using case total scores and then items across cases. When item dependence is present, an item level estimate of reliability will reflect this item dependence within a case and will generate a positively biased estimate of the ‘true’ reliability. The comparison of case level alpha (the ‘true’ observed reliability) with the item level alpha (inflated by case dependence) provided an indication of within case item dependence for these performance data.

**Generalizability analysis**

The generalizability (G) study model describing how the testing data was collected and analyzed for this pilot exam is a persons (p)-crossed-with-cases (c) \([p \times c]\) random model with no missing data. As cases contained a variable number of scored items, case scores were standardized to a common mean and standard deviation (mean = 10, SD = 2) to give each case an approximately equal weight. GENOVA® software was utilized to estimate the variance components and to conduct the D study.

For evaluating scores generated within the class assignment context, student learners were assigned different sets of cases to work up. For the scores generated in this less formal context, a case-nested-within-person \([c p]\) G study model was employed. All students engaged at least two cases. For the analysis, balanced samples of two random cases per student were selected for each of 143 students and the mean variance components across six random samples was used in a D study.

**Pilot study examinee perception survey**

A survey containing six questions (Table 4) related to examinees’ perception of the assessment was administered after completion of the eight cases in the pilot study. Summarizing and reporting of survey responses is performed with descriptive statistics.

**Results**

**Classical test theory and G study analysis**

The results of the classical test theory analysis of the pilot administration are summarized in Table 2. All eight case totals displayed a positive correlation with the overall across case total and displayed a mean discriminations index of 0.37 (range 0.11–0.56). Each discrimination index was calculated with the individual case removed from the total. Using this calculation method, three of the eight discrimination indexes were statistically significant at \(p < 0.10\). An alpha reliability coefficient of \(\alpha = 0.70\) for the eight-case test summary score was obtained. The alpha coefficient using items across cases was \(\alpha = 0.76\), indicating a moderate level of item dependence within cases.

The variance component estimates from the generalizability study are displayed in Table 3. The person variance component accounted for 22% of the total variance and the person-by-case (pc) variance component accounted for 78% of the total variance. Case variance was zero, reflecting the standardization of case scores. A decision study setting the number of cases to eight yielded a \(G = 0.70\) and was by definition equal to the observed alpha classical test theory estimate for the eight cases (Table 2). The D study additionally indicated that a total of 14 cases would be required to achieve \(G = 0.80\). With each case requiring an average of approximately 14 min, 3.27 h of testing time would be required to reach a reliability of 0.80.

For the G study of the responses collected from 143 students during the educational assignment, a G coefficient of \(G = 0.22\) for two cases was obtained. The D study demonstrated that with eight cases the expected G coefficient would be \(G = 0.53\). This was somewhat lower than the pilot test value of \(G = 0.70\), however, the assignment context in which the cases were administered was much less standardized and structured than the pilot test administration.

**Student perception survey**

Examinees in the pilot assessment gave consistently high ratings to their experience with the LabCAPS simulations. Table 4 displays the mean ratings and standard deviations for the six questions presented for the examinees to rate their perceptions of the eight case simulations. All examinees were asked to respond to each question. Two question responses were left unanswered. Each examinee ‘agreed’ or ‘strongly agreed’ that the

| Effect   | VC  | df | %  | SE  |
|----------|-----|----|----|-----|
| Person   | 0.8935 | 12 | 22 | 0.488 |
| Case     | (0.00) | 7  | 0  | 0.036 |
| PC       | 3.106  | 84 | 78 | 0.773 |

*Eight cases: \(G_p = 0.70\); 14 cases: \(G_{14} = 0.80\).
Note: \(N = 13\); VC, variance component; SE, standard error.
Table 4. Examinee perceptions

| Mean | SD  |
|------|-----|
| 1. Navigation is intuitive | 4.3 | 0.63 |
| 2. Case workup proceeds in logical order | 4.6 | 0.65 |
| 3. The simulations have relevance and fidelity to clinical practice | 4.2 | 0.83 |
| 4. LabCAPS content is appropriate for your level of training | 3.8 | 0.90 |
| 5. Interpretation and feedback sections are useful | 4.6 | 0.50 |
| 6. LabCAPS has the potential to enhance learning in your school's curriculum | 4.8 | 0.45 |

Note: 1 = Strongly disagree, 2 = Disagree, 3 = No opinion, 4 = Agree, 5 = Strongly agree; N = 13.

navigation was intuitive and that the cases proceeded in a logical order. Responses to items 5 and 6 (questions displayed in Table 4) indicate students believed the simulations made a positive educational contribution. The lowest rating was for item 4 and it appears that some students may have found the performance assessment too challenging.

Additionally the 143 learners in the class assignment context gave similarly high ratings to the overall teaching effectiveness of the LabCAPS exercises (mean 4.3; excellent = 5 and very good = 4).

Discussion

A validity argument for the scores generated by this simulation should be evaluated by considering both their generalizability and how well the scores reflect the target construct. In this case, we are interested in giving valid feedback and in generalizing to an examinee’s performance in ordering actual diagnostic laboratory tests in response to real clinical encounters. If our scores reliably summarize the quality of clinical decisions on the tasks presented, and are not a reflection of construct irrelevant variance, the similarity between the performance assessment task and the ‘real world’ task is a useful source of evidence to support a validity argument (7). On this simulation, examinees judged the case tasks to be similar to actual practice. In addition, examinees found the interface easy to navigate, suggesting that our summary scores were unlikely to contain a significant ‘technological aptitude’ factor.

Reliability is critically important to any validity argument. Although many simulations appear on the surface to require actions similar to those required in actual clinical practice, their inability to generate reliable scores mitigates their validity. This assessment yielded moderate levels of reliability and generalizability ($\alpha = G = 0.70$) with less than two hours of structured testing. Reliabilities are likely to significantly improve with the development of scoring methods for the remaining phases of the simulation. The pilot outcome suggests the potential to generate valid scores. Since this study employed a modest sample size, follow-up studies are needed with larger groups of examinees. However, the scoring and analysis methods employed in this study did not capitalize on sample dependent response characteristics. For example, although we conducted a case-level analysis of discriminations, we did not delete or change the scoring of any case based upon observed case discrimination findings. In addition, although the standard error for persons (Table 3) is large due to the small sample of examinees, the reliability obtained in the unstandardized presentation to the much larger sample of 143 students in the educational assignment situation, provided additional evidence that the simulation scores are capable of reflecting sizable person variance.

Future research using larger samples should focus on confirming answer key objectivity by assessing the expert consensus on a wide range of cases. Also, it will be important to further assess item dependence within cases. Depending on the magnitude of item dependence, item scaling using item-response theory should be investigated for its usefulness in scoring. The LabCAPS case simulation allows a scoring format, similar to what was reported here for the test ordering phase of the simulation, to be applied to the other three phases of the simulation (diagnostic hypothesis, diagnosis, and treatment) and further research is needed to develop and validate this aspect of the software. These additional scores, adding more information, are likely to significantly increase the reliability of the total performance score.

A valid and reliable automated score for all phases of the simulation will be essential for both educational feedback and for the formative assessment function of these simulations. This study suggests valid automated scores can be generated with this simulation design. In addition, because the simulation program, the case content, and the case editor can be shared freely with other medical schools and medical education organizations, the development of these cases has the potential to deliver economical and effective simulation-based medical education and performance assessment. The automated scoring requires little or no additional faculty time and these automated scores add a mechanism for accountability when using these simulations as part of course assignments.

Although the implementation reported in this study was with second-year (pre-clinical) medical students, the LabCAPS case simulations, with appropriate level case content, may also be appropriate for the clinical medical students, residents, and a variety of allied health students.
Conclusion
LabCAPS shows the potential to provide formative assessment of medical students’ skill at diagnostic test ordering and to provide valid feedback to learners. The perceived fidelity of the performance tasks and the statistical reliability findings support the validity of using the scores for formative assessment. LabCAPS cases appear well designed for use as a scored assignment, for stimulating discussions in small group educational settings, for self-assessment, and for independent learning. Extension of the formative pilot assessment study to a larger sample will be needed to confirm its reliability in other assessment applications.

Acknowledgements
Development and evaluation of LabCAPS is supported by a grant from the National Library of Medicine. We thank Dr. Ivan Damjanov, who assisted us in recruiting the University of Kansas students. The technology for this study was supported in part by grants from the National Library of Medicine.

Conflict of interest and funding
The authors received funding from the National Library of Medicine to conduct this study.

References
1. Dillon GF, Clauser BE. Computer-delivered patient simulations in the United States Medical Licensure Examination (USMLE). Simul Healthcare 2009; 4: 30-4.
2. Turner MK, Simon SR, Facemyer KC, Newhall LM, Vearch TL. Web-based learning versus standardized patients for teaching clinical diagnosis: a randomized, controlled, crossover trial. Teach Learn Medicine 2006; 18: 208–14.
3. Gursel Y, Musal B, Boulet J. Computer-based performance assessment in an undergraduate clinical clerkship. Med Educ 2009; 43: 483–4.
4. Cook DA, Triola MM. Virtual patients: a critical review and proposed next steps. Med Educ 2009; 43: 303–11.
5. Kreiter CD, Bergus G. The validity of performance-based measures of clinical reasoning and alternative approaches. Med Educ 2009; 43: 319–24.
6. Wainer H, Mee J. On assessing the quality of physicians’ clinical judgment. Eval Health Prof 2004; 27: 369–82.
7. Kane MT. Validation. In: Brennan RL, ed. Educational measurement. 4th ed. Westport, CT: American Council on Education and Praeger Publishers; 2006, pp. 17–64.

*Clarence D. Kreiter
Office of Consultation and Research in Medical Education
University of Iowa Carver College of Medicine
1204 Medical Education Building
Iowa City, IA 52242, USA
Email: clarence-kreiter@uiowa.edu