Is it live or is it Memorex? Student oral examinations and the use of video for additional scoring

Kenneth Burchard, M.D.[a,b,*], Horace Henriques, M.D.[a,b], Daniel Walsh, M.D.[a,b], Debra Ludington[b], Pamela A. Rowland, Ph.D.[a], Donald S. Likosky, Ph.D.[a,b]

[a]*Dartmouth Medical School, Hanover, NH, USA*
[b]*Department of Surgery, Dartmouth-Hitchcock Medical Center, One Medical Center Dr., Lebanon, NH 03756, USA*

Abstract

Background: Oral examination interrater consistency has been questioned, supporting the use of at least paired examiners and consensus grading. The scheduling flexibility of video recording allows more examiners to score performances. The purpose of this study was to compare live performance with video performance scores to assess interrater differences and the effect on grading.
Methods: A total of 283 consecutive, structured, videotaped 30-minute examinations were reviewed. A 5-point Likert scale ranked problem solving (2 cases), verbal skills, and nonverbal skills. Nonparametric paired analyses tested for differences.
Results: Live performance scores were higher for verbal and nonverbal skills and total scores. Video performance scores were higher for problem solving for the first presented case. The largest difference (.29 Likert point) was in nonverbal skills.
Conclusions: The minor yet statistical differences in several scores did not actually impact student grades. The use of video recording is sufficiently reliable to be continued and advocated. © 2007 Excerpta Medica Inc. All rights reserved.

Keywords: Oral examination; Interrater reliability; Clerkship; Student; Competency; Video recording

In 2003, the Accreditation Council for Graduate Medical Education outlined general competencies that are required for achievement during residency training and the accreditation process. This has influenced medical school faculty and several schools have developed similar concepts of capabilities that all medical students should acquire (Table 1). The oral examination is a tool that effectively measures such competencies as clinical problem solving, verbal communication, and nonverbal communication [1]. However, throughout the 20th century the interrater scoring consistency of oral examinations has been questioned, especially when examinees are performing poorly [2–7].

Because interrater difficulties may exist, a more reliable method is to use multiple examiners, similar to the format used by the American Board of Surgery. Unfortunately, the size of a medical school class, the frequency of such examinations (at least several times a year), and the constraints of

surgical practice make a process used by the Board of Surgery impractical for medical school faculty.

Video recording has been used to allow multiple raters to rank performances in a variety of evaluation circumstances, such as physical examinations, Objective Structured Clinical Examinations, and oral examinations [8–10]. For the past decade, the evaluation and grading of students at Dartmouth Medical School has included an oral examination that has been scored by a live reviewer (LR) as well as a video reviewer (VR).

The purpose of this study was to compare live with video oral examination performance scores to determine if this evaluation method is indeed consistent between raters, and to determine if any features of the examination process demand improvement.

Methods

Clerkship grading at Dartmouth Medical School uses a numeric scoring system that ranks ward performance (60% of the grade; maximum score, 45), written examination (20% of the grade; maximum score, 15), and an oral exam-

---

* Corresponding author. Tel.: +1-603-650-7903; fax: +1-603-650-8030.
E-mail address: Kenneth.W.Burchard@hitchcock.org

ination (20% of the grade; maximum score, 15). All scores are rounded to the nearest whole number.

The oral examination is a structured, 30-minute encounter in which 2 case scenarios are presented: the first case scenario is based on scheduled, faculty-led case studies; the second is based on the medical student's patient log. A 5-point Likert scale was used to rank 4 domains: problem solving for each case (2 domains), and verbal and nonverbal skills (1 domain each) (Table 2). The subcomponents of each of these domains (ie, diagnostic approach, management approach, and so forth) were averaged. Therefore, the maximum score for any examination was 20. For grading purposes the domain score for case 1 and case 2 were averaged, resulting in a grade score maximum of 15. This being understood, we chose to separate the case 1 and case 2 average domain scores for the comparisons described later.

A total of 283 consecutive examinations were accrued over 5 years (2001–2005). Four faculty rated performances, with one (K.B.) attending most live evaluations (84%) and a second evaluating (D.W.) most of the video performances (74%). The mean values for the individual domain scores were compared with paired *t* tests, and median values were compared with the Wilcoxon signed-rank test to evaluate the equality of the matched pairs [11]. The differences between scores (VR − LR) was calculated for graphic illustration.

Results

All values are listed as means/medians unless otherwise stated.

Problem solving for case 1

The VR score was higher for problem solving for case 1 (3.75/3.70 vs 3.65/3.66; *P* = .005). Differences were principally within 1 point (Fig. 1).

Problem solving for case 2

There was no difference in problem solving for case 2 (3.83/4.00 vs 3.83/4.00; *P* = .608). Differences were principally within 1 point (Fig. 2).

Table 1
General competencies for medical students

Brown University Medical School
Effective communication
Basic clinical skills
Using basic science to guide therapy
Diagnosis, management, and prevention
Lifelong learning
Self awareness, self-care, and personal growth
The social and community context of health care
Moral reasoning and ethical judgment
Problem solving
Dartmouth Medical School
Medical knowledge
Clinical skills
Communication skills
Professionalism
Practice-based learning and improvement
Systems-based practice

Table 2
Oral examination scoring

I. Problem solving				
Cases 1 and 2				
A. Knowledge about disease process				
Poor				Excellent
1	2	3	4	5
B. Diagnostic approach to disease process				
1	2	3	4	5
C. Management approach to disease process				
1	2	3	4	5
D. Average score for the 3 components _____				
II. Verbal communication skills				
A. Speech disturbances ("and ums," "you know")				
Frequent				Rare
1	2	3	4	5
B. Hesitation				
1	2	3	4	5
C. Level of vocabulary				
Poor				Excellent
1	2	3	4	5
Average score for the 3 components _____				
III. Nonverbal communication				
A. Direct eye contact				
Poor				Excellent
1	2	3	4	5
B. Professional dress				
1	2	3	4	5
C. Distracting behaviors (object manipulation, rocking, leg, arm, hand movements)				
Frequent				Rare
1	2	3	4	5
D. Engagement in conversation (attentive, listens, takes turns)				
Poor				Excellent
1	2	3	4	5
Average score for the 4 components _____				
Total score: (score case 1 + score case 2)/2 + verbal score + nonverbal score				

Verbal skills

The LR score was higher for verbal skills (4.17/4.25 vs 3.99/4.00; *P* = .000). Differences were principally within 1 point (Fig. 3).
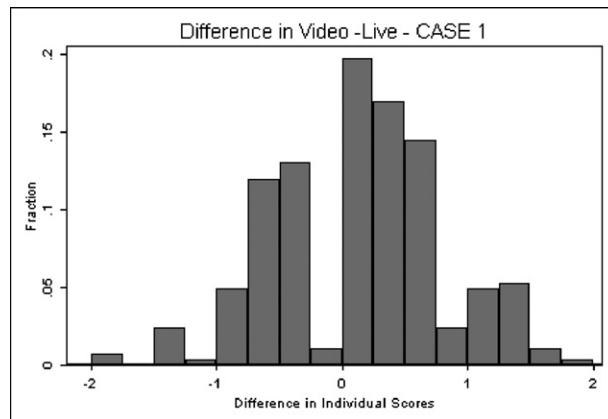

Fig. 1. Differences between VR and LR scores for case 1. The x-axis shows the difference of the individual scores. The y-axis shows the fraction of scores that were in this difference range. Most of the differences were within 1 point, although differences larger than 1 point were more frequent when the video score was higher.