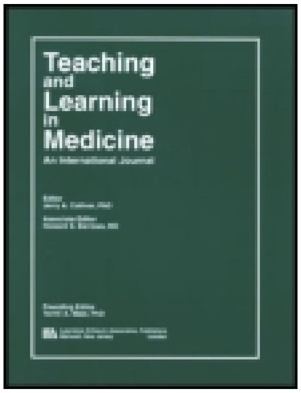

Teaching and Learning in Medicine: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/htlm20

Assessment of Clinical Skills With Standardized Patients: State of the Art Revisited

David B. Swanson [a] & Cees P.M. van der Vleuten [b]

[a] National Board of Medical Examiners , Philadelphia , Pennsylvania , USA

[b] Department of Educational Development and Research , Maastricht University , Maastricht , The Netherlands

Published online: 18 Nov 2013.

To cite this article: David B. Swanson & Cees P.M. van der Vleuten (2013) Assessment of Clinical Skills With Standardized Patients: State of the Art Revisited, Teaching and Learning in Medicine: An International Journal, 25:sup1, S17-S25, DOI: 10.1080/10401334.2013.842916

To link to this article: http://dx.doi.org/10.1080/10401334.2013.842916

Teaching and Learning in Medicine, 25(S1), S17–S25
Copyright © National Board of Medical Examiners®
ISSN: 1040-1334 print / 1532-8015 online
DOI: 10.1080/10401334.2013.842916

Routledge
Taylor & Francis Group

Assessment of Clinical Skills With Standardized Patients: State of the Art Revisited

David B. Swanson
National Board of Medical Examiners, Philadelphia, Pennsylvania, USA

Cees P.M. van der Vleuten
Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands

We would like to begin by thanking the Editors of *Teaching and Learning in Medicine* (*TLM*) for the invitation to contribute to this special issue of *Teaching and Learning in Medicine* commemorating its 25th year. In late 1989 (after a rejection from another journal because it was too long), we submitted a manuscript entitled "Assessment of Clinical Skills with Standardized Patients: State of the Art" to *TLM*. After benefiting from helpful reviews from Jerry Colliver and others, the published version appeared in *TLM* about 1 year later.[1] In this article, we address, in a limited way, what has (and has not) changed since the original publication.

The most obvious change, of course, is the sheer scale on which standardized-patient-based (SP-based) examinations and objective structured clinical examinations (OSCEs[*]) are now conducted. In addition to their use in national licensure and qualifying examinations in several countries (e.g., Australia, Canada, South Korea, Switzerland, Taiwan, the United Kingdom, and the United States), many schools now conduct large-scale OSCEs used in conjunction with making intramural promotion and graduation decisions, as well as exams in individual clerkships. Back in 1990, it was possible to do a reasonably thorough review of measurement research on OSCEs. That is certainly true no longer: Many dozen studies are published each year. Several large-scale systematic reviews of specific OSCE-related issues have appeared in the last few years.[2–4] In addition, Brian Hodges has written an interesting full-length socio-history[5] (in a Foucauldian historical research tradition) about the evolution and spread of the OSCE in the 30 years following its invention in 1975, and other historical treatments have been published as well.[6]

As a consequence, we do not attempt a comprehensive review of research on OSCEs and SPs in this article. Instead, we focus on a small, somewhat eclectic collection of methodological and substantive topics where we think we have something useful to say. Most were mentioned in the original *TLM* publication, and, for many, substantial progress has been made, often with new issues surfacing as a result of research done in the intervening years.

REPRODUCIBILITY (PRECISION) OF SCORES ON SP-BASED TESTS

In the original article, we said that "relatively little measurement error is introduced by training multiple SPs to play the same patient role, as long as examinees are randomly assigned to SPs." In retrospect, this statement may well have been incorrect; regardless, it would certainly have been useful if we had emphasized and explained the importance of the last portion of the sentence. It is now very common for OSCEs to be administered in multiple sessions that are spread out geographically, chronologically, or both. As well as raising security issues, this results in the addition of both random and systematic measurement error to scores and causes problems in the estimation of the reproducibility (precision) of scores and in placing scores from different sessions on the same scale (equating).[7] These are particularly important considerations when examination results are used to make high-stakes decisions. There are also potential ramifications for research on alternate approaches to scoring. Both are discussed in more detail next.

Use of Generalizability Theory to Analyze the Reproducibility of Scores

The original article emphasized the importance of using generalizability theory to analyze the reproducibility of scores because multiple sources of measurement error are present in

Correspondence may be sent to David B. Swanson, National Board of Medical Examiners, 3750 Market Street, Philadelphia, Pennsylvania, PA 19104, USA. E-mail: Dswanson@nbme.org

[*]We still prefer to use the term "OSCE" to refer to a test administration format in which examinees rotate through a series of stations, some of which may involve standardized patients, but it's clear that particular terminology battle was lost years ago.

OSCE scores, and the conceptual and computational tools it provides yield more and better information. An introduction to use of generalizability theory in conjunction with assessment methods commonly used in medical education is provided in a recent AMEE Guide[8] and in an older review specifically related to SP-based testing and inter-relationships (and trade-offs) among precision, equating, and security.[7]

To illustrate, it is helpful to have a concrete, real-world example to discuss. In a large-scale OSCE at a UK medical school, roughly 200 students completed a 25-station OSCE (predominantly 5-minute stations, plus a few 10-minute stations; this depended on station type). The OSCE was given in eight sessions over 2 days at two sites with morning and afternoon sessions at each site on each day. SPs were used at most stations, and an observing rater (mostly doctors) documented student performance on a checklist tailored to station content. Different SPs and raters were used for each station within a session, with little overlap across the eight sessions. All stations were scored on a percentage-of-possible-points basis, so that scores on individual stations varied between 0 and 100. Total scores were calculated as the mean of the station scores, and overall mean performance was roughly 70%.

An individual session can be viewed as a “crossed” design in which each student completed all 25 stations, with the same SP and rater involved throughout the session. Results of a generalizability analysis conducted for one of the sessions are reported in the top half of Table 1. Variance components are reported for persons (examinees), station difficulty, and error. The Persons variance component indicates the amount of “true (universe) score” variance; the square root of this value is a standard deviation that indicates how much examinee performance would vary if it could be measured without error. The value of 4.72 is on the percent-of-possible-points metric; the score range is typically 5 to 6 times the SD, so the overall true spread in scores is around 25%. The Stations variance component can be interpreted as an index of the true variation in station means; it reflects differences in inherent station difficulty, SP difficulty, and rater stringency, which are all confounded in this design. The standard deviation of 10.38 is on a percent-of-possible-points metric; it indicates that true stations means vary substantially (perhaps by 50 points), a fairly typical result. The largest variance component (145.837; 52.8% of the total variance) is for error. It also includes several confounded sources of variance, one of which is commonly termed “case (or content) specificity” in the medical problem-solving literature.[9,10] The table also provides several norm- and domain-referenced indices of reproducibility for a variety of test lengths. Domain-referenced indices treat differences in station difficulty as a component of measurement error; norm-referenced indices do not.

For this simple persons-by-stations design, coefficient alpha is computationally equal to the generalizability coefficient in the table. At the test length actually used, the values of the generalizability coefficient and coefficient alpha will match exactly because the latter is a special case of a generalizability coefficient when (a) only a single source of measurement error is present, (b) the examinees all take the same test form, and (c) the desired score interpretation is norm-referenced. For this design, the formula typically used to estimate the standard error of measurement from coefficient alpha also yields exactly the same value as the relative standard error of measurement in generalizability theory. In addition, the Spearman-Brown formula can be used to project what coefficient alpha would be at other test lengths; these results will also be identical.

In addition to providing domain-referenced indices, generalizability analysis yields additional information that is quite useful. Dividing the station difficulty standard deviation of 10.38 by five (the square root of the test length of 25 stations) indicates that the difficulty of 25-station forms constructed at random would have a standard deviation of approximately 2%. Multiplying this value by 5 provides a sense for the expected variation in form difficulty that might be seen across years of test administration. The variation in form difficulty is expected to be roughly twice as large as the person’s standard deviation. This is not good news, as it indicates that scores from different forms are likely to be on quite different scales, making it desirable to use an equating procedure to adjust scores from different forms so that they are on the same scale.[7]

Reproducibility of Scores: Two Thought Experiments

For large-scale OSCEs, two key conceptual issues are (a) the definition of a “form” and (b) the extent to which variation in form difficulty is caused by sampling of stations (which may be the same for all sessions during a test administration) versus sampling of SPs/raters (which are typically not the same). Partitioning these sources of variance is difficult with the “disconnected” test administration designs[*] often used for large-scale OSCEs.

To see why, consider the following thought experiment. Suppose that there are two additional sessions with station scores absolutely identical to those in the session just discussed, except the examinees were systematically better (by 4 points) in one session and systematically worse (by 4 points) in the other. Combining all of the scores into a single persons-by-stations data set and rerunning the generalizability analysis produces the variance components in the lower half of Table 1. Inspection of the variance components indicates that they are fairly similar, except the variance component for persons has increased by almost 50%. This makes sense, because the major change in the data set is that examinees are more variable as a result of systematically adding and subtracting 4 points to every station score. Similarly, the generalizability and dependability coefficients also increased substantially, though the standard errors are essentially unchanged.

[*] Disconnected designs occur when subgroups of examinees take test forms composed of non-overlapping sets of cases, SPs, and/or raters. See Swanson et al. for more detailed discussion of disconnected designs and approaches to connecting them.[7]

TABLE 1
Variance components for thought experiment

Original Variance Components

Variance Component	Estimate	% of Total	Square Root of Var Comp	Interpretation
Var(Persons)	22.274	8.1%	4.72	True differences in person ability
Var(Stations)	107.842	39.1%	10.38	True differences in station difficulty (multiple components)
Var(Error)	145.837	52.8%	12.08	Interaction, error (includes content specificity)
Total	275.953	100.0%	16.61	Total

For a test length of 25 stations:
Generalizability Coefficient = 22.274/(22.274 + (145.837/25)) = 0.792
Relative Standard Error of Measurement = SQRT(145.837/25) = 2.415
Dependability Coefficient = 22.274/(22.274 + *(107.842/25)* + (145.837/25)) = 0.687
Absolute Standard Error of Measurement = SQRT(*(107.842/25)* + (145.837/25)) = 3.185

No. of Stations in the OSCE	Norm-Referenced Interpretation		Domain-Referenced Interpretation	
	Generalizability Coefficient	Relative SEM (% of Points Scale)	Dependability Coefficient	Absolute SEM (% of Points Scale)
5	0.433	5.401	0.305	7.123
10	0.604	3.819	0.468	5.037
15	0.696	3.118	0.568	4.112
20	0.753	2.700	0.637	3.561
25	0.792	2.415	0.687	3.185

Variance Components for Thought Experiment

Variance Component	Estimate	% of Total	Square Root of Var Comp	Interpretation
Var(Persons)	32.547	11.4%	5.70	True differences in person ability
Var(Stations)	111.234	38.9%	10.55	True Differences in station difficulty (multiple components)
Var(Error)	142.445	49.8%	11.94	Interaction, error (includes content specificity)
TOTAL	286.226	100.0%	16.92	Total

No. of Stations in the OSCE	Norm-Referenced Interpretation		Domain-Referenced Interpretation	
	Generalizability Coefficient	Relative SEM (% of Points Scale)	Dependability Coefficient	Absolute SEM (% of Points Scale)
5	0.533	5.338	0.391	7.123
10	0.696	3.774	0.562	5.037
15	0.774	3.082	0.658	4.112
20	0.820	2.669	0.720	3.561
25	0.851	2.387	0.762	3.185

Note. OSCE = objective structured clinical examinations; SEM = standard error of measurement.

For the second thought experiment, again imagine there are two additional sessions with station scores absolutely identical to those in the session described in the previous section except, this time, the raters were systematically more lenient/dovish (by 4 points) in one session and systematically more stringent/hawkish (by 4 points) in the other. Generalizability analysis of these three sessions produces the same result as for the first thought experiment—as it should, because the data sets for the two thought experiments are completely identical. Examinee ability and station/rater/SP difficulty are completely confounded when different examinees and raters/SPs are used in each session (a disconnected design). As a consequence, a Persons × Items analysis of variance that ignores session structure does not produce interpretable estimates of variance components.* As the thought experiments make clear, an increase in the person variance component can be due to a true increase person variability (Thought Experiment 1), a systematic increase in error variance (Thought Experiment 2), or a combination of the two.

To avoid making unwarranted assumptions about the source of differences among sessions, it is desirable to run the analysis of variance so that sessions are explicitly represented in the design. Viewed across sessions, the correct design is (Persons : Sessions) × (Stations : Sessions), rather than Persons × Stations. This approach does not make any (tacit) assumptions about the equivalence of the sessions or the cause of performance differences across sessions, and it pools information across sessions in estimation of variance components. In addition, the analysis produces a variance component for sessions. If examinees and SPs/raters are randomly assigned to sessions, the value of the sessions variance component will often be zero (or negative). This does *not* mean that the sessions are equivalent and forms for each session are equal in difficulty; it simply means that the sessions are no more different than expected given random assignment. As illustrated in the previous section, it is still likely that forms vary in difficulty, and this variation can be substantial relative to examinee variance, particularly if the number of stations per test form is small. As discussed in a later section, ignoring the session structure can also produce misleading results in studies of alternate scoring approaches if the magnitude of the stations variance component, relative to other variance components, varies across approaches due to, for example, differences in subjectivity.

Related estimation problems occur even when an OSCE is administered in a single section if SPs and/or raters rotate in and out of stations when the design is (mis-)specified as persons by stations. For example, if teams of two SPs and two raters rotate in and out for some or all stations and this is ignored in analysis by using a Persons × Stations design, the estimated difficulty of the station will reflect the average across the teams. This will generally have the effect of reducing the magnitude of the stations variance component in generalizability analysis and increasing the error variance component. If the rotation is done so that the resulting design is connected, a more informative approach is to run the analysis so that the identities of the SPs and raters (nested in stations) are "visible" to the statistical package, yielding separate estimates for SP- and rater-related sources of measurement error.

Use of Station Scores versus Item Scores in Estimation of Reproducibility (Precision)

Occasionally, we see "station reliabilities" reported in the literature; these are calculated using the items on the checklist or global rating scale within a station as if they were independent items sampled from a larger domain of items. Except in a few situations discussed in the next paragraph, we do not think that this makes sense conceptually. The items on a checklist are not like a set of multiple-choice questions (MCQs) that are randomly sampled from a domain of items that could have appeared on a test. Items appear on a checklist because they are clinically appropriate given the overall case context, providing a basis for scoring performance.[11] The situation is similar for a set of rating scale items associated with a station: These are chosen because they are relevant for scoring examinee performance, spanning the aspects of performance to be scored. In both instances, it makes little sense to calculate a within-station reliability index based on item scores.

Exceptions occur when an individual station consists of independent assessments of a specific skill. For example, an OSCE could include an electrocardiogram (ECG)-reading station in which examinees are asked to read a series of unrelated ECGs. In this instance, it may make sense to calculate an index of the reproducibility of scores on the ECGs, particularly if an ECG score will be reported to examinees. This index will be interpretable in terms of the expected correlation between scores if the ECG-reading station was repeated with a different (randomly parallel) sample of ECGs spanning similar content.

On a related topic, some OSCEs make use of the same rating scale repeatedly on all or a subset of stations; this probably occurs most commonly for ratings of communication skills. In such situations, if a subscore for, for example, communication skills is calculated by averaging the associated ratings from different stations, it is quite sensible to calculate a reliability index for that subscore. This will be interpretable in terms of the expected correlation between communication subscores if communication skills were rated on a different (randomly parallel) sample of stations of the same size covering similar content. This is a different situation from the "within-station reliability" previously discussed.

USE OF CHECKLISTS VERSUS RATING SCALES IN SCORING EXAMINEE PERFORMANCE

A number of studies have explored the psychometric properties of using checklists versus rating scales in scoring

*This will be true whether assignment of examinees, SPs, and raters to sessions is random or systematic (e.g., raters observe at the clinical sites at which they normally work): The resulting variance components will not be readily interpretable if the underlying session structure is ignored in the analysis.

examinee performance.[12–18] Checklists commonly consist of detailed items of performance like "washes hands," "introduces oneself to the patient," or "measures blood pressure with a cuff of appropriate size." Checklist items are typically scored as "done" or "not done." Rating scales are more broadly formulated items that require more holistic judgment of the associated item. For example, interpersonal skills could be one item followed by a Likert scale that expresses a quality range of the performance. Although there is probably an underlying continuum[17] ranging from most holistic to most analytic (atomistic), checklist items tend to capture whether some action occurred, whereas rating scales require the interpretation of actions. Checklists were originally proposed for their objectivity and the potential bias that can result when judgment is required. This paved the way for use of lay assessors (e.g., SPs), particularly in North American implementations of SP-based testing.

Research on use of checklists versus rating scales in station scoring has provided some seemingly clear messages. In terms of the reproducibility of scores, global rating scales do just as well and often better than checklists.[12–14] A recent review reported a mean "inter-station reliability" (which appeared to be equivalent to the mean interstation correlation) of .207 versus .168.[2] Projected to a test length of 15 stations, this yields total test reliabilities of roughly 0.80 and 0.75, respectively—a practically important difference in reproducibility. Although interstation reliability tends to be superior for rating scales, interrater reliability tends to be lower than for checklists. Apparently, although greater subjectivity is present for holistic judgments, those subjective judgments appear to capture skills that are more generalizable across stations. Similar results are seen whether ratings are provided by medically qualified raters or SPs.[3] This somewhat paradoxical outcome has also been reported for other assessment methods.[19,20] This research suggests that gathering a large set of subjective judgments from different judges across different contexts can result in a reliable overall outcome.[21] This result has provided a portion of the rationale for moving toward less standardized assessments embedded in the workplace.[22]

From an educational perspective, some have warned for a potential negative effect of checklists on learning.[19,23,24] Checklists may reward trivialization, thoroughness, and rote-learning—students memorizing checklists leading to mechanical performance—while rating scales may reward learning for understanding and encourage integration of knowledge and skills.

Do we therefore generally recommend the universal adoption of global rating scales? The answer is no. The use of checklists or rating scales depends on the purpose of the assessment. In situations where mastery is important checklists are appropriate. For example, at junior levels of training it might be appropriate to use checklists. This may lead to more thorough preparation for and mastery of specific clinical skills like individual physical exam maneuvers. At more senior levels, mastery may also be important. Some clinical and procedural skills (e.g., resuscitation, intubation) may require precision in execution of prescribed steps, and checklists may then be appropriate. In assessment situations in which holistic performance is important, probably at more senior levels of training, rating scales are advisable.

There have been some recurring methodological problems in research on alternate scoring approaches. In many studies comparing checklists and rating scales, the same observer has completed both the checklist and the rating scale, and the psychometric characteristics of the two are then compared. Although such studies may provide useful information about the psychometric characteristics of checklists versus rating scales when *both* are completed, they do not necessarily provide information about the likely results when one or the other (but not both) is completed. Intuitively, it seems likely that checklists provide raters with information about what the station authors considered important in scoring; this may well serve to improve the quality of the ratings, as well as promoting greater comparability in use of the rating scale across raters grading performance on the same station in large-scale OSCEs involving multiple sessions.

There have also been some analytic problems in research reports. It is common for published reports to provide estimates of coefficient alpha for each scoring method. As discussed in an earlier section, this is less than ideal because generalizability analyses are more informative. In particular, for large-scale OSCEs involving multiple sessions, the magnitude of the stations variance component is important, and one might expect that the subjectivity inherent in rating scales could increase the relative magnitude of this source of error variance. Many of the published reports also appear to have used a simple Persons × Stations design in analysis, ignoring session structure. If, in fact, the stations variance component is relatively larger for rating scales than checklists, the observed superior reliability of rating scales may, at least in part, be due to misspecification of the design in statistical analysis of results. Further research should shed more light on this issue.

At this point, we interpret the conclusions of research on checklists versus rating scales to be somewhat permissive. It is sensible for OSCE developers to make a utilitarian choice of the method(s) used to capture and score performance. The original objectivity argument for use of checklists has turned out to be too simplistic to provide a good basis for choice. The skills to be assessed, together with the psychometric and educational consequences, should be considered in choosing a scoring approach. It also seems quite possible that station scoring should consist of a blend of methods, with checklists, in a sense, helping to calibrate raters so that rating scales are used comparably by different raters. We think that additional research on alternate scoring approaches would be worthwhile; such work should seek to make incremental quality improvements in scoring rather than simple comparisons of grossly different methods. As discussed in the next section, the approaches that work best may need to be tailored to the type of individual rating performance and the skills to be rated.

USE OF DOCTORS VERSUS SPS TO RATE EXAMINEE PERFORMANCE

Given the cost and availability of physician examiners, nonphysician lay examiners are often used. In the literature, three categories of nonexpert examiners are found: SPs who rate performance after playing the patient role in the encounter, trained lay examiners observing the encounter, and similarly trained medical students. Although there is some variability across studies (e.g., Zanetti et al.[25]), the majority seems to show that SPs may score the performance as reliably as experts.[26–29] The same holds for lay examiners[30] and for students.[1,32] In a study by Moineau et al., lay examiner scores (medical students) were equally well accepted by examiners (30) and seemed to have similar credibility. There is a long history of "gynecologic teaching associates" and "patient instructors" contributing to medical student education in the United States[32,33] and elsewhere.[34,35] Some might argue whether this is really assessment;[36] see the next section for related discussion. Additional work investigating how best to promote the acceptability and utility of SP feedback may merit further study.

Although one might expect a drop in reliability when rating scales are completed by lay examiners, this does not seem to be the case in practice.[27] Global ratings differentiated better between levels of expertise for physician assessors in one study;[37] in another study, physician ratings did correlate with an external written exam while ratings from lay examiners (SPs) did not.[28] Lay examiners do tend to be more lenient than physicians,[25,26,28] though this is not uniformly true: The opposite was observed in a study of ratings of more expert performance.[37]

Overall, it seems lay examiners can be used in SP-based exams both when checklists and rating scales are used. There are trade-offs. When lay examiners are used, one should be mindful of leniency effects, and physician examiners seem to benefit most from use of global rating scales. The skills to be rated also should play a role. SPs clearly have a reasonable perspective for making judgments about whether examinees are, for example, understandable linguistically; probably this is also true for communication skills more generally. For physical exam skills, with appropriate training, SPs also have access to somatosensory and other perceptual information unavailable to physicians observing exam maneuvers. However, if a station is intended, in part, to look at the sequencing of questions in history taking, that judgment may be more difficult for SPs to make, even with training. Such considerations should be weighed, along with availability and cost trade-offs, in deciding what types of raters are effective for which kinds of stations/skills.

ASSESSMENT FOR LEARNING VERSUS ASSESSMENT OF LEARNING: OSCES IN ASSESSMENT SYSTEMS

Although the psychometric literature tends to focus on test reliability and the validity of inferences from test scores, increasing attention is now being given to the formative value of assessment and the impact on learning outcomes.[38] Rather than optimizing the accuracy of scores and decisions based on those scores, the latter is more concerned with optimizing learning outcomes in the longer term. This shift in perspective is sometimes characterized as a shift from assessment *of* learning to assessment *for* learning.[39–43] The format of feedback from assessments and impact on learning outcomes receive central attention.

Another recent movement is to look beyond research on individual assessment methods toward design of full assessment systems. The reasoning is that an individual assessment method will not meet all needs, but the purposeful combination of methods in a program of assessment may.[19,44] The intent is for a carefully designed program of assessment, consisting of a purposeful mix of assessment components, to promote learning, as well as support decision making and program evaluation/improvement. Recently, a theoretical model for such a program has been proposed,[45] identifying the elements that need be realized to serve the multiple purposes of assessment (i.e., learning, decision making, program evaluation, public accountability). Next we briefly discuss both feedback and programmatic assessment in relation to SP-based testing.

In studying the impact of feedback on learning outcomes, a distinction is made among preassessment effects, pure assessment effects, and postassessment effects.[46] It is now well-established that instruction connected to testing results in better learning outcomes than instruction without testing.[47] This is termed the *pure assessment effect* and has also been documented for SP-based testing.[48] The *preassessment effect* refers to the way that students prepare for an assessment. For example, in one of the early studies, introduction of an OSCE resulted in students spending increased time in the wards and less time reading in the library.[49] Negative effects have been reported as well,[50] particularly in relation to memorizing checklists leading to artificial improvements in performance by trainees. Similarly, SP-based tests that follow the traditional OSCE format using very short stations scored with detailed checklists may lead to fragmentation and trivialization of learning.

This preparation behavior can be identified and explained after the fact[42] but is more difficult to predict because of dependencies on specific local circumstances. In general terms, to prevent the adverse effects on learning outcomes caused by reductionist trivialization of the tasks posed in stations, a good strategy is to make the challenges posed by individual stations as realistic as possible, requiring integration of the skills needed in the real clinical situation; this often requires more time per station. Use of rating scales focused holistically on key features of performance should also both improve assessment of learning and encourage better study behavior by students. A *postassessment effect* is achieved through feedback on performance on the assessment. This feedback can take many different forms. One can structure the testing time within a station to permit direct face-to-face feedback, copies of the completed checklist and/or rating form can be provided, debriefing sessions can be held after the exam, and subscores based on subsets of stations (or skills measured repeatedly across stations) can be calculated and reported in performance profiles for individual examinees. Consistent with the feedback literature,[51] the simple provision of feedback with SP-based testing is no guarantee of the actual use of the feedback. Harrison et al. developed an online feedback report system and documented the use of the feedback.[52] He came to the saddening conclusion that those who needed the feedback most were least inclined to use it. It appears best for feedback to be scaffolded with educational interventions that require learners to analyze the information provided.[53,54]

SP-based testing is best viewed in the context of an overall program of assessment. Thinking from a programmatic perspective[45] may have consequences for the design and implementation of SP-based testing. The focus in programmatic assessment is on information gathering. This means that an SP-based test should always provide meaningful feedback. A well-designed score reporting system is useful, permitting results to be broken down to relevant domains, competencies, or entrustable professional activities.[55] Having examiners provide narrative feedback during or after each encounter is another potentially powerful source of feedback.

To allow for a meaningful aggregation across different assessment methods, the scoring (or narrative information) should be structured according to some consistent overarching framework (e.g., performance domains or competencies). Information from the SP-based tests may then be triangulated against other sources of assessment information that are different in format but focused on similar skills (e.g., ratings of performance during clerkships). Using the assessment information in a programmatic way may also lead to breaking up SP-based tests into multiple parts administered at different points of time. This may be particularly relevant for mastery of skills that can then be "signed off" as mastery is achieved. Of course, the shorter the test becomes, the lower the reliability. If the purpose of testing is formative and high-stakes decisions will not be based results, this does not matter much. In programmatic assessment, stakes and number of data points are related: High-stake decisions can be based on the combination of assessment information across multiple occasions and methods.[56]

CONCLUSIONS AND RECOMMENDATIONS

As noted at the beginning of this article, much has changed since our *TLM* article first appeared. We find the growth in SP-based testing over the past 25 years to be both astonishing and encouraging: Greater focus on assessment of clinical skills benefits both students and patients. At the same time, some of the methodological recommendations from the original paper still merit attention.

Articles occasionally include a statement along the lines of "OSCEs have been demonstrated to be a reliable and valid method for assessment of . . .." For several reasons, we do *not* think this is appropriate. First, reliability is strongly related to test length (both number of stations and total testing time) and the mix of station formats used. Although some OSCEs are long enough to yield reproducible scores, it is fairly common that tests are too short, and the resulting scores are not reproducible enough to make high-stakes decisions, at least in isolation. Second, it is always desirable to run an appropriate generalizability analysis and calculate indices of reproducibility. A large body of related research indicates that a well-designed OSCE *can* yield reliable scores, but this is worth checking; the results can also be informative for improving the test design and making more efficient use of testing resources. Third, an assessment method is not valid or invalid: Validity is a property of inferences from scores.[11] The same test can be valid for some purposes and not for others, depending upon the nature of the intended inferences from scores.

Published reports are often vague about psychometrically important details of test administration (e.g., whether multiple sessions were conducted, if multiple SPs played the same patient role; how examinees were assigned to raters, SPs, and stations). The format and content of checklists and rating forms and the procedures used for scoring are rarely described in enough detail for the reader to have a good feel for the instrumentation; this makes it difficult for test designers to build on the work of others in improvement of instrumentation. More work is needed in this area, beginning with better descriptions of methods already in use.

Procedures used for reliability estimation are often suboptimal and, occasionally, appear to be wrong. Use of generalizability theory in analysis is absolutely required, because multiple sources of measurement error are commonly present. Variance component estimates should be reported, so that readers can explore alternative uses of testing resources, and researchers can integrate results across studies. The design used in the analysis should accurately reflect the actual conditions of test administration and sources of measurement error present. In particular, if multiple sessions are used with different (or partially overlapping) SPs/raters in each session, it is absolutely essential that these are appropriately reflected in analysis; this is particularly true for studies of alternate scoring methods.

There is growing interest in assessment for learning, the impact of assessment on learning outcomes, alternate approaches to providing assessment feedback, and the design of assessment systems. The literature has (accurately, we think) emphasized the importance of these areas, but relatively little systematic research has been reported—and this is apt to be methodologically challenging to systematically investigate. The educational impact of SP-based tests has been a major factor in their increased use, and more work on optimizing their impact on learning outcomes (in the context of overall system of assessments) seems desirable, particularly given the high costs of SP-based tests.

REFERENCES

1. Van der Vleuten CPM D, Swanson D. Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine* 1990;2:58–76.

2. Patricio MF, Juliao M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Medical Teacher* 2013;35:503–14.
3. Patricio MF. Is the OSCE a reliable and valid tool to assess clinical competence in undergraduate medical education? Manuscript under review.
4. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Medical Education* 2011;45:1181–9.
5. Hodges BD. The objective structured clinical examination: A socio-history. Koln, Germany: Lambert Academic, 2009.
6. Wallace P. Following the threads of an innovation: The history of standardized patients in medical education. *Caduceus-Springfield* 1997;13:5–28.
7. Swanson DB, Clauser BE, Case SM. Clinical skills assessment with standardized patients in high-stakes tests: A framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education* 1999;4:67–106.
8. Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher* 2012;34:960–92.
9. Swanson DB. A measurement framework for performance-based tests. In I Hart, R Harden (Eds.), *Further developments in assessing clinical competence* (pp. 13–45). Montreal: Can-Heal, 1987.
10. Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Science Education* 1996;1:41–67.
11. Kane M. Validation. I. In RL Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: ACE/Praeger, 2006.
12. Van Luijk SJ, Van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In *Current developments in assessing clinical competence* (pp. 357–82). Montreal: Can-Heal, 1992.
13. Cunnington JPW, Neville AJ, Norman GR. The risk of thoroughness: reliability and validity of global ratings in checklists in an OSCE. *Advances in Health Sciences Education* 1996;1:227–33.
14. Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine* 1998;73:993–7.
15. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Academic Medicine* 1999;74:1129–34.
16. Govaerts MJ, van der Vleuten CP, Schuwirth LW. Optimising the reproducibility of a performance-based assessment test in midwifery education. *Advances in Health Sciences Education: Theory and Practice* 2002;7:133–45.
17. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Medical Education* 2003;37:1012–6.
18. Hodges B. Validity and the OSCE. *Medical Teacher* 2003;25:250–4.
19. Van der Vleuten CP, Schuwirth LW. Assessing professional competence: From methods to programmes. *Medical Education* 2005;39:309–17.
20. Van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology* 2010;24:703–19.
21. Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education* 1991;25:110–8.
22. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher* 2007;29:855–71.
23. Boursicot K, Etheridge L, Setna Z, Sturrock A, Ker J, Smee S, et al. Performance in assessment: Consensus statement and recommendations from the Ottawa conference. *Medical Teacher* 2011;33:370–83.
24. Gormley G. Summative OSCEs in undergraduate medical education. *Ulster Medical Journal* 2011;80:127–32.
25. Zanetti M, Keller L, Mazor K, Carlin M, Alper E, Hatem D, et al. Using standardized patients to assess professionalism: A generalizability study. *Teaching and Learning in Medicine* 2010;22:274–9.
26. Heine N, Garman K, Wallace P, Bartos R, Richards A. An analysis of standardised patient checklist errors and their effect on student scores. *Medical Education* 2003;37:99–104.
27. Han JJ, Kreiter CD, Park H, Ferguson KJ. An experimental comparison of rater performance on an SP-based clinical skills exam. *Teaching and Learning in Medicine* 2006;18:304–9.
28. McLaughlin K, Gregor L, Jones A, Coderre S. Can standardized patients replace physicians as OSCE examiners? *BMC Medical Education* 2006;6:12.
29. Park J, Ko J, Kim S, Yoo H. Faculty observer and standardized patient accuracy in recording examinees' behaviors using checklists in the clinical performance examination. *Korean Journal of Medical Education* 2009;21:287–97.
30. Moineau G, Power B, Pion AMJ, Wood TJ, Humphrey-Murto S. Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. *Medical Education* 2011;45:183–91.
31. Chenot JF, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, et al. Can student tutors act as examiners in an objective structured clinical examination? *Medical Education* 2007;41:1032–8.
32. Kretzschmar RM. Evolution of the Gynecology Teaching Associate: An education specialist. *American Journal of Obstetrics and Gynecology* 1978;131:367.
33. Stillman P, Ruggill J, Rutala P, Sabers D. Patient instructors as teachers and evaluators. *Journal of Medical Education* 1980;55.
34. Cleland JA, Abe K, Rethans J-J. The use of simulated patients in medical education: AMEE Guide No 42 1. *Medical Teacher* 2009;31: 477–86.
35. Bokken L, Rethans JJ, Scherpbier AJ, van der Vleuten CP. Strengths and weaknesses of simulated and real patients in the teaching of skills to medical students: A review. *Simulation in Healthcare* 2008;3:161–9.
36. Bokken L, Linssen T, Scherpbier A, van der Vleuten C, Rethans JJ. Feedback by simulated patients in undergraduate medical education: a systematic review of the literature. *Medical Education* 2009;43:202–10.
37. Regehr G, Freeman R, Robb A, Missiha N, Heisey R. OSCE performance evaluations made by standardized patients: Comparing checklist and global rating scores. *Academic Medicine* 1999;74(10 Suppl):S135–7.
38. Wiliam D. Education: The meanings and consequences of educational assessments. *Critical Quarterly* 2000;42:105–27.
39. Boud D, Falchikov N. Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education* 2006;31:399–413.
40. Boud D, Falchikov N. *Rethinking assessment in higher education: Learning for the longer term*. New York: Routledge, 2007.
41. Klenowski V. Assessment for learning in the accountability era: Queensland, Australia. *Studies in Educational Evaluation* 2011;37:78–83.
42. Cilliers FJ, Schuwirth LW, Herman N, Adendorff HJ, van der Vleuten CP. A model of the pre-assessment learning effects of summative assessment in medical education. *Advances in Health Sciences Education: Theory and Practice* 2012;17:39–53.
43. Wiliam D. What is assessment for learning? *Studies in Educational Evaluation* 2011;37:3–14.
44. Dijkstra J, Galbraith R, Hodges BD, McAvoy PA, McCrorie P, Southgate LJ, et al. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education* 2012;12: 20.
45. van der Vleuten CP, Schuwirth LW, Driessen EW, Dijkstra J, Tigelaar D, Baartman LK, et al. A model for programmatic assessment fit for purpose. *Medical Teacher* 2012;34:205–14.
46. Dochy F, Segers M, Gijbels D, Struyven K. Assessment engineering: Breaking down barriers between teaching and learning, and assessment. In D Boud, N Falchikov (Eds.), Rethinking assessment in higher education: Learning for the longer term (pp. 87–100). Oxford: Routledge, 2007.
47. Karpicke JD, Roediger HL, 3rd. The critical importance of retrieval for learning. *Science* 2008;319:966–8.
48. Kromann CB, Jensen ML, Ringsted C. The effect of testing on skills learning. *Medical Education* 2009;43:21–7.

49. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983;17:165–71.
50. Van Luijk SJ, Van der Vleuten CPMvd, Van Schelven RM. Observer and student opinions about skills tests. In W Bender, RJ Hiemstra, AJJA Scherpbier, RP Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 497–502). Groningen: Boekwerk, 1990.
51. Hattie J, Timperley H. The power of feedback. *Review of Educational Research* 2007;77:81–112.
52. Harrison CJ, Konings KD, Molyneux A, Schuwirth LW, Wass V, van der Vleuten CP. Web-based feedback after summative assessment: How do students engage? *Medical Education* 2013;47:734–44.
53. Nicol D. From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education* 2010;35:501–17.
54. Archer JC. State of the science in health professional education: Effective feedback. *Medical Education* 2010;44:101–8.
55. ten Cate O, Scheele F. Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? *Academic Medicine* 2007;82:542–7.
56. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher* 2011;33:478–85.