Validity, Reliability, Feasibility, Acceptability and Educational Impact of Direct Observation of Procedural Skills (DOPS)

Naghma Naeem

ABSTRACT

Direct observation of procedural skills (DOPS) is a new workplace assessment tool. The aim of this narrative review of literature is to summarize the available evidence about the validity, reliability, feasibility, acceptability and educational impact of DOPS. A PubMed database and Google search of the literature on DOPS published from January 2000 to January 2012 was conducted which yielded 30 articles. Thirteen articles were selected for full text reading and review. In the reviewed literature, DOPS was found to be a useful tool for assessment of procedural skills, but further research is required to prove its utility as a workplace based assessment instrument.

Key words: *Clinical competence. Educational assessment. Measurement. Reproducibility of data. Validity. Reliability. Feasibility. Educational Impact. Direct observation of procedural skills.*

INTRODUCTION

Competence forms the foundation of practice in any profession.[1] Competence is defined as "the degree to which an individual can use the knowledge, skills and judgment associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice."[1] Competence is usually inferred from performance.[2] Decisions about professional competence are, therefore, based on observation of the proficiency of trainees performing authentic tasks related to the practice of medicine. Miller provided a simple description of the hierarchical nature of professional competence as a pyramid of increasing performance proficiency.[3] In postgraduate training, assessment of clinical competence is receiving increasing attention.[4-6]

A review study on assessment of procedural skills found that there were no validated methods of procedural performance assessment described in the literature.[7] Many authors have expressed concern over a lack of rigorous testing of procedural skills for doctors in training.[8] Historically, other instruments such as 'Objective Structured Clinical Examination' (OSCE), log-books and supervisor evaluations have been used for the assessment of procedural skills. OSCE leads to assessment of procedural skills in a fragmented manner with no opportunity to observe the examinee carrying out a complete procedure.[9,10] Also, emergency procedures cannot be assessed in OSCE and examinee attitudes are difficult to assess during an OSCE. Another limitation of OSCE is cost and resource intensiveness. The cost of OSCE per examinee ranges from $21 to $100.[11] Similarly logbooks have been criticized as being a measure of expected competence rather than performance. They have also not been found to be reliable. Therefore, the quest for a more reliable and valid tool for workplace based assessment of procedural skills led to the development of DOPS.

The direct observation of procedural skills commonly referred to as DOPS is among the workplace based assessment (WBA) instruments piloted in the United Kingdom as part of the new assessment tools for the "Foundation Programme."[5] Literature on the foundation Programme describes it as a new instrument. In reality, the concept of assessment involving direct observation of procedural skills has existed in a non-structured format for a long-time.

The use of DOPS is most prevalent amongst surgical residents due to the higher frequency of procedures performed by them, but less frequent in internal medicine and general practice. DOPS is the observation and evaluation of a procedural skill performed by a trainee on a real patient. The procedural skills assessed using DOPS may range from simple and common to complex. The key features of DOPS include assessment of procedural skills, evaluation of a specific patient encounter, performance of procedure on actual patient, immediate feedback on performance.[6]

DOPS is not widely used in undergraduate medical education or for assessing working doctors except by the Royal Australasian College of Physicians as part of its maintenance of professional standards program since 1994.[6]

King Saud University Chair for Medical Education Research and Development, Riyadh, KSA.

Correspondence: Dr. Naghma Naeem, King Saud University Chair for Medical Education Research and Development, College of Medicine, King Saud University, P.O. Box 2925, Riyadh 11461, Kingdom of Saudi Arabia.
E-mail: naghma.naeem@gmail.com

Received July 05, 2011; accepted March 09, 2012.

DOPS is a learner centered assessment which promotes self directed learning by allowing the trainee to identify learning needs, select the procedure and the assessor and schedule the assessment. Each DOPS represents a different procedure and trainees sample across the core skills identified in the curriculum by the end of the year. The assessor's evaluation is recorded on a structured form either a checklist of defined tasks, a global rating scale, or a combination of both. Assessors can provide post-encounter feedback based on ratings and narrative comments. DOPS offers opportunity for teaching, supervision and feedback.

DOPS is increasingly being used to assess the competencies of residents; however, there is a dearth of literature regarding the utility of DOPS as an assessment instrument.[12]

The objective of this study was to determine the validity, reliability, feasibility, acceptability and educational impact of workplace based instrument DOPS through a systematic narrative review of literature.

METHODOLOGY

A search of the PubMed database and Google for papers on DOPS instruments published between January 2000 to January 2012 was conducted. The search aimed to identify papers related to validity, reliability, feasibility, acceptability and educational impact of DOPS instruments. For this search we used the following search terms:

Clinical competence (medical subject heading [MeSH] term and text word) OR educational measurement (MeSH term and text word) OR educational measurements (text word) OR clinical skills (text word) AND medical students (MeSH term and text word) OR clinical clerkship (MeSH term and text word) OR internship and residency (MeSH terms) OR internship (text word) OR residency (text word) OR medical education (MeSH term and text word) OR preceptorship (MeSH term and text word) AND observation (text word) OR observe (text word) OR observed (text word) OR reproducibility of results OR feasibility studies OR psychometrics OR evaluation studies (MeSH term, publication type and text word) OR validation studies (publication type and text word) AND direct observation of procedural skills or DOPS.

In addition, the reference lists of the included articles for relevant literature were manually searched.

Inclusion criteria were that the instrument was used by medical/health professionals to assess directly observed performance, in authentic patient encounters, in a post-graduate or undergraduate medical/nursing programme. Also included were review studies including systematic reviews and meta-analysis. Exclusion criteria were instrument being used for peer, patient or self-assessment and no availability of abstract.

The researcher selected and judged articles based on the research question and the inclusion and exclusion criteria. Each article was analyzed to determine whether validity, reliability, feasibility, acceptability and educational impact of DOPS were addressed. The final selection was made after reading the full text of the selected articles.

RESULTS

The initial search yielded 484 articles, out of these 30 articles met the inclusion and exclusion criteria. After reading the abstract, 13 articles were selected and reviewed.[13-25] A flow diagram of literature search and article selection is provided in Figure 1. A summary of reviewed articles and main findings related to validity, reliability, feasibility is provided in Table I.

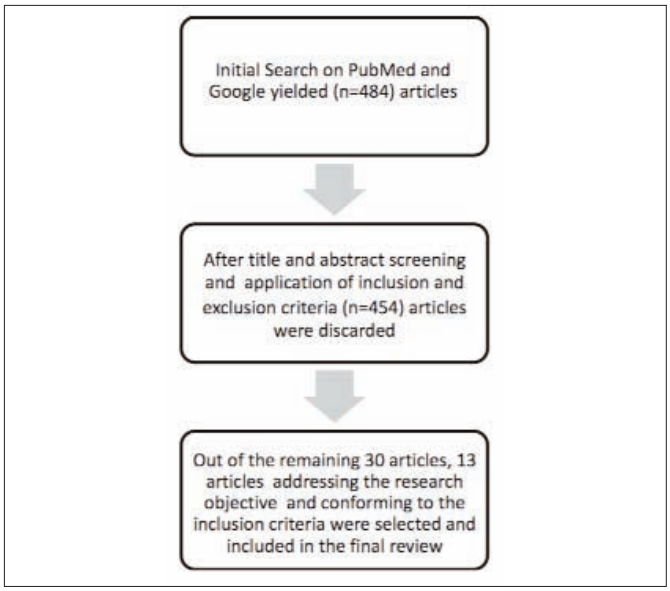

Figure 1: Flow chart of article search and selection strategy.

DISCUSSION

Validity: Validity is a unitary multifaceted concept that cannot be measured, but is inferred. Historically, several facets of validity have been described such as face, content, predictive, concurrent, construct, implying that multiple sources of evidence are required to evaluate validity.[26]

DOPS is seen as a high quality instrument as it tests the "DOES" level of the Miller's Pyramid.[3]

Direct observation of an individual's procedural skills has high face validity as trainees are observed in their workplace performing routine procedures on real patients and all the items on the rating scale or checklist are related to the performance of procedural skills. The validity of DOPS is claimed to be high based on perceptions of assessees and assessors, but no hard evidence is available to support this claim.[13,14] Only one study has demonstrated a positive relationship between

Table I: Summary of articles reviewed with main findings.

Authors (year)	Methodology	Utility aspects addressed *					Main findings
		V	R	F	A	EI	
Barton JP, Corbette S, van der Vleuten CP. (2012)	Survey, generalizability analysis and D-study	√	√	√	-	-	G = 0.81. The reliability could be improved by increasing cases or assessors per assessment, but it is currently strong enough and acceptable in terms of cost and practicability. DOPS scores correlated highly with global expert assessment. 73.6% of candidates and 88.1% of assessors thought DOPS was valid or very valid, while 17.3% of candidates thought that it was somewhat valid.
Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. (2011)	Review	√	-	√	√	√	In most studies content validity was not established by accepted scientific method. Little data on feasibility, acceptability, and educational impact was available.
Mitchell C, Bhat S, Herbert A, Baker P. (2011)	Retrospective observational study	√	-	-	-	-	There was no statistical association between mean scores on DOPS and trainees in difficulty and hence the scores do not appear to predict lack of competence.
Yang YY, Lee FY, Hsu HC, Huang CC, Chen JW, Cheng HM, Lee WS, Chuang CL, Chang CC, Huang CC. (2011)	Experimental design	√	√	-	-	-	The small-scale Objective Structured Clinical Examination (OSCE) scores correlated well with the 360-degree evaluation scores ($r = 0.37$, $p < 0.021$). The addition of DOPS scores to small-scale OSCE scores increased its correlation with 360-degree evaluation scores of PGY (1) residents ($r = 0.72$, $p < 0.036$). Further, combination of Internal Medicine-In Training Examination (IM-ITE) score with small-scale OSCE+ DOPS scores markedly enhanced their correlation with 360-degree evaluation scores ($r = 0.85$, $p < 0.016$). The strong correlations between 360-degree evaluation and small-scale OSCE+DOPS+ IM-ITE composited scores suggest that these methods measure the same construct.
Memon MA, Brigden D, Subramanya MS, Memon B. (2010)	Review	-	√	-	-	-	Inter-case variations in DOPS decreases reliability because of poor content sampling and significant variation in case difficulty.
Miller A, Archer J. (2010)	Review	-	-	-	-	√	No hard evidence to support improvement in performance.
Ma IWY, Zalunardo N, Brown M, Pachev G, Hatala R. (2009)	Experimental design	-	√	-	-	-	Internal consistency of DOPS was 0.94. The inter-rater reliability of the overall rating of competence on the DOPS was 0.81.
Kogan J, Holmboe ES, Hauer KE. (2009)	Review	√	-	-	-	√	Strongest validity evidence available for Mini CEX. Less validity and educational impact evidence for DOPS.
Cohen SN, Farrant PB, Taibjee SM. (2009)	Survey research	-	-	√	√	-	Time-consuming and difficult, carried a degree of stress but trainees appreciated feedback.
Shahgheibi Sh, Pooladi A, BahramRezaie M, Farhadifar F, Khatibi R. (2009)	Experimental design	√	-	-	-	√	Students' scores for each skill in the intervention group had significantly improved more than control group ($p=0.000$). Comparing the means of students' averages for all skills before and after intervention of intervention group (49.49 vs 86.03, $p < 0.0001$) with those of control group (49.99 vs 77.43,

Continued on next page.....

Authors (years)	Methodology	Utility aspects addressed *					Main findings
		V	R	F	A	EI	
							p < 0.0001) showed that the intervention group performed significantly better than the control group (36.54 vs. 27.44, p < 0.0001).
Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. (2008)	Survey, generalizability analysis and D-study	√	-	√	-	-	Mean time for DOPS varied according to procedure. In general DOPS required the length of the procedure plus 20-30% of the procedure time for feedback. DOPS scores increase as trainee progresses in training. DOPS scores do not depend upon procedure.
Morris A, Hewitt J, Roberts C. (2006)	Observational study and survey research	-	-	√	-	√	Perception that DOPS improves clinical skills and it will improve future careers. Opportunity to perform some skills more difficult than others.
Chisholm C, Whenmouth L, Daly E, Cordell W, Giles B, Brizedine E. (2004)	Observational study	-	-	√	-	-	Direct observation by faculty ranged from a high of 6% for Emergency Medicine Residents years 2 and 3 in the critical care areas of the Emergency Department (95% CI = 3% to 9%) to a low of 1% (95% CI = 0% to 2%) on internal medicine wards.

* *Key: V = Validity; R = Reliability; F = Feasibility; A = Acceptability; EI = Educational impact.*

DOPS.[23] This is in contrast to the findings of a retrospective observational study which concluded that there is no statistical association between mean scores on DOPS and trainees in difficulty.[15] Concurrent validity of DOPS has also not been established, which is in contrast to another workplace based assessment instrument Mini-CEX for which the validity is supported by strong and significant correlations with other valid assessment instruments.[27] A review of tools for direct observation and assessment of clinical skills of medical trainees also reported that the strongest validity evidence has been established for the Mini Clinical Evaluation Exercise (Mini-CEX).[28]

The validity and authenticity of DOPS has also been challenged on the grounds that the instrument is obtrusive and may lead to much better performance than in real life introducing a bias in measurement. Further, work is required to establish validity evidence about DOPS.

Reliability: Four studies in the current review discussed the reliability of DOPS.[13,16,17,19] There are several reliability issues associated with DOPS. A review identified inter-case variation as one of the factors affecting reliability of DOPS.[17] Case specificity can also lower reliability. However, a recent study reported that DOPS scores did not appear to depend upon the procedure.[23] Another issue is determining the number of procedures that need to be observed to achieve adequate reliability. A generalizability analysis and D-study reported that a trainee should be observed by at least three different assessors observing at least two procedures each to achieve good reliability.[23] A generalizability coefficient (G) of 0.81 showing good reliability has been reported in one of the study despite the assessment being based on two cases.[13] Another study reported the internal consistency of DOPS as 0.94 and inter-rater reliability of the overall rating of competence on DOPS as 0.81 providing evidence of good reliability of DOPS.[19] In contrast, for the Mini-CEX the item-total correlation has been reported to be between 0.7 and 0.8.[28,29] and the inter-item correlation between the six domains of Mini-CEX has been reported as between 0.5 and 0.8.[29]

In general, good reliability with DOPS can be achieved with relatively fewer cases and assessors as compared to MiniCEX.

Feasibility: There are several feasibility issues concerning implementation of DOPS such as finding willing and trained assessors to participate in DOPS.[17] Implementation of DOPS also poses time constraints for the trainers as well as the trainees[18] and resource constraints for administration. On an average the time taken to complete the DOPS was equal to the length of the procedure and an additional 20-30% of the procedure time for providing feedback.[23]

Wilkinson *et al*. concluded that without adequate time and resources the feasibility of DOPS would be significantly reduced.[23] However, van der Vleuten stated that reliable, formal assessment in real clinical situations is achievable;[30] and Hamilton *et al*. also reported that DOPS is feasible since only one assessor is needed for each observation.[31]

The true costs of running DOPS are unknown as none of the reviewed literature has reported any findings. Costs are calculated using assessor and trainee time, and

added time in clinics or theatres as proxy measures.[32] The American Board of Medical Specialties used an instrument similar to DOPS as part of its re-certification procedures; however, its use was discontinued due to high costs.[6]

A study by Morris found that it was easier to arrange DOPS for common procedures but some procedures were not frequently required as such it was difficult to find opportunities to observe these skills.[24] The emergency department and routine operating lists are common places where DOPS can be performed. In addition, continuous faculty presence in the emergency department should facilitate the use of direct observation as an assessment technique.[25] However, contrary to this expectation, it was seen that faculty in Emergency Medicine rarely performed observations despite physical presence.[25]

All the above feasibility and logistic issues need to be acknowledged and addressed to support implementation of DOPS.

Acceptability: Acceptability may be defined in terms of the number of completed assessment forms, average time for completion of assessments and user (assessor and assessee) satisfaction with the assessment tool.[31]

Trainees perceived DOPS as stressful but appreciated the feedback.[21] Trainees generally welcome the opportunity to be observed by someone more experienced than them and to be given immediate feedback. Greaves and Grant surveyed a small group of anaesthetists and found that they felt that the procedural skills of trainees could be accurately assessed by more senior physicians.[32] DOPS forms were completed by only 33% of trainees in a study which highlights a feasibility problem.[23] Interviews with study co-ordinators revealed that lack of time and not lack of acceptability was the main factor preventing completion.[23]

Overall DOPS appears to be acceptable to both examinees and examiners.

Educational Impact: It is perceived that DOPS has a high educational impact as it provides opportunity for continuous developmental feedback from consultants, highlighting areas of strengths and weaknesses and leads to an agreed upon action plan to address developmental needs. Further, it provides opportunity for reflection.[30] In a review of tools of direct observation it was noted that few studies examined educational effects by measuring improvement of clinical skills or the quality of patient care.[18] Other educational outcomes crucial for the evaluation of educational impact such as change in learner behaviour, transfer of skills to workplace and improvement of patient care have also not been investigated.

An observational survey describing the implementation of direct observation of procedural skills, mini-clinical evaluation exercise, and multisource feedback in a London Hospital provides some data on the educational impact of direct observation of procedural skills.[24] The feedback survey returned by pre-registration house officers reported that 70% of respondents felt that direct observation helped to improve clinical skills.[24]

However, literature does not provide clear evidence beyond subjective survey reports that direct observation of procedural skills actually leads to objective performance improvement.[18,24,29] Only one experimental study reported that mean scores for skills in the intervention group of students using DOPS was significantly better than in the control group.[22]

More experimental studies are needed to investigate the positive impact of DOPS on student's learning and skill acquisition, workplace practice and patient health outcomes.

Articles were included in this review based on the research question and inclusion criteria and not based on any standard criteria for quality of research which might influence the results and conclusions drawn.

CONCLUSION

DOPS is a useful tool for assessment of procedural skills, but further research is required to prove its utility as a workplace based assessment instrument. DOPS has good reliability and acceptability. However, objective evidence about its validity, feasibility and educational impact is not currently available.

Acknowledgement: This research was supported by the College of Medicine Research Centre, Deanship of Scientific Research, King Saud University, Riyadh, Kingdom of Saudi Arabia.

REFERENCES

1. Hays R, Davies H, Beard J. Selecting performance assessment methods for experienced physicians. *Medical Educ* 2002; **36**: 910-7.
2. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009; **302**:1316-26.
3. Miller GE. The assessment of clinical skills/competence/ performance. *Acad Med* 1990; **65**:S63.
4. Waas V, Van Der Vleuten CP, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001; **357**:945-9.
5. General Medical Council. Workplace based assessment: a guide for implementation. [Internet]. [cited 2010 Apr 17]. Available from: http://www.gmc-uk.org/Workplace_Based_Assessment.pdf_31300577.pdf
6. CIPHER. Review of work-based assessment methods. Sydney: *Centre for Innovation in Professional Health Education & Research;* 2007.
7. Wilkinson J, Benjamin A, Wade W. Assessing the performance of doctors in training. *BMJ* 2003; **327**:S91-2.

8. Kneebone R, Nestel D, Yadollahi F, Brown R, Nolan C, Durack J. Assessing procedural skills in context: exploring the feasibility of an Integrated Procedural Performance Instrument (IPPI). *Med Educ* 2006; **40**:1105-14.
9. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979; **13**:41-54.
10. Kirby RL, Curry L. Introduction of an objective structured clinical examination (OSCE) to an undergraduate clinical skills programme. *Med Educ* 1982; **16**:362-4.
11. Carpenter JL. Cost analysis of objective structured clinical examinations. *Acad Med* 1995; **70**:828-33.
12. Van der Vleuten CP. The assessment of professional competence: development, research and practical implications. *Adv Health Sci Educ* 1996; **1**:41-67.
13. Barton JR, Corbett S, van der Vleuten CP. English Bowel Cancer Screening Programme and the UK Joint Advisory Group for Gastrointestinal Endoscopy. The validity and reliability of direct observation of procedural skills assessment tool: assessing colonoscopic skills of senior endoscopists. *Gastrointestinal Endoscopy* 2012; **75**:591-7. Epub 2012 Jan 9.
14. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. *Am J Surg* 2011; **202**:469-80.
15. Mitchell C, Bhat S, Herbert A, Baker P. Workplace-based assessments of junior doctors: do scores predict training difficulties? *Med Educ* 2011; **45**:1190-8.
16. Yang YY, Lee FY, Hsu HC, Huang CC, Chen JW, Cheng HM, *et al*. Assessment of first-year post-graduate residents: usefulness of multiple tools. *J Chinese Med Assoc* 2011; **74**:531-8.
17. Memon MA, Brigden D, Subramanya MS, Memon B. Assessing the surgeon's technical skills: analysis of the available tools. *Acad Med* 2010; **85**:869-80.
18. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ* 2010; **341**:c5064.
19. Ma I. Comparing two methods to evaluate technical competence in central venous catheterization on simulators [Internet]. 2012. Available from: http://crmcc.medical.org/meetings/Sep25_View Royal_1030_Ma_ICRE2009.pdf
20. Kogan J, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009; **23**:1316-26.
21. Cohen SN, Farrant PB, Taibjee SM. Assessing the assessments: U.K. dermatology trainees' views of the workplace assessment tools. *Br J Dermatol* 2009; **161**:34-9.
22. Shahgheibi Sh, Pooladi A, BahramRezaie M, Farhadifar F, Khatibi R. Evaluation of the effects of direct observation of procedural skills (DOPS) on clinical externship students' learning level in obstetrics ward of Kurdistan University of Medical Sciences. *J Med Educ* 2009; **13**:29-32.
23. Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008; **42**: 364-73.
24. Morris A, Hewitt J, Roberts C. Practical experience of using directly observed procedures, mini clinical evaluation examinations, and peer observation in preregistration house officer (FY1) trainees. *Postgrad Med J* 2006; **82**:285-8.
25. Chisholm C, Whenmouth L, Daly E, Cordell W, Giles B, Brizedine E. An evaluation of emergency medicine resident interaction time with faculty in different teaching venues. *Acad Emerg Med* 2004; **11**:149-55.
26. Bari V. Direct observation of procedural skills in radiology. *AJR Am J Roentgenol* 2010; **195**:14-8.
27. Pelgrim EA , Kramer AW, Mokkink HG, van den Elsen L, Grol RP, van der Vleuten CP. In-training assessment using direct observation of single-patient encounters: a literature review. *Adv Health Sci Educ Theory Pract* 2011; **16**:131-42.
28. Torre DM, Simpson DE, Elnicki DM, Sebastian JL, Holmboe ES. Feasibility, reliability and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students. *Teach Learn Med* 2007; **19**:271-7.
29. Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Acad Med* 2003; **78**:S33-5.
30. Van Der Vleuten C. Work-based assessment [Internet]. 2009. [cited 2010 Apr 17]. Available from www.fdg.unimaas.nl/./Taiwan/ Work-based%20assessment%20Taiwan.ppt
31. Hamilton K, Coates V, Kelly B, Boore J, Cundell J, Gracey J, *et al*. Performance assessment in health care providers: a critical review of evidence and current practice. *J Nurs Manag* 2007; **15**:773-91.
32. Senta Z, Jha V, Boursicot KA, Roberts TE. Evaluating the utility of workplace-based assessment tools for speciality training. *Best Pract Res Clin Obstet Gynaecol* 2010; **24**:767-82.