Instruments to measure the ability to self-reflect: A systematic review of evidence from workplace and educational settings including health care

Julie Williams (Corresponding author); Julie.Williams@bristol.ac.uk;
Tel no: 0117 342 9610; Bristol Dental School, Faculty of Health Sciences, University of Bristol, Bristol, UK
Tony Ireland; Tony.Ireland@bristol.ac.uk
Bristol Dental School, Faculty of Health Sciences, University of Bristol, Bristol, UK
Sheena Warman; Sheena.Warman@bristol.ac.uk
Bristol Veterinary School, Faculty of Health Sciences, University of Bristol, Bristol, UK
Martin A. Cake; M.Cake@murdoch.edu.au
School of Veterinary and Biomedical Sciences, Murdoch University, Perth, Australia
David Dymock; D.Dymock@bristol.ac.uk
Bristol Dental School, Faculty of Health Sciences, University of Bristol, Bristol, UK
Ellayne Fowler; Ellayne.Fowler@bristol.ac.uk
Centre for Medical Education, University of Bristol, Bristol, UK
Sarah Baillie; Sarah.Baillie@bristol.ac.uk
Bristol Veterinary School, Faculty of Health Sciences, University of Bristol, Bristol, UK;

Acknowledgements

The review team are grateful for the help and guidance of the subject librarian Mrs Cath Borwick with searching the literature for this review. We would also like to thank Miss Ali Lloyd and Miss Nicky Newcombe for clerical support during the early stages of the project and Professor Martyn Sherriff for statistical advice.

Abstract

Introduction: Self-reflection has become recognised as a core skill in dental education, although the ability to self-reflect is valued and measured within several professions. This review appraises the evidence for instruments available to measure the self-reflective ability of adults studying or working within any setting, not just healthcare.

Materials and Methods: A systematic review was conducted of 20 electronic databases (including Medline, ERIC, CINAHL and Business Source Complete) from 1975 to 2017, supplemented by citation searches. Data was extracted from each study and the studies graded against quality indicators by at least two independent reviewers, using a coding sheet. Reviewers completed a utility analysis of the assessment instruments described within included studies, appraising their reported reliability, validity, educational impact, acceptability and cost.

Results: 131 studies met the inclusion criteria. 18 were judged to provide higher quality evidence for the review and three broad types of instrument were identified, namely: rubrics (or scoring guides), self-reported scales and observed behaviour.

Conclusions: Three types of instrument were identified to assess the ability to self-reflect. It was not possible to recommend a single most effective instrument due to under reporting of the criteria necessary for a full utility analysis of each. The use of more than one instrument may therefore be appropriate dependent on the acceptability to the faculty, assessor, student and cost. Future research should report on the utility of assessment instruments and provide guidance on what constitutes thresholds of acceptable or unacceptable ability to self-reflect, and how this should be managed.

Introduction

The ability to self-reflect is an acknowledged core skill for all healthcare professionals, including dentists (1-5). With the emergence of outcomes-based education in healthcare (6, 7), there is an increasing focus on describing a number of wide-ranging competences that need to be assessed prior to registration. This includes continuous, systematic and self-directed reflection on practice, with appropriate action to improve patient care (2, 5, 8, 9). However, assessing and measuring primarily metacognitive processes, such as self-reflection, presents unique challenges, as they cannot by their very nature, be observed directly. Whilst there is understandable controversy regarding whether it is appropriate to view reflective ability as a technical competence that can be robustly assessed (10), the expectation of outcomes-based curricula means that curriculum developers must try to address this issue. An initial search of the literature failed to find a systematic review of instruments available to assess the ability to self-reflect. However, it became clear that cultivating individuals who have the ability to self-reflect is considered essential within a number of disciplines in addition to healthcare, including management and organisational research (11, 12), accountancy (13), the ministry (14), teaching within primary or secondary education (15, 16), social work (17), higher education (18, 19) and leadership (20, 21). There was broad agreement across disciplines that self-reflection is an important skill which is difficult to observe and complex to assess (22-25). Current approaches to reflective practice in the education of healthcare professionals has been challenged as being out of step with the theory of reflection (10), with there being a trend towards ritualistic or utilitarian written reflection (26, 27). A wider search for instruments used to measure self-reflection in disciplines outside healthcare, including an evaluation of the evidence available to support their use, may further inform the assessment of self-reflection within dental education.

To date the assessment of self-reflection within healthcare education has largely focused on the importance and application of reflection (24, 28) rather than the utility of available instruments to measure it. Utility describes, predicts or explains the usefulness of

decision options, for example in the estimation of the value of human resource programmes in business (29). One such model of utility used to evaluate assessments of competency in healthcare has been described by van der Vleuten (30). This model comprises five criteria, namely: reliability, validity, educational impact, acceptability (to stakeholders) and cost (in terms of the resources required, including time). It is recognised that the evaluation of the utility of any assessment tool against a list of predefined criteria such as this, would be helpful in the education of healthcare professionals (31).

An initial appraisal of the literature informed our research question, namely: What instruments are available to measure self-reflection and what is the evidence to support their use? This systematic review aimed to identify which assessment instruments are currently available to measure the ability to self-reflect in either workplace or educational settings for any vocation, not just healthcare; to explore and synthesise currently available evidence relating to the assessment instruments and to promote 'best evidence' approaches to assessment of self-reflection that could be applied to dental education. The utility of each instrument was assessed against five criteria suggested by van der Vleuten (30), namely: reliability, validity, educational impact, acceptability and cost.

Materials and Methods

Twenty electronic databases were searched for the period January 1975-August 2017, which included six core electronic databases, namely: Medline, Embase, CINAHL (Cumulative Index to Nursing and Allied Health Literature), ERIC (Education Resource Information Centre), British Education Index and PsycINFO, and 14 additional electronic databases. The following terms were used in the search strategy (Figure 1) in various combinations in order to address the research question: education, teaching, university (or) college, (or) further education, (or) school; personnel management, staff development, management, educational measurement (including assessment); tool (or) instrument, (or) scale, (or) test. Further terms were truncated (*) to allow for variation of the root word, for

example train*, in-service train*, work*, self-reflect*, self-aware*, self-regulat*. Slight variation of the search strategy was required to match the structure of each database (Appendix 1). Ancestral searches were performed to check for instruments or tools mentioned within citations.

Selection procedure - All citations retrieved were imported to Endnote X7.5 reference management software (Thomson Reuters, Philadelphia) and duplicate citations removed. Firstly, the title and abstract of all identified citations were screened for inclusion against the selection criteria (Table 1) by one reviewer (JCW), with 10% independently screened by a second reviewer (MC). The inter-rater agreement determined using weighted Kappa was acceptable (0.61) (32). Secondly, the full text of articles identified in the first stage were assessed against the selection criteria by one reviewer (JCW) and once again 10% were independently assessed by a second reviewer (MC). The inter-rater agreement for inclusion of studies at the second stage, determined using weighted Kappa, was 0.88. Disagreement on whether to include articles in the review was resolved by discussion between the two reviewers.

The search was intentionally broad to include instruments used in all workplace and educational settings, within and beyond healthcare. Studies of participants aged 16 years or less were excluded as the focus of the review was on instruments that could be applied to undergraduate dental professionals. For sources within healthcare, studies were excluded if the self-reflective activity was undertaken by a patient. Opinion pieces, commentary articles, studies that could not be retrieved and studies without primary data were also excluded.

Data extraction and analysis - A data extraction coding sheet was developed and piloted by three members of the review team to report: the types of assessment instrument, the context of use, who was tested, who rated the quality of self-reflection, details of any other criteria measured at the same time (such as insight), evidence for repeated testing of the same students and the utility of the instrument (reliability, validity, educational impact,

acceptability and cost). Following minor modification, the final electronic coding sheet was created in Microsoft Excel (Appendix 2).

Quality assessment of studies – Each of the 131 studies that met the inclusion criteria was assessed independently by at least two reviewers and the quality of evidence was scored using a five point scale, based on the work of Harden et al. (33) and Cake et al. (34). The scores ranged from 1 (low quality evidence) to 5, (high quality evidence) as shown below:

1 - No clear conclusions can be drawn. Not significant LOW QUALITY

2 - Results ambiguous, but there appears to be a trend

3 - Conclusions can probably be based on the results

4 - Results are clear and very likely to be true

5 - Results are unequivocal HIGH QUALITY

The strengths and weaknesses of each study were noted as comments on the coding sheet. If there was disagreement between the quality scores of two reviewers of two or more points, a third reviewer was asked to review and score the study independently and any discrepancies were discussed until consensus was reached. Only studies which received a score of four or more from two independent reviewers were referred to as higher-scoring studies and the utility of the instruments used in these studies forms the best evidence for this review.

Evidence synthesis - The heterogeneity of the study designs in the higher-scoring studies precluded meta-analysis of the quantitative data. Therefore, a descriptive synthesis was undertaken by one reviewer (JCW) using the comments entered on the coding sheet by all reviewers.

Results

This section is presented in 5 parts, namely: study selection process; methodological characteristics of the higher-scoring studies; instruments used in the higher-scoring studies; synthesis of evidence for each instrument and the reported utility of these instruments.

Study selection process - The PRISMA flow chart of the literature search and selection process is shown in Figure 2. The primary database search yielded 9599 records and a further 51 records, not found in the primary search, were identified for screening following ancestral searching of citations from included studies. After de-duplication, the remaining 3898 records were then screened by title and abstract, leading to a total of 519 full-text articles being retrieved. Of these, four studies could not be accessed and a further 384 failed the inclusion criteria, leaving 131 studies included in the review. The most common reason for exclusion was the absence of an instrument to measure self-reflection (n=212). The country of origin of included studies was the USA (49 studies, 37%), UK (19 studies, 15%), Australia (14 studies, 11%), the Netherlands (12 studies, 9%), Brazil and Canada (5 studies each, 4%), Taiwan (4

studies, 3%), Belgium, Hong Kong, South Africa and Sweden (3 studies each, 2%) and other countries (11 studies, 8%).

Despite searching for studies published after 1975, the earliest included paper was 1988. The majority of studies (98/131) had been published over the decade 2007-2017. As expected from the initial literature search, included studies were drawn from a wide variety of professional fields, namely Medicine or Surgery (60 studies), Nursing (16 studies), Teaching (9 studies), Occupational Therapy or Physical Therapy or Physiotherapy (7 studies), Dentistry, Pharmacy and Psychology (6 studies each), Dental hygiene or therapy (4 studies), Management and Social work (3 studies each), Tennis Athletes, Royal Navy, Veterinary Science (2 studies each), Chaplaincy, Midwifery, Music, Public Health and Counselling (1 study each). The included studies were published in 70 academic journals.

The weighted kappa value for inter-rater agreement on the quality of evidence score between each pair of initial reviewers was 0.66 (95% CI 0.57,0.75), with 95% of coding being either the same grade (64/131 studies) or within one grade (63/131 studies). The four remaining studies, where the difference in the scores was 2 or more, were coded by a third reviewer and the final grade decided by discussion to reach consensus. There were 18 higher-scoring studies where both reviewers independently judged the study as score 4 (i.e. the results are clear and very likely to be true) or score 5 (i.e. the results are unequivocal). The 18 higher-scoring studies provided the best evidence for the review whilst the remaining lower-scoring studies (n=113) were judged to provide supporting evidence.

Methodological characteristics of the higher-scoring studies - The criteria that typified the 18 higher-scoring studies were:

- Larger sample sizes (14, 35-38).
- A clear study design (14, 35-51).
- An experimental (39-42) or pretest-posttest design (14, 38, 43-46).
- Measurements at multiple time-points during training (47-49).

- Measure of performance compared with measures of self-reflection. Such performance measures included communication and professionalism (50), pastoral skills (14), writing and story-telling skills (40), clinical judgement and diagnosis (49), scores for case-solving and other skills/knowledge based tests (51), grades for written examinations and objective structured clinical examinations (45), adherence to clinical guidelines (38), self-perceived stress, coping behaviour and self-reported nursing competence (37).
- The use of previously-validated measures (35, 39, 42, 43, 46) along with recorded inter-rater (39, 41, 42, 44) or both inter-rater and intra-rater reliability (36).

The characteristics of the higher-scoring studies are displayed in Table 2. The participants within these studies were all undertaking some form of training within a healthcare programme including dentistry (49), medicine (35,36,39-45, 51), nursing (37), physical therapy (38) and chaplaincy in healthcare (14).

Assessment instruments used in the higher-scoring studies - This review identified three broad types of instruments within the higher-scoring studies that have been used to measure the ability to self-reflect (Figure 2), namely:

- Rubric (or scoring guide) (n=9)

  This is an instrument used by another person (a rater or assessor) to evaluate a participant's (e.g. student, trainee clinician) response(s) to a real or simulated situation and against a set of pre-agreed criteria. The response is in the form of reflective writing and the situation might be a clinical case, written vignette or video.

- Self-reported scale (n=7)

  This instrument comprises a questionnaire for the participant to complete before and/or after a period of study, with responses recorded using a Likert scale.

- Observed behaviour (n=2)

  This is a measurement by an observer (rater or assessor) of a clinical performance, using a scale, following self-reflection by the participant and a one-to-one discussion about the

clinical performance. Alternatively, both the observer(s) and the participant can score the participant's clinical performance. The level of agreement between the self-score and the observer's score is then used to determine the level of insight.

Synthesis of evidence for each type of assessment instrument - None of the higher-scoring papers reported on all five utility criteria (reliability, validity, educational impact, acceptability and cost) required for a full utility analysis of the three instruments. A synthesis of reported evidence for each of the three instruments from the 18 higher-scoring studies is shown in Table 3. The summary below describes the numbers of higher-scoring studies using each instrument, who assessed or rated the level of self-reflection, and whether studies used the instrument in a pretest-posttest study design. The educational impact of the instruments was not reported within any of the higher-scoring studies. There was also limited information regarding their acceptability to the raters (39, 49) or other stakeholders e.g. students or patients. None of the higher-scoring studies reported the cost of instruments. However, the review team looked for details of the time required to prepare material for each instrument, to train raters and to conduct the assessment. These details informed an estimate of the costs for each instrument.

1. Rubrics (or scoring guides)

Six rubrics (referred to in the remainder of the text as Rubrics A to F) were used within nine higher-scoring studies to measure the self-reflective writing ability of medical students or doctors undertaking postgraduate training. In each study the assessors were clinical educators, although one study also engaged and trained two fourth-year medical students as raters (42). Two of the six rubrics, A and C, were used within a pretest-posttest study design. Rubric A was used to score reflective writing prior to a 9-hour medical ethics course, with repeated measurement four weeks later (44), and Rubric C was used to score reflective writing before and after ten months of undergraduate medical training (41). No significant change in reflection scores was reported over time in either study (41, 44). One study (42) compared two types of rubrics (Rubrics C and F).

Only two of the higher-scoring studies included information about the resources required. One study reported the time taken to score reflective writing using Rubric E as an average of four minutes per case (36), plus 30 minutes for rater training and five hours to train simulated patients prior to filming the cases. The length of time taken for script-writing or filming with the same rubric was not recorded (36). The time taken to train raters to score with Rubric C was reported to be 2 hours, whilst initial training with Rubric F took 4 hours (42). Further rater training was required for Rubric F due to drift in scoring between raters, resulting in a total training time of 6 hours (42). The review team estimated that rubrics require considerable time e.g. for rater training and for raters to read and mark assignments, and were therefore medium cost. In addition to the six rubrics (Rubrics A to F) described in the higher-scoring studies, 28 different rubrics were described in 32 lower-scoring studies (17, 23, 52-81). Of these only three employed a pretest-posttest design with a rubric to evaluate reflective skills (53, 62, 80).

2. Self-reported scales

Seven of the higher-scoring studies used self-reported scales to measure self-reflective ability, with no additional human judgement required. Although the format used to administer the scale was not always reported (43, 45), some were web-based (35, 38). Each

higher-scoring study used one of three self-reported scales namely: Reflection-in-Learning Scale (RiLS) first described by Sobral (43), Self-Reflection and Insight Scale (SRIS) first reported by Grant et al. (82), or Reflective Thinking Questionnaire (RTQ) described by Kember et al. (83). Within the higher-scoring studies all three scales have been reported as being used for pretest-posttest study designs. RiLS was used at the start and end of one 15-week term of an undergraduate medical programme to measure the impact of a voluntary course to teach learning skills (43). SRIS was used to score self-reflection and insight before and after one year of an undergraduate medical programme (45), a course of continued professional education (14) and a six-month programme of case discussion with or without peer assessment (38). RTQ was completed before and after a randomised controlled trial of a smartphone app to document "learning moments" (46). RiLS, SRIS and RTQ comprised 11, 20 and 16 items respectively. The review team estimated that the use of these existing self-reported scales is feasible and relatively cheap to undertake in terms of overall resources and are therefore of low cost.

Ten other self-reported scales, in addition to RiLS, SRIS and RQT, were used in 43 studies to assess self-reflection (43, 82-123). SRIS was the most frequently used self-reported scale. Three other self-reported scales, not including those in the higher-scoring studies, were used as pretest-posttest measures. These were the Groningen Reflective Ability Scale developed by Aukes et al. (89) and used in studies by Aukes et al. (90), Nakamura et al.(109), Duke et al. (112) and van Vliet et al. (123); the Rumination-Reflection Questionnaire developed by Trapnell and Campbell (124) and used by Sutton et al. (115) and the Teaching Reflection Scale developed by Kayapinar and Erkus (125) and used by Armutcu and Yaman (93).

3. Observed behaviour

Two higher-scoring studies described the observation of clinicians-in-training during specific patient encounters, and by an experienced clinician. The assessment instruments were introduced to the users during a group meeting with senior clinicians, described as two hours

of departmental time by Roach et al. (47), and as a short briefing by Prescott-Clements et al. (49). The time taken to observe the procedure was not recorded in either case, although the maximum time to complete the online assessment instrument was reported as 3 minutes (47). It was difficult to estimate the cost for observation, as it will depend on the time taken and also the nature of the observation. The review team estimated observation assessment instruments to be high cost compared to other instruments.

A further 16 lower-scoring studies described the assessment of self-reflection by observation of the participant (126-141). One study used observation of behaviour in a pretest-posttest design to measure the impact of 11 months of surgical residency training on levels of self-awareness (129). The raters in this study were standardised patients and scores were compared with surgical residents' self-scoring.

Discussion

The ability to self-reflect is considered a core skill for all healthcare professionals (1-5), and in an era of outcomes-based healthcare education, where this ability needs to be assessed (6,7), it presents a challenge. The principal aim of this systematic review was to identify the most effective assessment instrument that has been used to measure the ability to self-reflect in adults, in any workplace or educational setting, including healthcare. The broad search strategy, to include any profession that values and assesses self-reflection, was intentional following a scoping search of the literature.

Of the 18 higher-scoring studies identified in the review, three types of assessment instrument were subsequently identified, namely: rubrics (or scoring guides), self-reported scales and observed behaviour. These three types of instrument were used within different healthcare training programmes, including dentistry, medicine, nursing, physical therapy and chaplaincy (in healthcare).

As part of the review, each instrument was assessed for utility based on the model by van der Vleuten (30), which comprises five criteria: reliability, validity, educational impact,

acceptability and cost. This utility model has also been used in other reviews in medical education (142, 143). Two of these criteria, namely reliability and validity, are comprised of more than one component, although they may not be relevant to every type of instrument. Reliability comprises internal consistency ("whether items within a test that are intended to measure the same construct produce consistent scores" (144)), inter-rater reliability and intra-rater reliability. Internal consistency is an important component that should be possessed by all three types of instrument, but was only demonstrated by the three self-reported scales and rubrics A (44) and E (36). Inter-rater and intra-rater reliability are important in the case of rubrics and observed behaviour instruments, but they are not relevant to self-reported scales. Validity comprises content validity (the degree to which elements of an assessment are relevant to and representative of the targeted construct (145)) and construct validity (whether a scale adequately measures the reported construct (146)). Both of these should be possessed by all three types of instrument, but were only reported or implied for each of the self-reported scales and rubrics A, C and E.

The higher-scoring studies did not report the educational impact of the instruments and acceptability to the raters was only reported in two studies (39, 49) without reference to other stakeholders such as students or patients.

Although costs were not fully described, the review team were able to make an estimate for each type of assessment instrument. This was done by considering the time required by the student to undertake the self-reflection, the assessor to score or rate the student, the requirement for training prior to the use of the instrument, and the setting e.g. clinical or simulated clinical scenario (e.g. video or vignette). Using these criteria, the review team considered the observed behaviour instruments to be the most costly, due to the 1:1 nature of the assessment and the fact it was used in a clinical setting. In this case the assessor either participated in the clinical procedure, as occurred with the surgical trainees in theatre (47), or observed the trainee directly without participating in the clinical procedure in the case of recent dental graduates (49). In both instances, the assessor would be required to use some

of their own clinical time in observing the student's performance within the clinical setting, adding considerably to costs. With rubrics the greatest costs centred on rater training, the preparation of the materials on which the student had to reflect e.g. clinical cases or vignettes, and scoring the student's assignments, e.g. responses to a clinical simulation, a written assignment or e-portfolio. The review team considered the cost of rubrics to be less than that of the observed behaviour instruments, but greater than that of self-reported scales. Self-reported scales were considered to have the lowest costs because they require the least amount of assessor time during the assessment process, although time would still be required to analyse the results.

The most frequently cited instrument within the higher-scoring studies was the self-reported scale, Self-Reflection and Insight Scale (SRIS) (82). It was also one of only five instruments amongst the higher-scoring studies used for pretest-posttest (to assess student performance before and after an intervention or period of study), the others being rubrics A and C, and the other self-reported scales RiLS and RTQ. It is unclear why SRIS was the most frequently cited instrument, but as with the other self-reported scales it may have been due to relatively low cost involved and therefore its suitability for use with large cohorts.

Effective self-reflection comprises three elements: an awareness of the need to self-reflect, a willingness to engage with the process (24, 82) and an ability to self-reflect. Within the review it was difficult to determine whether or not each instrument measured all three elements. The first two, an awareness and a willingness to engage, were reportedly measured using the three self-reported scales RiLS (43), SRIS (82), and RTQ (83). Even though a willingness to engage with self-reflection was inferred in all of the higher-scoring studies, it was not always clear whether student participation was voluntary or compulsory. The third element of the process, the ability to self-reflect, was measured in the case of two of the self-reported scales RiLS (43) and RTQ (83), and four of the six rubrics, namely rubrics C (42, 50), D (48), E (36) and F (42). It is acknowledged that healthcare professionals find accurate self-assessment challenging (147), and being able to reflect with insight is an important

component of reflection. Five of the higher scoring studies used the SRIS scale (14, 35, 37, 38, 45), which specifically measures insight. The study by Prescott-Clements et al. (49) included a rating by trainers of the level of insight shown by dental graduates during a feedback session about observations of clinical cases. The measurement of insight was considered a valuable part of the overall process, and so consideration should perhaps be given to its inclusion in instruments designed to assess self-reflection.

Triangulation of information from multiple sources with multiple methods has previously been recommended within medical education for the assessment of professional competence (148,149). For the assessment of a complex skill such as self-reflection, it would therefore seem reasonable to apply more than one instrument, or type of instrument to the task. Within the higher-scoring studies only two self-reported scales, RiLS (43) and RTQ (83), purported to measure all three elements of effective self-reflection (awareness, willingness and ability). Although even here it was assumed that participation was voluntary and therefore there was a willingness on the part of the students to participate in self-reflection. This need for triangulation has also been highlighted by Miller-Kulhmann et al. (42) comparing the two rubrics Reflection on Action (Rubric C) and REFLECT (Rubric F). The authors described how the moderate correlation in the scores obtained using the two instruments was better than they had expected, given the fact the two instruments are not only different in form and intention, but have different origins, with Reflection on Action being derived from education and REFLECT derived largely from medicine. However, given that the two instruments were measuring the same construct, they also commented the correlations should have been even greater. Although they suggested the two rubrics were perhaps measuring related and overlapping variations of self-reflective ability, rather than measuring exactly the same ability. A more comprehensive assessment of self-reflection might therefore be gained through triangulation. However, the use of two or more instruments at the same time is not without its problems. The additional costs to the faculty, assessor and student, in terms of both time and money, may not justify the potential benefit if any. Particularly if these

are to be used at frequent intervals during an educational programme. An alternative might be to consider using different instruments at different times during a course of study. For example, a self-reported scale might be used during a period of self-study for a large cohort of students, whereas an observed behaviour instrument might be used at the same time as the 1:1 assessment of another clinical competence e.g. a Directly Observed Practical Skill (DOPS), mini-CEX or Case-based Discussion (CBD).

Another consideration, related to both triangulation and utility, is whether the assessment of self-reflection using any of the identified instruments should be formative or summative. None of the higher-scoring studies described a pass mark or score for any of the instruments above which a participant would be deemed to be sufficiently self-reflective, or below which they might benefit from support, remediation or further reflection. Even following 1:1 feedback, as was the case with the observed behaviour instrument and rubric C (41, 47, 49), it was still not clear if a low score would have educational consequences, such as remediation or re-assessment. The description of an expected passing standard, the route by which the standard is determined, and the consequences for the faculty, assessor and student if the standard is not met, would be useful additions to any future study reporting the use of an instrument to assess the ability to self-reflect. Alternatively, the assessment of reflective skills perhaps requires a more qualitative approach rather than the numeric, quantitative values that are often used for assessment of other competences.

The strengths of the review include the representation within the team of three healthcare professions (dentistry, medical education and veterinary medicine) and the expertise of several members in writing systematic reviews (34, 150). In addition, the utility model and subsequent analysis was an important and valuable framework for this review. However, the lack of evidence across all five utility criteria, in particular educational impact, acceptability and cost, is a limitation for educators wishing to make decisions about which instrument would be most appropriate to adopt in their setting. A recommendation as to the most effective instrument is therefore not possible, as there was inadequate information to

perform a full utility analysis. The decision to review studies from any profession (not just healthcare) was also a strength. However, the higher-scoring studies were all from healthcare-related professions, which was surprising. It might be argued that selection bias was present as the review team comprised healthcare in education professionals, but we felt this was minimised by the use of pre-determined quality indicators in scoring the studies

This study confirmed that self-reflective skills are valued and assessed within a wide variety of professions including healthcare disciplines. A standardised approach to reporting studies using the identified assessment instruments or newly-devised tools, to include the reliability, validity, educational impact, acceptability and cost would facilitate a more comprehensive utility analysis of their effectiveness. Further research is required to identify which instruments are acceptable to all stakeholders in the educational process, for example patients, trainees, clinicians, mentors and tutors. Pretest and posttest study designs, repeated sampling over time and triangulation of scores from different types of assessment instruments would also be helpful in monitoring changes in self-reflective ability in individual healthcare practitioners during training. It might then be possible to provide a threshold score, below which an individual could be offered support to become sufficiently self-reflective.

Conclusion

This review identified 3 types of instrument that can be used to assess ability to self-reflect, namely rubrics (scoring guides), self-reported scales and observation of behaviour. Under reporting of the criteria necessary for a full utility analysis meant it was not possible to make recommendations as to the most appropriate instrument(s) to be used to assess this ability. As a result, the use of more than one instrument might be appropriate. Authors of future work should be encouraged to report on the five criteria necessary for comprehensive analysis of utility. It would also be of value to include guidance as to what would constitute a good outcome in the assessment of the ability to self-reflect, and perhaps more importantly

what would constitute a poor outcome and the impact this might have on the student and their subsequent training.

References

(* = Higher-scoring papers)

1.Sandars J. The use of reflection in medical education: AMEE Guide No. 44. *Med Teach* 2009: 31(8): 685-95.

2.R.C.V.S. Day One Competences, Setting Veterinary Standards 2014 [Online]. Available from: https://www.rcvs.org.uk/document-library/day-one-competences/. [Accessed 13 December 2018]

3.N.M.C. NMC Standards for Competence of Registered nurses. 2014. [Online].Available from: https://www.nmc.org.uk/standards/standards-for-nurses/standards-of-proficiency-for-registered-nurses/ [Accessed 13 December 2018]

4.G.D.C. Preparing for practice. Dental team learning outcomes for registration. 2015. [Online]. Available from: https://www.gdc-uk.org/professionals/students-and-trainees/learning-outcomes. [Accessed 13 December 2018]

5.G.M.C. Outcomes for graduates. 2015. [Online]. Available from: https://www.gmc-uk.org/-/media/documents/Outcomes_for_graduates_Jul_15_1216.pdf_61408029.pdf [Accessed 13 December 2018]

6.Norcini J, Talati J. Assessment, surgeon, and society. *Int J Surg* 2009: 7(4): 313-7.

7.Holmboe ES, Sherbino J, Englander R, Snell L, Frank JR, Collaborators I. A call to action: The controversy of and rationale for competency-based medical education. *Med Teach* 2017: 39(6): 574-81.

8.AAMC. Recommendations for Clinical Skills Curricula for Undergraduate Medical Education 2008. [Online]. Available from: https://www.aamc.org/download/130608/data/clinicalskills_oct09.qxd.pdf.pdf. [Accessed 13 December 2018]

9.G.D.C. CPD for Dental Professionals 2018. [Online]. Available from: https://www.gdc-uk.org/professionals/cpd/enhanced-cpd [Accessed 13 December 2018]

10.Ng SL, Kinsella EA, Friesen F, Hodges B. Reclaiming a theoretical orientation to reflection in medical education research: a critical narrative review. *Med Educ* 2015: 49(5): 461-75.

11.Dulewicz V, Higgs M. Can emotional intelligence be measured and developed? *LODJ* 1999: 20(5): 242-52.

12.van Seggelen-Damen I, van Dam K. Self-reflection as a mediator between self-efficacy and well-being. *J Manage Psychol* 2016: 31(1): 18-33.

13.Lucas U, Tan PL. Assessing levels of reflective thinking: the evaluation of an instrument for use within accounting and business education. 1st Pedagogic Research in Higher Education Conference; Liverpool Hope University,Liverpool, 2006.

*14.Jankowski KR, Vanderwerker LC, Murphy KM, Montonye M, Ross AM. Change in pastoral skills, emotional intelligence, self-reflection, and social desirability across a unit of CPE. *J Health Care Chaplain* 2008: 15(2): 132-48.

15.Collins JB, Pratt DD. The Teaching Perspectives Inventory at 10 Years and 100,000 Respondents: Reliability and Validity of a Teacher Self-Report Inventory. *AEQ* 2011 61(4): 358-75.

16.Clarà M. What Is Reflection? Looking for Clarity in an Ambiguous Notion. *J Teach Educ* 2014: 66(3): 261-71.
17.Bogo M, Regehr C, Katz E, Logie C, Mylopoulos M. Developing a Tool for Assessing Students' Reflections on Their Practice. *J Soc Work Educ* 2011: 30(2): 186-94.
18.Ryan M, Ryan M. Theorising a model for teaching and assessing reflective learning in higher education. *HERD* 2013: 32(2): 244-57.
19.Van Beveren L, Roets G, Buysse A, Rutten K. We all reflect, but why? A systematic review of the purposes of reflection in higher education in social and behavioral sciences. *Educ Res Rev* 2018: 24: 1-9.
20.Byrne A, Crossan M, Seijts G. The Development of Leader Character Through Crucible Moments. *J Manage Educ* 2017: 42(2): 265-93.
21.Vitello-Cicciu JM, Weatherford B, Gemme D, Glass B, Seymour-Route P. The effectiveness of a leadership development program on self-awareness in practice. *J Nurs Adm* 2014: 44(3): 170-4.
22.Osipova A, Prichard B, Boardman AG, Kiely MT, Carroll PE. Refocusing the lens: Enhancing elementary special education reading instruction through video self-reflection. *LDRP* 2011: 26(3): 158-71.
23.Field J, Vernazza C. Developing a grading matrix for reflection. *Med Educ* 2013: 47(5): 531.
24.Nguyen QD, Fernandez N, Karsenti T, Charlin B. What is reflection? A conceptual analysis of major definitions and a proposal of a five-component model. *Med Educ* 2014: 48(12): 1176-89.
25.Ixer G. The concept of reflection: is it skill based or values? *Soc Work Educ* 2016 35(7): 809-24.
26.Birden HH, Usherwood T. "They liked it if you said you cried": how medical students perceive the teaching of professionalism. *Med J Aust* 2013: 199(6): 406-9.
27.Furmedge D. Written reflection is dead in the water. BMJ Careers [Internet]. 2016. [Online]. Available from: http://careers.bmj.com/careers/advice/Written_reflection_is_dead_in_the_water - ref2. [Accessed 13 December 2018]
28.Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. *Adv Health Sci Educ Theory Pract* 2009: 14(4): 595-621.
29.Boudreau JW. Utility analysis for decisions in human resource management. CAHRS Working Paper #88-211988.
30.Van Der Vleuten C. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996: 1:41-67.
31.Norcini JJ, McKinley DW. Assessment methods in medical education. *JTTE* 2007: 23(3): 239-50.
32.Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 1977: 33(1): 159-74.
33.Harden RM, Grant J, Buckley G, Hart IR. BEME Guide No. 1: Best Evidence Medical Education. *Med Teach* 1999: 21(6): 553-62.

34 Cake MA, Bell MA, Williams JC, Brown FJL, Dozier M, Rhind SM, et al. Which professional (non-technical) competencies are most important to the success of graduate veterinarians? A Best Evidence Medical Education (BEME) systematic review: BEME Guide No. 38. *Med Teach* 2016: 1-14.
*35.Roberts C, Stark P. Readiness for self-directed change in professional behaviours: factorial validation of the Self-Reflection and Insight Scale. *Med Educ* 2008: 42(11): 1054-63.
*36.Koole S, Dornan T, Aper L, De Wever B, Scherpbier A, Valcke M, et al. Using video-cases to assess student reflection: development and validation of an instrument. *BMC Med Educ* 2012a:12: 22.
*37.Eng CJ, Pai HC. Determinants of nursing competence of nursing students in Taiwan: the role of self-reflection and insight. *Nurse Educ Today* 2015: 35(3): 450-5.
*38.Maas MJ, van der Wees PJ, Braam C, Koetsenruijter J, Heerkens YF, van der Vleuten CP, et al. An innovative peer assessment approach to enhance guideline adherence in physical therapy: single-masked, cluster-randomized controlled trial. *Phys Ther* 2015: 95(4): 600-12.
*39.Driessen EW, Muijtjens AM, van Tartwijk J, van der Vleuten CP. Web- or paper-based portfolios: is there a difference? *Med Educ* 2007: 41(11): 1067-73.
*40.Aronson L, Niehaus B, DeVries CD, Siegel JR, O'Sullivan PS. Do writing and storytelling skill influence assessment of reflective ability in medical students' written reflections? *Acad Med* 2010: 85(10 Suppl): S29-32.
*41.Aronson L, Niehaus B, Hill-Sakurai L, Lai C, O'Sullivan PS. A comparison of two methods of teaching reflective ability in Year 3 medical students. *Med Educ* 2012: 46(8): 807-14.
*42.Miller-Kuhlmann R, O'Sullivan PS, Aronson L. Essential steps in developing best practices to assess reflective skill: A comparison of two rubrics. *Med Teach* 2016: 38(1): 75-81.
*43.Sobral DT. An appraisal of medical students' reflection-in-learning. *Med Educ* 2000: 34(3): 182-7.
*44.Boenink AD, Oderwald AK, De Jonge P, Van Tilburg W, Smal JA. Assessing student reflection in medical practice. The development of an observer-rated instrument: reliability, validity and initial experiences. *Med Educ* 2004: 38(4): 368-77.
*45.Carr SE, Johnson PH. Does self reflection and insight correlate with academic performance in medical students? *BMC Med Educ* 2013: 13: 113.
*46.Könings KD, van Berlo J, Koopmans R, Hoogland H, Spanjers IA, ten Haaf JA, et al. Using a Smartphone App and Coaching Group Sessions to Promote Residents' Reflection in the Workplace. *Acad Med* 2016: 91(3): 365-70.
*47.Roach PB, Roggin KK, Selkov G, Jr., Posner MC, Silverstein JC. Continuous, data-rich appraisal of surgical trainees' operative abilities: a novel approach for measuring performance and providing feedback. *J Surg Educ* 2009: 66(5): 255-63.
*48.McNeill H, Brown JM, Shaw NJ. First year specialist trainees' engagement with reflective practice in the e-portfolio. *Adv Health Sci Educ Theory Pract* 2010: 15(4): 547-58.

*49.Prescott-Clements LE, van der Vleuten CP, Schuwirth L, Gibb E, Hurst Y, Rennie JS. Measuring the development of insight by dental health professionals in training using workplace-based assessment. *Eur J Dent Educ* 2011: 15(3): 159-64.
*50.Learman LA, Autry AM, O'Sullivan P. Reliability and validity of reflection exercises for obstetrics and gynecology residents. *Am J Obstet Gynecol* 2008: 198(4):461.e1-8: discussion .e8-10.
*51.Koole S, Dornan T, Aper L, Scherpbier A, Valcke M, Cohen-Schotanus J, et al. Does reflection have an effect upon case-solving abilities of undergraduate medical students? *BMC Med Educ* 2012b: 12: 75.
52.Hatton N, Smith D. Reflection in teacher education:towards definition and implementation. *JTTE* 1995: 11(1): 33-49.
53.Duke S, Appleton J. The use of reflection in a palliative care programme: a quantitative study of the development of reflective skills over an academic year. *J Adv Nurs* 2000: 32(6): 1557-68.
54.Williams RM, Sundelin, G., Foster-Seargeant, E., Norman, G.R. Assessing the Reliability of Grading Reflective Journal Writing. *J Phys Ther Educ* 2000: 14(2): 23-6.
55.Fakude LP, Bruce JC. Journaling: a quasi-experimental study of student nurses' reflective learning ability. *Cura* 2003: 26(2): 49-55.
56.Cise JS, Wilson CS, Thie MJ. A qualitative tool for critical thinking skill development. *Nurse Educ* 2004: 29(4): 147-51.
57.Rees CE, Sheard CE. The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience. *Med Educ* 2004: 38(2): 138-44.
58.Ward JR, McCotter SS. Reflection as a visible outcome for preservice teachers. *Teach Teach Educ* 2004: 20(3): 243-57.
59.Adams CL, Nestel D, Wolf P. Reflection: a critical proficiency essential to the effective development of a high competence in communication. *J Vet Med Educ* 2006: 33(1): 58-64.
60.Amsellem-Ouazana D, Van Pee D, Godin V. Use of portfolios as a learning and assessment tool in a surgical practical session of urology during undergraduate medical training. *Med Teach* 2006: 28(4): 356-9.
61.Driessen EW, Overeem K, van Tartwijk J, van der Vleuten CP, Muijtjens AM. Validity of portfolio assessment: which qualities determine ratings? *Med Educ* 2006: 40(9): 862-6.
62.Wallman A, Lindblad AK, Hall S, Lundmark A, Ring L. A categorization scheme for assessing pharmacy students' levels of reflection during internships. *Am J Pharm Educ* 2008: 72(1): 05.
63.Briceland LL, Hamilton RA. Electronic reflective student portfolios to demonstrate achievement of ability-based outcomes during advanced pharmacy practice experiences. *Am J Pharm Educ* 2010: 74(5): 15.
64.Haffling A-C, Beckman A, Pahlmblad A, Edgren G. Students' reflections in a portfolio pilot: Highlighting professional issues. *Med Teach* 2010: 32(12): e532-e40.
65. Hanson K, Alexander S. The influence of technology on reflective learning in dental hygiene education. *J. Dent Educ* 2010: 74: 644-53.

66.Wittich CM, Beckman TJ, Drefahl MM, Mandrekar JN, Reed DA, Krajicek BJ, et al. Validation of a method to measure resident doctors' reflections on quality improvement. *Med Educ* 2010: 44(3): 248-55.
67.Wittich CM, Reed DA, Drefahl MM, West CP, McDonald FS, Thomas KG, et al. Relationship between critical reflection and quality improvement proposal scores in resident doctors. *Med Educ* 2011a: 45(2): 149-54.
68.Hudson JN, Rienits H, Corrin L, Olmos M. An innovative OSCE clinical log station: a quantitative study of its influence on Log use by medical students. *BMC Med Educ* 2012: 12: 111-9.
69.Wald HS, Borkan JM, Taylor JS, Anthony D, Reis SP. Fostering and evaluating reflective capacity in medical education: developing the REFLECT rubric for assessing reflective writing. *Acad Med* 2012: 87(1): 41-50.
70.Goodyear HM, Bindal T, Wall D. How useful are structured electronic portfolio templates to encourage reflective practice? *Med Teach* 2013: 35(1): 71-3.
71.Koole S, Vanobbergen J, De Visschere L, Aper L, Dornan T, Derese A. The influence of reflection on portfolio learning in undergraduate dental education. *Eur J Dent Educ* 2013: 17(1): e93-9.
72.Canniford LJ, Fox-Young S. Learning and assessing competence in reflective practice: Student evaluation of the relative value of aspects of an integrated, interactive reflective practice syllabus. *Collegian* 2014: 22(3): 291-7.
73.McKay FH, Dunn M. Student reflections in a first year public health and health promotion unit. *Reflective Practice* 2015: 16(2): 242-53.
74.Yusuff KB. Does self-reflection and peer-assessment improve Saudi pharmacy students' academic performance and metacognitive skills? *Saudi Pharm J* 2015: 23(3): 266-75.
75.Hoffman LA, Shew RL, Vu TR, Brokaw JJ, Frankel RM. Is Reflective Ability Associated With Professionalism Lapses During Medical School? *Acad Med* 2016: 91(6): 853-7.
76.McEvoy M, Pollack S, Dyche L, Burton W. Near-peer role modeling: Can fourth-year medical students, recognized for their humanism, enhance reflection among second-year students in a physical diagnosis course? *Med Educ Online* 2016: 21(1): 31940.
77.Teply R, Spangler M, Klug L, Tilleman J, Coover K. Impact of Instruction and Feedback on Reflective Responses during an Ambulatory Care Advanced Pharmacy Practice Experience. *Am J Pharm Educ* 2016: 80(5): 81.
78.Devi V, Abraham RR, Kamath U. Teaching and Assessing Reflecting Skills among Undergraduate Medical Students Experiencing Research. *J Clin Diagn Res* 2017: 11(1): JC01-JC5.
79.King AE, Joseph AS, Umland EM. Student perceptions of the impact and value of incorporation of reflective writing across a pharmacy curriculum. *Curr Pharm Teach Learn* 2017: 9(5): 770-8.
80.Nagro SA, deBettencourt LU, Rosenberg MS, Carran DT, Weiss MP. The Effects of Guided Video Analysis on Teacher Candidates' Reflective Ability and Instructional Skills. *Teacher Education and Special Education* 2017: 40(1): 7-25.

81.Tsingos-Lucas C, Bosnic-Anticevich S, Schneider CR, Smith L. Using Reflective Writing as a Predictor of Academic Success in Different Assessment Formats. *Am J Pharm Educ* 2017: 81(1): 8.
82.Grant AM, Franklin J, Langford P. The Self-Reflection and Insight Scale:A New Measure of Private Self-Consciousness. *Soc Behav Pers* 2002: 30(8): 821-36.
83.Kember D, Leung D, Jones A, Loke A, McKay J, Sinclair K, et al. Development of a Questionnaire to Measure the Level of Reflective Thinking. *Assess Eval High Educ* 2000: 25(4): 381-95.
84.Shain L, Farber BA. Female identity development and self-reflection in late adolescence. *Adolescence* 1989: 24(94): 381-92.
85.Mamede S, Schmidt HG. The structure of reflective practice in medicine. *Med Educ* 2004;38(12):1302-8.
86.Pearson M, Kayrooz C. Enabling Critical Reflection on Research Supervisory Practice *Int J Acad Dev* 2004;9(1).
87.Mamede S, Schmidt HG. Correlates of reflective practice in medicine. *Adv Health Sci Educ* 2005: 10(4): 327-37.
88.Sobral DT. Medical students' mindset for reflective learning: a revalidation study of the reflection-in-learning scale. *Adv Health Sci Educ* 2005: 10(4): 303-14.
89.Aukes LC, Geertsma J, Cohen-Schotanus J, Zwierstra RP, Slaets JP. The development of a scale to measure personal reflection in medical practice and education. *Med Teach* 2007: 29(2-3): 177-82.
90.Aukes LC, Geertsma J, Cohen-Schotanus J, Zwierstra RP, Slaets JP. The effect of enhanced experiential learning on the personal reflection of undergraduate medical students. *Med Educ Online* 2008: 13: 15.
91.Phan HP. Exploring students' reflective thinking practice, deep processing strategies, effort, and achievement goal orientations. *Educ Psychol* 2009: 29(3): 297-313.
92.Akbari R, Behzadpoor F, Dadvand B. Development of English language teaching reflection inventory. *System* 2010: 38: 211-27.
93.Armutcu N, Yaman S. ELT pre-service teachers' teacher reflection through practicum. *Procedia Soc Behav Sci* 2010: 3: 28-35.
94.Hsu LL. Metacognitive Inventory for nursing students in Taiwan: instrument development and testing. *J Adv Nurs* 2010: 66(11): 2573-81.
95.Dunn L, Musolino GM. Assessing reflective thinking and approaches to learning. *J Allied Health* 2011: 40(3): 128-36.
96.Harrington R, Loffredo DA. Insight, Rumination, and Self-Reflection as Predictors of Well-Being. *J Psychol* 2011: 145(1): 39-57.
97.Leung GSM, Lam DOB, Chow AYM, Wong, D.F.K., Chung, C.L.P., Chan, B.F.P. Cultivating reflexivity in social work students: A course-based experience. *J Practice Teach Learn* 2011: 11(1): 54-74.
98.Silvia PJ, Phillips, A.G. Evaluation self-reflection and insight as self-conscious traits. *Pers Individ Dif* 2011: 50: 234-7.

99.Xu X. Self-Reflection, Insight, and Individual Differences in Various Language Tasks. *Psychol Rec* 2011: 61(1): 41-57.
100.Wittich CM, Lopez-Jimenez F, Decker LK, Szostek JH, Mandrekar JN, Morgenthaler TI, et al. Measuring faculty reflection on adverse patient events: development and initial validation of a case-based learning system. *J Gen Intern Med* 2011b: 26(3): 293-8.
101.de Groot E, Jaarsma D, Endedijk M, Mainhard T, Lam I, Simons RJ, et al. Critically reflective work behavior of health care professionals. *J Contin Educ Health Prof* 2012: 32(1): 48-57.
102.Devi V, Mandal T, Kodidela S, Pallath V. Integrating students' reflection-in-learning and examination performance as a method for providing educational feedback. *J Postgrad Med* 2012: 58(4): 270-4.
103.Asselin ME, Fain JA. Effect of reflective practice education on self-reflection, insight, and reflective thinking among experienced nurses: a pilot study. *J Nurses Prof Dev* 2013: 29(3): 111-9.
104.Lethbridge K, Andrusyszyn M-A, Iwasiw C, Laschinger HKS, Frernando R. Assessing the psychometric properties of Kember and Leung's Reflection Questionnaire. *Assess Eval High Educ* 2013: 38(3): 303-25.
105.Mamede S, Loyens S, Ezequiel O, Tibirica S, Penaforte J, Schmidt H. Effects of reviewing routine practices on learning outcomes in continuing education. *Med Educ* 2013: 47(7): 701-10.
106.Wittich CM, Pawlina W, Drake RL, Szostek JH, Reed DA, Lachman N, et al. Validation of a method for measuring medical students' critical reflections on professionalism in gross anatomy. *Anat Sci Educ* 2013: 6(4): 232-8.
107.Ambrose LJ, Ker JS. Levels of reflective thinking and patient safety: an investigation of the mechanisms that impact on student learning in a single cohort over a 5 year curriculum. *Adv Health Sci Educ* 2014: 19(3): 297-310.
108.Kalk K, Taimalu M, Täht K. Validity and reliability of two instruments to measure reflection: a confirmatory study. *Trames* 2014: 18(2): 121-34.
109.Nakamura M, Altshuler D, Binienda J. Clinical skills development in student-run free clinic volunteers: a multi-trait, multi-measure study. *BMC Med Educ* 2014: 14: 250.
110.Stein D, Grant AM. Disentangling the Relationships Among Self-Reflection, Insight, and Subjective Well-Being: The Role of Dysfunctional Attitudes and Core Self-Evaluations. *J Psychol* 2014: 148(5): 505-22.
111.van Dulmen SA, Maas M, Staal JB, Rutten G, Kiers H, Nijhuis-van der Sanden M, et al. Effectiveness of peer assessment for implementing a Dutch physical therapy low back pain guideline: cluster randomized controlled trial. *Phys Ther* 2014: 94(10): 1396-409.
112.Duke P, Grosseman S, Novack DH, Rosenzweig S. Preserving third yeat medical students' empathy and enhancing self-reflection using small group "virtual hangout" technology. *Med Teach* 2015: 37(6): 566-71.

113.Miksza P, Tan L. Predicting collegiate wind players' practice efficiency, flow, and self-efficacy for self-regulation: An exploratory study of relationships between teachers' instruction and students' practicing. *J Res Music Educ* 2015: 63(2): 162-79.
114.Pai H-C. The Effect of a Self-Reflection and Insight Program on the Nursing Competence of Nursing Students: A Longitudinal Study. *J Prof Nurs* 2015: 31(5): 424-31.
115.Sutton A, Williams HM, Allinson CW. A longitudinal, mixed method evaluation of self-awareness training in the workplace. *EJTDS* 2015: 39(7): 610-27.
116.Tricio J, Woolford J, Escudier M. Dental students' reflective habits: is there a relation with their academic achievements? *Eur J Dent Educ* 2015: 19: 113-21.
117.Cowden RG, Meyer-Weitz A. Self-Reflection and Self-Insight Predict Resilience and Stress in Competitive Tennis. *Soc Behav Pers* 2016: 44(7): 1133-49.
118.Pai H-C. An integrated model for the effects of self-reflection and clinical experiential learning on clinical nursing performance in nursing students: A longitudinal study. *Nurse Educ Today* 2016: 45: 156-62.
119.Sutton A. Measuring the Effects of Self-Awareness: Construction of the Self-Awareness Outcomes Questionnaire. *Eur J Psych* 2016: 12(4): 645-58.
120.Tricio JA, Woolford MJ, Escudier MP. Fostering Dental Students' Academic Achievements and Reflection Skills Through Clinical Peer Assessment and Feedback. *J Dent Educ* 2016: 80(8): 914-23.
121.Cowden RG. On the mental toughness of self-aware athletes: Evidence from competitive tennis players. *S Afr J Sci* 2017: 113(1-2): 50-5.
122.Ratelle JT, Bonnes SL, Wang AT, Mahapatra S, Schleck CD, Mandrekar JN, et al. Associations between teaching effectiveness and participant self-reflection in continuing medical education. *Med Teach* 2017: 39(7): 697-703.
123.van Vliet M, Jong M, Jong MC. Long-term benefits by a mind-body medicine skills course on perceived stress and empathy among medical and nursing students. *Med Teach* 2017: 39(7): 710-9.
124.Trapnell PD, Campbell JD. Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. *J Pers Soc Psychol* 1999: 76(2): 284-304.
125.Kayapinar U, Erkus A. Measuring Teacher Reflection: Development of the Teacher Reflection Scale (TRS). *Egitim Arastirmalari-EJER* 2009: 37: 144-58.
126.Wohlers AJ, London M. Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Pers Psychol* 1989: 42(2): 235-61.
127.Berr SA, Church AH, Waclawski J. The right relationship is everything: Linking personal preferences to managerial behaviors. *Human Resource Devel Quart* 2000: 11(2): 133-57.
128.Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ* 2003; 326(7388): 546-8.

129.Hassett JM, Zinnerstrom K, Nawotniak RH, Schimpfhauser F, Dayton MT. Utilization of standardized patients to evaluate clinical and interpersonal skills of surgical residents. *Surgery* 2006: 140(4): 633-8; discussion 8-9.
130.Moorthy K, Munz Y, Adams S, Pandey V, Darzi A, Imperial College--St. Mary's Hospital Simulation G. Self-assessment of performance among surgical trainees during simulated procedures in a simulated operating theater. *Am J Surg* 2006: 192(1): 114-8.
131.Young M, Dulewicz V. Relationships between emotional and congruent self-awareness and performance in the British Royal Navy. *J Manage Psychol* 2006: 22(5): 465-79.
132.Eva KW, Regehr G. Knowing when to look it up: a new conception of self-assessment ability. *Acad Med* 2007: 82(10 Suppl): S81-4.
133.Papinczak T, Young L, Groves M, Haynes M. An analysis of peer, self, and tutor assessment in problem-based learning tutorials. *Med Teach* 2007: 29(5): e122-32.
134.Chen LP, Gregory JK, Camp CL, Juskewitch JE, Pawlina W, Lachman N. Learning to lead: self- and peer evaluation of team leaders in the human structure didactic block. *Anat Sci Educ* 2009: 2(5): 210-7.
135.Young M, Dulewicz V. A study into leadership and management competencies predicting superior performance in the British Royal Navy. *J Manage Dev* 2009: 28(9): 794-821.
136.Calhoun AW, Rider EA, Peterson E, Meyer EC. Multi-rater feedback with gap analysis: an innovative means to assess communication skill and self-insight. *Patient Educ Couns* 2010: 80(3): 321-6.
137.Gow KW. Self-evaluation: how well do surgery residents judge performance on a rotation? *Am J Surg* 2013: 205(5): 557-62; discussion 62.
138.Jepsen RM, Spanager L, Lyk-Jensen HT, Dieckmann P, Ostergaard D. Customisation of an instrument to assess anaesthesiologists' non-technical skills. *Int J Med Educ* 2015: 6: 17-25.
139.Kamangar F, Davari P, Parsi KK, Li CS, Wang Q, Mathis S, et al. 360-degree Evaluations on Physician Performance as an Effective Tool for Interprofessional Teams: A critical analysis of physician self-assessment as compared to nursing staff and patient evaluations of providers. *Dermatol Online J* 2016: 22(7): 15.
140.Lo K, Osadnik CR, Leonard M, Maloney SR. Student-clinician agreement in clinical competence as a predictor of clinical placement performance in Australian undergraduate physiotherapy students. *Physiother Theory Prac* 2016: 32(1): 63-8.
141.Lyle B, Borgert AJ, Kallies KJ, Jarman BT. Do Attending Surgeons and Residents See Eye To Eye? An Evaluation of the Accreditation Council For Graduate Medical Education Milestones in General Surgery Residency. *J Surg Educ* 2016: 73(6): e54-e8.
142.Rees EL, Hawarden AW, Dent G, Hays R, Bates J, Hassell AB. Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A BEME systematic review: BEME Guide No. 37. *Med Teach* 2016: 38(5): 443-55.

143.Mortaz Hejri S, Jalili M, Shirazi M, Masoomi R, Nedjat S, Norcini J. The utility of mini-Clinical Evaluation Exercise (mini-CEX) in undergraduate and postgraduate medical education: protocol for a systematic review. *Sys Rev* 2017: 6: 146.
144.Tang W, Cui Y, Babenko O. Do we really know what it is and how to assess it? *J Psychol Behav Sci* 2014: 2(2): 2015-220.
145.Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychol Assess* 1995: 7: 238-47.
146.Cronbach L, Meeh P. Construct validity in psychological tests. *Psychol Bull* 1955: 52: 281-302.
147.Kustritz M, Molgaard L, Rendahl A. Comparison of student self-assessment with faculty assessment of clinical competence. *J Vet Med Educ* 2011: 38: 163-70.
148.Holsgrove G, Davies H. Assessment in the foundation Programme. In: Jackson ND, Jamieson A, Khan A, editors. Assessment in medical education and training : a practical guide. Abingdon: Radcliffe; (2007) p. 44.
149.van der Vleuten CP, Schuwirth LW, Driessen EW, Dijkstra J, Tigelaar D, Baartman LK, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012: 34(3): 205-14.
150.Rhind SM, Baillie S, Brown F, Hammick M, Dozier M. Assessing competence in veterinary medical education: where's the evidence? *J Vet Med Educ* 2008: 35(3): 407-11.

List of Tables

Table 1 Inclusion and exclusion criteria

Table 2 Characteristics, authors and frequency of the 18 higher-scoring studies

Table 3 Synthesis of evidence from the higher-scoring studies for each instrument type

List of Figures

Figure 1 The search strategy applied in Ovid MEDLINE

Figure 2 PRISMA flow diagram

Appendices

Appendix 1 Definitive additional database search strategies

Appendix 2 Data extraction coding sheet

Table 1 Inclusion and exclusion criteria

	Inclusion criteria	Exclusion criteria
Setting	Educational setting Workplace setting for any vocation	
Population	Young adults, adults Undergraduate or postgraduate students	Children ≤ 16 years of age Clinical studies with patients as participants in studies of clinical conditions e.g. weight loss, management of diabetes, mental health
Intervention	Instrument for the assessment of self-reflection or self-awareness for example (but not exclusively) scale, tool, measure, questionnaire Studies with self-reported scales that were then given a score by computer or another person	Studies where students were asked directly to describe their own self-reflective ability, unless part of a validated externally scored scale was used Tests of clinical reasoning, personality, cultural awareness or emotional intelligence
Evaluation	Quantitative, qualitative and mixed-methods studies which provide primary data about both the assessment instrument and the participants who were assessed.	Descriptive studies outlining benefits of e.g. a training intervention with no description of an assessment instrument
Limits	In English Published between 1st January 1975 and 1st August 2017	Not in English