A review of instruments for assessing complex vocational competence

Jackie Greatorex, Martin Johnson and **Victoria Coleman** Research Division

Introduction

Complex competences integrate a variety of skills. For example, models of professional proficiency or intelligent practice often incorporate the ability of a person to construct a holistic view of a problem or situation (Dreyfus & Dreyfus, 1980; Eraut, 1994). There is evidence that observational methods can be used to capture the integration of skills, knowledge and attitudes that pertain to higher level work (Eraut & Steadman, 1998). The aim of this research was to explore the measurement qualities of prevalent approaches to observation (checklists and Global Rating Scales [GRSs]) in the context of assessing complex competence.

According to Lester (2000), the assessment of complex competence is possible if the performances assessed are approached holistically rather than in an instrumental or piece-by-piece fashion. At the same time this presents a challenge to assessment. Watson (1994) argues that for observation-based assessment to be reliable, fair, generally practicable and cost-effective it needs to include adequate quality control to ensure consistency across assessors, and involve sensible decisions about the range and number of observations of performance that are required to make a reliable judgement about competence.

Since the holistic assessment of complex competence is a challenge, it is useful to look more closely at cases where assessment models are used to capture complex competence and are considered to be trustworthy. Training in the medical field is a safety-critical professional context involving the assessment of important competences. Moreover, these assessment processes are highly respected since they result in the certification of practice in a very high-stakes professional domain. This article looks more closely at aspects of the assessment processes used in this context to explore how observation-based assessment is used to assess complex competences without concern that the assessments compromise validity. The discussion of assessment of complex competence is foregrounded with a review of human judgement research.

One assessment approach is the checklist approach. This approach involves the development of a checklist of features that are used as the basis for observations of performance. Checklists require raters to indicate the performance or omission of directly observable actions with a separate checklist required for each task (Ilgen, Ma, Hatala, & Cook, 2015). The items are scored for presence or absence.

A concern with this approach is that it leads to an atomistic construction of competence, which narrows the scope of initiative and field of responsibility of professional practitioners, and fails to encompass matters such as maturity, critical thinking, group work, and complex skills (Winter, 1995). In addition, since assessment can be a significant influence on learning, such an assessment approach could also lead to the construction of learning situations where the notion of simple competence dominates.

A second approach is the use of a GRS. These scales require raters to judge participants' overall performance or to provide impressions of performance on subtasks or traits (Ilgen et al., 2015; Norcini, 2005). There is no rule about how many points should be in the scale. The GRS is applied to several traits, such as physical examination and history taking, as well as in several situations such as in Accident and Emergency, and in General Practice. One distinguishing characteristic of a GRS is it is used for multiple situations and traits. It must be noted however that the term 'global rating' is not always used consistently, with greater clarification needed across research in how it is defined and distinguished from other scoring instruments (Boursicot et al., 2011).

In the absence of examples of tasks with mark schemes in the form of checklists and a GRS, we developed the following fictional assessment tasks and extracts from mark schemes for illustrative purposes only. The first task involved the candidate leading a meeting about the progress of a project. The second task involved the candidate giving a 15-minute presentation about the completed project to a group of 20 peers who have not been involved in the project. Both performances were to be observed and rated. The fictional mark schemes provided are:

- GRS for use with Task 1 and 2 (Table 1 on page 36)
- Checklist for use with Task 1 (Table 2 on page 36)
- Checklist for use with Task 2 (Table 3 on page 36).

Human judgement

There is a literature about the strengths and weaknesses of using human judgement. Here we offer a short exploration of the pros and cons of the use of GRSs and checklists to inform human judgement.

GRS and human judgement

Research has indicated that GRSs can give an accurate overview of students' abilities. For instance, the surgery skills of medical students were assessed through ratings on a GRS on ten specific traits (Pulito, Donnelly, & Plymale, 2007). Students were also assigned a grade summarising their performance which was based on the examiner's perception of the student's overall performance, considering any additional factors, and weighting their performance of the ten traits as they felt was appropriate. It was found that the rating on any of the ten specific traits was 75–80 percent accurate in predicting a student's overall grade. Thus this shows that scores on a GRS were able to accurately reflect judgements of students' overall performance. However, it also indicates that examiners tend to make single overall judgements on a student's performance rather than considering each trait separately, suggesting that using multi-item GRSs is unnecessary. That said, there was evidence of some variation between traits with the non-cognitive aspects rated higher overall compared to cognitive aspects. The limitation of this and similar studies is that there is no

Table 1: Fictional GRS for use with Assessment Tasks 1 and 2

Rate the candidate's performance as *Unacceptable, Improvement needed, Adept, Very Good* or *Outstanding* for each of the following traits:

	Level of performance				
	Unacceptable	*Improvement needed*	*Adept*	*Very Good*	*Outstanding*
	1	*2*	*3*	*4*	*5*
People management Negotiating allocation of tasks and resources to appropriate staff. Rewarding achievement, giving credit where due and challenging underperformance. Maintaining good relationships.					
Time management Ensuring activities meet deadlines and fit allocated time windows.					
Branding Communications are brand appropriate.					
Written communication Text is clear, succinct and engaging. Sentences and paragraphs are well constructed and build up to an overall conclusion. Text is augmented by varied and imaginative images, graphics and other media which reinforce the message. Images, graphics and other media are accessible and appropriately labelled.					
Expertise Using facts and credible evidence to inform analysis and evaluation which are used to draw conclusions. The content is original. No content is sexist, racist, ageist, homophobic or inflammatory in nature.					
Overall performance					

Table 2: Fictional checklist for use with Assessment Task 1

Tick items which were achieved. All items must be present to gain a pass.

Trait	*Meeting*
Time management	☐ Started and ended on time ☐ Each section started and ended on time ☐ The purpose(s) of the meeting was/were clear ☐ Appropriate timespans were given to each agenda item ☐ Meeting papers, agenda, minutes of previous meeting were received well before the meeting ☐ All agenda items were covered in the meeting ☐ Project activities/stages met deadlines ☐ Work progress was checked against milestones ☐ Necessary changes to timelines were made
Branding	☐ Organisational template was used ☐ All images/text and so on met brand guidelines ☐ Copyright permissions gained as necessary ☐ Copyright notice added as required
Written communication	☐ Tables/figures images/graphics were accessible and augmented the message ☐ The text was grammatically accurate ☐ The text was correctly spelt ☐ Paragraphs had an introduction to the topic, gave evidence about the topic and had a concluding sentence, as appropriate
People management	☐ Active listening was exercised ☐ Questions were answered ☐ Appropriate responses were given to questions and comments ☐ Credit was attributed where due ☐ All meeting attendees had the opportunity to contribute as relevant ☐ Discussion focused on the issues to hand ☐ All relevant perspectives were considered before agreeing a way forward
Expertise	☐ Expert knowledge was demonstrated ☐ Conclusions were drawn via analysis of facts or evaluation of evidence ☐ Content was devoid of sexist, racist, ageist, homophobic or inflammatory content

Table 3: Fictional checklist for use with Assessment Task 2

Tick items which were achieved. All items must be present to gain a pass.

Trait	*Presentation*
Time management	☐ Started and ended on time ☐ Each section started and ended on time ☐ The purpose(s) of the presentation was/were clear ☐ Actions and deadlines/milestones were agreed and recorded
Branding	☐ Organisational template was used ☐ All images/text and so on met brand guidelines ☐ Copyright permissions gained as necessary ☐ Copyright notice added as required ☐ Organisational authorisation gained as needed
Written communication	☐ Tables/figures/images/graphics were accessible and augmented the message ☐ The text was grammatically accurate ☐ The text was correctly spelt ☐ Paragraph had an introduction to the topic, gave evidence about the topic and had a concluding sentence, as appropriate
People management	☐ Active listening was exercised ☐ Questions were answered ☐ Appropriate responses were given to questions and comments ☐ Credit was attributed where due ☐ Questions and comments were requested ☐ Attendees were attentive
Expertise	☐ Expert knowledge was demonstrated ☐ Conclusions were drawn via analysis of facts or evaluation of evidence ☐ Content was devoid of sexist, racist, ageist, homophobic or inflammatory content

independent measure against which to compare individual trait scores and overall judgements and thereby determine which is more accurate.

Furthermore, a comparison of a single-trait GRS with a multi-trait GRS found that whilst there was significant correlation between the two, a single-trait GRS was not able to reflect the differences found between different traits, such as the finding that ratings tended to be higher on humanistic traits compared to technical ones (Domingues, Amaral, & Zeferino, 2009). Additionally, the ratings on technical traits correlated particularly well with the single-trait GRS scores. This demonstrates that certain traits may have a greater impact on single-trait GRS scores, and that single-trait GRSs are limited as they cannot reflect variation within performance on specific traits (Domingues et al., 2009). This finding may be due to the psychological phenomenon that people can be good at judging individual traits and less good at combining information into an overall judgement (Einhorn, 1972; Laming, 2004).

Overall, it appears that a GRS can be used to generate scores for specific traits which reflect judgements of the student's overall performance. Additionally, the use of a multi-item GRS enables a more in-depth understanding, although examiners do often give fairly uniform responses across these (Pulito et al., 2007).

Checklists and human judgement

Prior research shows that experts can successfully identify the characteristics in a checklist, but they are poor at combining the decisions from each point in the checklist into an overall judgement. For instance, Eining, Jones, and Loebbecke (1997) evaluated the effectiveness of cue processing aids in fraud detection. The aids were:

- A checklist
- A statistical model (using data collected by humans using a checklist)
- An expert system (using data collected by humans using a checklist)
- Unaided judgement (when auditors make an overall judgement using the evidence available).

The most superior fraud assessment was achieved by the statistical model and the expert system; here unaided judgement was inferior but better than using the checklist alone. Later, Boritz and Timoshenko (2014) reviewed related studies and argued that humans can effectively respond to each item on a checklist, but that mechanical combination of the decisions on each checklist item (statistical model/expert system) is superior to human combination of the decisions on each point on the checklist. Additionally, it is noteworthy that all forms of cue processing aids rely on high-quality checklists which contain all the key traits (Boritz & Timoshenko, 2014).

Combining different types of evidence

There is a host of research about how humans integrate evidence from several sources to make a judgement and the quality of those judgements – examples include Kahneman (2011) and Laming (2004). Here we focus on work comparing human and mechanical approaches to integrating evidence.

Highhouse and Kostek (2013) reviewed research on college admissions and employee selection in the US. Generally the studies compared predictions of college success or achievement in a job from:

- Human integration of information into an overall judgement
- Mechanical integration of evidence.

An illustrative example is that in a police assessment centre each assessor scored each candidate's performance on each exercise, and the assessors jointly provided an overall rating for each candidate (Feltham, 1988). A statistical combination of the scores on various exercises was a better predictor of success as a police officer than the consensus overall judgement. In four of the seven studies about college admissions, a mechanical combination of evidence outperformed human integration of evidence (Highhouse & Kostek, 2013). In 6 of the 13 studies about employee selection, a mechanical combination of evidence outperformed human integration of evidence, and 3 gave the reverse result. Together the research shows that mechanical combination of evidence tends to be better than human judgements which integrate a variety of evidence.

Methods for mechanically combining assessment outputs (scores, grades etc.) are many and varied. An example follows by way of illustration. Many (post)graduate degrees assess aspects of complex competence at many intervals (Janssen et al., 2016). In the case of Medicine these can involve scores, grades (and equivalent) as well as the textual comments of the assessors. One way of combining the scores, grades and text is a *Multi-Entity Bayesian network* (Janssen et al., 2016). A *Bayesian network* is a statistical model that uses Bayesian methods to estimate the parameters of the posterior distribution (probability distribution of an unknown quantity treated as a random variable conditional on the data provided). A Multi-Entity Bayesian network goes beyond Bayesian networks to form complex situation-specific Bayesian networks, and as more data is accrued in the database the network and outputs are updated. In other words, the Multi-Entity Bayesian network can account for the assessment context, which other types of Bayesian models cannot. The data fed to the model are scores, grades (equivalents) and sentiment levels derived from a sentiment analysis of assessors' textual comments (Janssen et al., 2016). The Multi-Entity Bayesian network combines the information and estimates the true present level of performance. Output from the model is posterior probability tables for multiple variables, such as level of motivation. These analytics are interpreted by experts to make decisions about degree classifications, learning needs to be addressed and so on.

Why can mechanical combination be better than human combination of evidence?

To explain why mechanical combination can outperform human integration of evidence we return to theory. Kahneman (2011) explains that there are two reasoning systems controlling human judgement. *System 1* is intuitive, unconscious, automatic and fast. System 1 thinking associates new information with established thought patterns and understandings, rather than noting the uniqueness of the current situation. For example, when a doctor encounters a case of measles and uses System 1 thinking he/she recalls cases he/she previously experienced rather than recognising the distinguishing characteristics of this case. System 1 thinking quickly amalgamates new information into a model (script/schema) based on prior experience, and potentially overlooks key new data. *System 2* thinking is deliberate, conscious, laboured and slow. System 2 thinking integrates information using a coherent judgement model, and can be used to make considered and logical decisions. System 1 thinking often obstructs System 2 thinking, which may influence the quality of human judgement. Both systems must be useful otherwise they would have disappeared through evolutionary processes.

Arguably, human judgement involves simplifying heuristics (Gilovich, Griffin, & Kahnemann, 2002). Generally these heuristics are helpful and provide accurate judgements, however, they can lead to unintentional biases (Gilovich & Griffin, 2002). For example, Tversky and Kahneman (1982) found that sometimes people appraise the likelihood of an event by the ease with which incidences can be recalled. This mental short cut is known as the *availability heuristic*. Often the availability heuristic is successful because recurring events are brought to mind more effortlessly than infrequent events. But the availability heuristic can result in biased judgements, for example, biases due to the retrievability of instances. One group might be judged larger than another, even though the two groups are of equal size. The bias occurs because the group of familiar instances is more easily brought to mind and therefore seems larger.

This theory applies to the situation of an assessor combining performance evidence to give an overall performance rating. The assessor's experience might be that students who are good at physical examinations are sound doctors. That is, the assessor has a script that students who are reasonable at physical examinations are able doctors. Therefore, when the assessor is integrating evidence from physical examinations, professionalism and so on, they give greatest weight to the students' performance on the physical examination. In other words, they used System 1 thinking. If the assessor's experience is correct, then the System 1 thinking was successful. If however, the assessor's schema were factually incorrect, then the System 1 judgement is biased. Some mark schemes might circumvent such biases by requiring assessors to judge each trait separately and then judgements are mechanically combined to give an overall score.

In the following sections we consider how these issues extend into assessing complex competence by focusing on a widely used GRS (*mini-CEX*) and an area where checklists are popular (essential skills).

The mini-CEX: A Global Rating Scale

The *Clinical Evaluation Exercise* (*CEX*) was designed as a practical assessment of trainee doctors' clinical skills (Norcini, Blank, Duffy, & Fortna, 2003). The CEX involved trainees carrying out a two- hour full history and physical examination of an inpatient, being observed and assessed on their clinical skills by a supervising clinician using a GRS. Whilst the CEX enabled the assessment of a trainee's clinical skills with a real patient, it had limited generalizability beyond this specific context, only involved a single assessor, and was not representative of normal doctor-patient interactions (Norcini, 2005).

The mini-CEX is a modification of the original CEX that was developed by the American Board of Internal Medicine and has since been used in a variety of countries including the UK. It is a GRS which assesses the clinical skills of trainee doctors across a number of settings and scenarios (Norcini, 2005). It involves a higher trained physician assessing the trainees' performance of clinical skills on a routinely conducted clinical task. Trainees are assessed on seven domains: history taking; physical exam; professionalism; clinical judgement; communication skills; organisation/efficiency; and overall clinical care (Norcini et al., 2003). This is done on a rating scale, which ranges from six to nine points with the bottom of the scale representing unsatisfactory/below expectations and the top of the scale superior/above expectations. Assessors are often required to complete this form online and to provide feedback to trainees immediately afterwards, noting particular strengths or weaknesses (Norcini, 2005). The mini-CEX lasts approximately 15–20 minutes and is carried out during normal clinical activities. Six are carried out in each of the first and second year of the UK foundation programme for trainee doctors. They are organised by the trainee doctors themselves, are spaced out throughout the year, and conducted by a variety of different assessors in different scenarios and settings (Norcini, 2005). It is not necessary for all domains to be assessed on each mini-CEX if they are not relevant to particular scenarios.

When all six mini-CEX are completed the data is collated and returned to the trainee. It was designed as a formative learning and development tool, enabling trainees to reflect on their strengths and weaknesses, rather than having a summative function of measuring proficiency levels, and was not intended as a tool to compare trainees (Norcini, 2005; Weston & Smith, 2014; Yates, 2013). That said, it is frequently used in a summative manner (Hawkins, Margolis, Durning, & Norcini, 2010).

Research evidence about the mini-CEX

The mini-CEX shows good feasibility and can form part of normal clinical practice (Pelgrim, 2010; Yates, 2013). There is evidence that assessors view the mini-CEX favourably (Norcini et al., 2003). Whilst one literature review suggested there was evidence of learner engagement in the mini-CEX (Yates, 2013), others found that trainees did not consider it a useful part of their training which may relate to a lack of understanding of its formative purpose (Weston & Smith, 2014).

Evidence that individuals' mini-CEX rating scores in all domains appear to increase over time is supportive of its construct validity (Hawkins et al., 2010; Pelgrim, 2010). However, factors beyond clinical competence may also influence mini-CEX ratings. Assessors make social judgements when assessing trainees and differences in these judgements impact rating scores (Gingerich, van der Vleuten, Eva, & Regehr, 2014). Differences have been found in the ratings given by assessors who were residents (doctors holding certain degrees who are not yet fully licensed) compared to those who were faculty members (Al Ansari, Ali, & Donnon, 2013). Generalizability of the mini-CEX results may be limited by the influence of examiner factors on reliability, with examiner factors accounting for 23–40% of variance compared to trainee ability which accounts for 4–17% of variance (Yates, 2013).

There is also evidence that, when assessing clinical competence in a real-life setting such as this or in more complex situations, assessors may give overinflated rating scores thus limiting validity (Hawkins et al., 2010; Norcini et al., 2003). Evidence of criterion validity has been inferred by comparing the mini-CEX with other assessment of clinical skills, such as oral and written exams or performance evaluations (Al Ansari et al., 2013; Hawkins et al., 2010; Pelgrim, 2010).

Finally, research suggested that mini-CEX scores from ten encounters produces good reliability (Norcini et al., 2003). Inter-rater reliability of the mini-CEX is influenced by the number of points on a scale, with greater inter-rater reliability on nine-point scales compared to five (Yates, 2013). There is good internal consistency between the ratings given to the different domains of the mini-CEX, with a Cronbach's alpha of 0.79 (Weston & Smith, 2014).

It is noteworthy that in the 2012 update to the National Health Service (NHS) Foundation Programme curriculum for trainee doctors, the UK mini-CEX was updated to remove the tick boxes. Therefore the mini-CEX has moved away from using a GRS and instead focuses on

written feedback of strengths and weaknesses and the development of action plans (Weston & Smith, 2014). This was done in order to return the focus on the use of the mini-CEX as a formative tool. That said, the mini-CEX in its GRS form remains in use elsewhere.

Essential Skills Clusters (ESCs): A checklist approach

The standards for pre-registration nurses and midwives are set out by the Nursing & Midwifery Council (NMC) (2010). The standards incorporate, amongst other things, a set of mandatory *Essential Skills Clusters* (ESCs) which, according to Borneuf and Haigh (2010), developed out of concerns about skill deficits in earlier proficiency requirements. The standards state that the ESCs are to be used as guidance and should be incorporated into all pre-registration nursing and midwifery programmes, although the nature of programme incorporation is left to local determination (NMC, 2010).

ESCs encompass a broad set of interconnecting skills, knowledge and attitudes that are used to observe and assess trainee nurses and midwives. The ESCs comprise five skills clusters: care, compassion and communication; organisational aspects of care; infection prevention and control; nutrition and fluid management; and medicines management. Within these clusters there is a mixture of soft skills and knowledge content. For example, there are soft skills requirements to evidence that trainees "Form appropriate and constructive professional relationships with families and other carers" and "Manage and diffuse challenging situations effectively" (NMC, 2010, p.105 – *Care, compassion and communication ESC*). In other clusters there are requirements to demonstrate content knowledge such as "Recognises potential signs of infection and reports to relevant senior member of staff" (NMC, 2010, p.124 – *Infection prevention and control ESC*) and "Takes and records accurate measurements of weight, height, length, body mass index and other appropriate measures of nutritional status" (NMC, 2010, p.130 – *Nutrition and fluid management ESC*).

A variety of methods are used to assess ESCs, and these are characterised by a number of common elements. ESC assessment arrangements include:

- Mentor observation, with this usually organised around three meeting points (pre-, during-, and post-practice). This arrangement ensures that the assessment process performs both formative and summative functions
- The assessment materials articulate the criteria that are the basis for assessment
- Self-assessment is a key element of the assessment process. The assessment materials include space where the trainee is expected to record reflections on their practice and learning, and is in keeping with the tradition that reflection on practice has an important role in professional development, for example, Schön (1983)
- The assessment materials have an accountability function:
    - They are a record of attendance. This is because there are requirements that trainees complete a number of hours of practice that are attested to by the mentor.
    - They are a record of competence that is signed off by the mentor. The form of competence reporting for sign-off differs. Some ask the mentor to make a pass/fail judgement of competence, others ask for a judgement of whether competent performance has been achieved in context(s), or ask for a judgement on the level of competence in terms of the trainee's participation involvement and the degrees of assistance required.

Comparing checklists and Global Rating Scales using systematic reviews

In this section we discuss accrued evidence about the advantages and disadvantages of checklists and GRSs. Results from many studies can be statistically combined in a systematic review, when studies meet particular quality criteria. Therefore, systematic reviews are useful for drawing evidence-based conclusions.

There are three systematic reviews which are key to our research topic. Ilgen et al. (2015) aimed to compare the reliability and validity of checklists and GRSs, as well as the correlation between scores from the two different scales. Their work was undertaken in the context of simulation-based assessment in health professionals' education. Their final analysis included 45 studies. McKinley et al. (2008) aimed to quantify the extent to which existing checklists allow for assessing both the humanistic and technical competencies needed in procedural competencies in the context of clinical procedures (tasks directly related to the care of a single patient, excluding physical examination). Their final analysis covered 75 studies. Finally, Ahmed, Miskovic, Darzi, Athanasiou, and Hanna (2011) aimed to identify assessment instruments and evaluate their validity and reliability in the context of direct observation of procedural/technical skills assessment in Medicine, (e.g., surgical skills). Such assessments may be work-based or simulations. Their final analysis included 106 studies.

The outcomes of an individual systematic review may not be generalizable to the assessment of complex competence across all professions. Furthermore, the outcome of the systematic reviews cannot be quantitatively combined or compared. Unfortunately, Ahmed et al. (2011) found that they could not statistically amalgamate results from different studies due to the diverse study designs. However, together, Ilgen et al. (2015) and McKinley et al. (2008) provide a solid evidence base from which to draw key comparisons between checklists and GRSs.

Reliability

Inter-rater reliability was substantial for both checklists and GRSs (Ilgen et al., 2015). Inter-item reliability was substantial for GRSs and lower for checklists. Interstation reliability was good for GRSs and suboptimal for checklists. Broadly speaking, the literature points towards GRSs achieving slightly better reliability than checklists.

Validity

Validation and development can be intense for task-specific checklists as each requires validation (Ilgen et al., 2015). In contrast, a GRS can be validated using evidence from many tasks yielding robust validity evidence, which can be less intense (Ilgen et al., 2015).

Ilgen et al. (2015) found that there was no difference between the content validity of checklists and GRSs. To evaluate content validity, researchers referred to previous instruments and expert consensus. On the other hand, McKinley et al. (2008) reviewed 88 checklists and

found that the inclusion of key competencies varied. The proportion of checklists including each competency was as follows:

Preparation: 74%,

Infection control: 32%,

Communication and working with the patient: 36%,

Teamworking: 15%,

Safety: 51%,

Procedural competence: 97%,

Post-procedural care: 27%.

Therefore, McKinley et al. (2008) argued that a GRS with a descriptor for each of these themes would have greater content validity than many checklists. Together, this information is a reminder that the quality of individual assessment instruments varies with several factors, including style of assessment (checklist or GRS).

Regarding criterion validity, Ilgen et al. (2015) found that the criterion validity was equivalent for checklists and GRSs in 11 studies, and higher for GRSs in a further 6 studies. Furthermore, Ilgen et al. (2015) reported there was a correlation of 0.76 between checklist and GRS measures, denoting that they measured somewhat similar traits. On balance, checklists and GRSs may measure similar traits, but GRSs generally have higher criterion validity.

The outcomes of rater training were under reported (Ilgen et al., 2015). To be specific, one study about checklists and two about GRSs reported rater training outcomes. This resonates with the point made earlier that there is little research about rater training for the mini-CEX. Therefore, rater training is likely to be an area requiring further research.

The scope of systematic reviews

There are several factors which were not included in either systematic review. These included cognitive validity (whether the raters or test takers used the intended cognitive activities). An example of a single study that addresses cognitive validity is McIlroy, Hodges, McNaughton, and Regehr (2002). They found that students adapt their behaviours according to their perceptions and expectations of the measurement tool being used to assess them. A total of 57 medical students assigned to 2 groups were primed to expect that they were being assessed on a 10-station Objective Structured Clinical Examination (OSCE) with either a GRS or checklist measure. McIlroy et al. (2002) found a significant interaction between the type of OSCE measure and the measure students expected to be used. Those in the group anticipating a checklist attained higher checklist scores but lower GRS scores than those in the group anticipating a GRS assessment, although the effect size was small. They also found higher interstation reliability coefficients for the GRS ratings than for the checklist scores across all students, thus suggesting that overall GRS ratings show higher reliability regardless of students' perceptions. The difference in interstation reliability between the two measures was greater for the students expecting to be assessed using a GRS, which also showed lower interstation reliability for both measures. The researchers speculated that when students expect to be assessed using GRSs, their performance became more heterogeneous across stations. This may be because the students are less able to rely on a 'script' and so their performance varies according to their content-specific expertise on each station, thus decreasing reliability. This study shows that the remit of the systematic reviews is somewhat limited. Nonetheless, the systematic reviews provide rich and solid evidence regarding many validity and reliability issues. Together the systematic reviews reveal that GRSs tend to achieve greater validity than checklists.

Conclusions

The aim of our research was to explore the measurement qualities of checklists and GRSs in the context of assessing complex competence. Firstly, we reviewed the literature about the affordances of human judgement and mechanical combination of human judgements. Secondly, we considered examples of assessment instruments (checklists and GRSs) used to assess complex competence in highly regarded professions. These examples served to contextualise and illuminate assessment issues. Finally, we compiled research evidence from the outcomes of systematic reviews which compared advantages and disadvantages of checklists and GRSs. Our research has caveats: for example, focusing on healthcare may restrict the generalizability of the findings. However, merging the research on human judgement, mini-CEX, essential skills and systematic reviews provides a nuanced and firm evidence base for drawing key conclusions.

Reliability

The weight of evidence signifies that GRSs generally achieve better reliability than checklists. Furthermore, human judgement research tends to confirm that accuracy is enhanced by humans judging individual traits and those judgements being mechanically combined to gain an overall assessment. Technology in this area is ever advancing, including deriving sentiment levels from assessors' comments and combining them with other quantitative data to report assessment outcomes (Janssen et al., 2016). Hence, we recommend that human judgements focus on judging individual traits and that these judgements are combined by computer, when practicable.

Validity

Together the systematic reviews suggest that GRSs tend to achieve greater validity than checklists. However, validity is a multifaceted concept and the picture is nuanced. There is no difference between the content validity of checklists and GRSs; however the content validity of individual instruments varies. Whilst checklists and GRSs can measure similar traits, the criterion validity of GRSs is generally slightly higher. In summary, it is recommended that GRSs are considered preferable to checklists, although a high-quality checklist is better than a poor-quality GRS.

Social bias

Concerns about assessor bias are a common feature to both checklist and GRS approaches. For example, studies of mini-CEX show that rating scores appear to be influenced by the nature of the assessor, such as whether they were a resident or a faculty member (Al Ansari et al., 2013). Social biases can also extend to the contextual features that surround an assessment. In the case of the mini-CEX, evidence

suggests that assessing clinical competence in a real-life setting may result in more lenient judgements compared with simulated task environments (Hawkins et al., 2010). For ESC assessment there are concerns that assessors' dual practice and assessment roles can interfere with the assessment process as maintaining interpersonal relations can potentially influence assessor judgements (Heaslip & Scammell, 2012). This has parallels with findings in other vocational areas, for example, Colley and Jarvis (2007) and Yaphe and Street (2003). Broadly speaking there are three ways of guarding against social bias: rater training; moderation; and scaling. It is recommended that such safeguards are employed.

Practicalities

The practicability of assessment is also a feature that influences the use of both checklist and GRS assessment approaches. In general it is considered that holistic judgements can work well in contexts that afford frequent and close observations of learner performance, for example, Curtis (2004). At the same time, contextual considerations can undermine the enactment of multiple assessment observations. Assessment in professional contexts can be resource intensive. For example, it is suggested that the validity of the mini-CEX requires different assessors to assess a range of clinical skills over time in a number of contexts and scenarios, and that this should involve 10 encounters and between 6 and 10 different assessors (Norcini et al., 2003). Similarly, the assessment of ESCs often involves mentor observations that are organised around three meeting points during a placement. Evidence suggests that it is sometimes difficult to ensure that devolved mentor assessment responsibilities are carried out at the appropriate time during the placement (Shaw, 2016). This issue has led, in part, to the development of e-portfolio tools to support the assessment process. Such systems often also facilitate combining human judgements on multiple traits and assessments, as we have mentioned. It is recommended that those making assessment judgements are involved in designing assessments to increase the manageability of the assessments.

Evidence quality

The validity of using observation as an assessment tool links to the notion that the method elicits characteristics of performance that are indicative of 'true' capability. To support this, it has been noted that past (observed) performance can be taken as a good indicator of future performance – see Adams (2012), cited in Marks (2014). The quality of an assessment relates to the quality of the evidence that is elicited through an assessment task. Therefore any justification for using observation as an assessment tool relates to the quality of the instruments that support that observation process. For both GRS and checklist approaches, there are variances around the practices that are found in different contexts, and this can undermine the confidence of assessment outcomes. For example, the reliability of the mini-CEX assessment is influenced by the number of points on the rating scale (Yates, 2013). In the case of ESC assessment, it is noted that the form of competence reporting for sign-off can differ but that the role of competent professionals (i.e., mentors) is generally limited to a sign-off function that attests to task completion rather than the quality of performance. This highlights the importance of including validation in the development and review processes.

References

Adams, R. (2012). *National Partnership Agreement on Literacy and Numeracy reporting: Measures and models for reporting gain over time*. Sydney, Australian: Council of Australian Governments Reform Council.

Ahmed, K., Miskovic, D., Darzi, A., Athanasiou, T., & Hanna, G. B. (2011). Observational tools for assessment of procedural skills: a systematic review. *The American Journal of Surgery*, *202*, 469–480. Available online from doi: 10.1016/j.amjsurg.2010.10.020

Al Ansari, A., Ali, S. K., & Donnon, T. (2013). The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. *Academic Medicine*, *88*(3), 413–420. Available online from doi: 10.1097/ACM.0b013e318280a953

Boritz, J. E., & Timoshenko, L., M. (2014). On the Use of Checklists in Auditing: A commentary. *Current Issues in Auditing*, *8*(1), C1–C25. Available online from doi: 10.2308/ciia-50741

Borneuf, A. M., & Haigh, C. (2010). The who and where of clinical skills teaching: A review from the UK perspective. *Nurse Education Today*, *30*(2), 197–201. Available online from doi: 10.1016/j.nedt.2009.07.012

Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smee, S., & Sambandam, E. (2011). Performance in assessment: Consensus statement and recommendations from the Ottawa conference. *Medical Teacher*, *33*(5), 370–383. Available online from doi: 10.3109/0142159X.2011.565831

Colley, H., & Jarvis, J. (2007). Formality and informality in the summative assessment of motor vehicle apprentices: a case study. *Assessment in Education: Principles, Policy & Practice*, *14*(3), 295–314. Available online from doi: 10.1080/09695940701591883

Curtis, D. D. (2004). The assessment of generic skills. In J. Gibb (Ed.), *Generic skills in vocational education and training: research findings* (pp.136–156). Station Arcade, Australia: National Centre for Vocational Education Research.

Domingues, R. C. L., Amaral, E., & Zeferino, A. M. B. (2009). Global overall rating for assessing clinical competence: what does it really show? *Medical Education*, *43*(9), 883–886. Available online from doi: 10.1111/j.1365-2923.2009.03431.x

Dreyfus, S. E., & Dreyfus, H., L. (1980). *A Five-Stage Model of the Mental Activities Involved in Direct Skill Acquisition:* Operations Research Center, University of California.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behaviour and Human Performance*, *7*, 86–106. Available online from doi: 10.1016/0030-5073(72)90009-8

Eining, M., Jones, D. R., & Loebbecke , J. (1997). Reliance on decision aids: An examination of auditors' assessment of management fraud. *Auditing*, *16*(2), 1–19.

Eraut, M. (1994). *Developing Professional Knowledge and Competence*. London: Falmer Press.

Eraut, M., & Steadman, S. (1998). *Evaluation of Level 5 Management S/NVQs*. Brighton: University of Sussex Institute of Education.

Feltham, R. (1988). Assessment centre decision making: judgement vs mechanical. *Journal of Occupational Psychology*, *61*, 237–241.

Gilovich, T., & Griffin, D. (2002). Introduction – heuristics and biases: then and now. In T. Gilovich, D. Griffin, & D. Kahnemann (Eds.), *Heuristics and biases: the psychology of intuitive judgement*. (pp.1 to 18). Cambridge: Cambridge University Press.

Gilovich, T., Griffin, D., & Kahnemann, D. (2002). *Heuristics and biases: the psychology of intuitive judgement*. Cambridge: Cambridge University Press.

Gingerich, A., van der Vleuten, C. P., Eva, K. W., & Regehr, G. (2014). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic Medicine*, *89*(11), 1510–1519. Available online from doi: 10.1097/ACM.0000000000000486

Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. *Academic Medicine*, *85*(9), 1453–1461. Available online from doi: 10.1097/ACM.0b013e3181eac3e6

Heaslip, V., & Scammell, J. M. (2012). Failing underperforming students: The role of grading in practice assessment. *Nurse Education in Practice*, *12*(2), 95–100. Available online from doi: 10.1016/j.nepr.2011.08.003

Highhouse, S., & Kostek, J., A. (2013). Holistic assessment for selection and placement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA Handbook of Testing and Assessment in Psychology* (Vol. 1. Test theory and Testing and Assessment in Industrial and Organizational Psychology). Washington, DC, US: American Psychological Association.

Ilgen, J. S., Ma, I. W. Y., Hatala, R., & Cook, D. A. (2015). A systematic review of validity evidence for checklists versus global rating scales in simulation based assessment. *Medical Education in Review*, *49*, 161–173. Available online from doi: 10.1111/medu.12621

Janssen, D., Holthuijsen, M., Clarebout, G., Donkers, J., Slof, B., & van der Schaaf, M. (2016). *Learning Analytics enhanced E-portfolios for Workplace Based Assessment*. Paper presented at the European Association for Research on Learning and Instruction (EARLI) Conference, SIG 1 Assessment and Evaluation, Universität München, Munich.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Doubleday.

Laming, D. (2004). *Human judgment: the eye of the beholder*. Hong Kong: Thomson Learning.

Lester, S. (2000). The Professional Accreditation of Conservator-Restorers: Developing a competence-based professional assessment system. *Assessment & Evaluation in Higher Education*, *25*(4), 407–419. Available online from doi: 10.1080/713611439

Marks, G. N. (2014). Demographic and socioeconomic inequalities in student achievement over the school career. *Australian Journal of Education*, *58*(3), 223–247. Available online from doi: 10.1177/0004944114537052

McIlroy, J. H., Hodges, B., McNaughton, N., & Regehr, G. (2002). The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. *Journal of Medical Education*, *77*(7), 725–728. Available online from doi: 10.1097/00001888-200207000-00018

McKinley, R. K., Strand, J., Ward, L., Gray, T., Lun-Jones, T., & Miller, H. (2008). Checklists for assessment and certification of clinical procedural skills omit essential competencies: a systematic review. *Medical Education*, *42*, 338–349. Available online from doi: 10.1111/j.1365-2923.2007.02970.x

Norcini, J. J. (2005). The Mini Clinical Evaluation Exercise (mini-CEX). *The Clinical Teacher*, *2*(1), 25–30. Available online from doi: 10.1111/j.1743-498X.2005.00060.x

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The Mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine*, *138*(6), 476–481. Available online from doi: 10.7326/0003-4819-138-6-200303180-00012

Nursing & Midwifery Council. (2010). *Standards for Pre-Registration Nursing Education*. London: NMC.

Pelgrim, E. A. M. (2010). In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education*, *16*(1), 131–142. Available online from doi: 10.1007/s10459-010-9235-6

Pulito, A. R., Donnelly, M. B., & Plymale, M. (2007). Factors in faculty evaluation of medical students' performance. *Medical Education*, *41*(7), 667–675. Available online from doi: 10.1111/j.1365-2923.2007.02787.x

Schön, D. (1983). *The Reflective Practitioner: How professionals think in action*. New York: Basic Books.

Shaw, S. (2016). *Go-Electronic practice assessment in action*. Paper presented at the Electronic Practice Assessment – Nurses and Midwives Conference, Anglia Ruskin University, Cambridge. Retrieved from https://www.youtube.com/watch?v=Ji8EFfqyDms&list=PLl7A9-faRl96oce_7qR5pefrqeWNJXk7M&index=4

Tversky, A., & Kahneman, D. (1982). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & T. A. (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp.3–22). Cambridge: Cambridge University Press.

Watson, A. (1994). Strategies for the Assessment of Competence. *The Vocational Aspect of Education*, *46*(2), 155–165. Available online from doi: 10.1080/0305787940460205

Weston, P. S., & Smith, C. A. (2014). The use of mini-CEX in UK foundation training six years following its introduction: lessons still to be learned and the benefit of formal teaching regarding its utility. *Medical Teacher*, *36*(2), 155–163. Available online from doi: 10.3109/0142159X.2013.836267

Winter, R. (1995). The assessment of professional competences: the importance of general criteria. In A. Edwards & P. Knight (Eds.), *Assessing Competence in Higher Education*. London: Kogan Page.

Yaphe, J., & Street, S. (2003). How do examiners decide? A qualitative study of the process of decision making in the oral component of the MRCGP examination. *Medical Education*, *37*(9), 764–771. Available online from doi: 10.1046/j.1365-2923.2003.01606.x

Yates, P. J. (2013). The Mini-CEX is not Valid or Reliable in Assessing the Clinical Competence of Higher Surgical Trainees. *The Bulletin of the Royal College of Surgeons of England*, *96*(8), 1–4. Available online from doi: 10.1308/147363513X13690603820144