RMD Open
Rheumatic & Musculoskeletal Diseases

Original research

Assessment of competences in rheumatology training: results of a systematic literature review to inform EULAR points to consider

Alessia Alunno,[1] Aurélie Najm,[2] Francisca Sivera,[3,4] Catherine Haines,[5] Louise Falzon,[6] Sofia Ramiro[7,8]

To cite: Alunno A, Najm A, Sivera F, *et al*. Assessment of competences in rheumatology training: results of a systematic literature review to inform EULAR points to consider. *RMD Open* 2020;**6**:e001330. doi:10.1136/rmdopen-2020-001330

► Supplemental material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/rmdopen-2020-001330).

Received 20 May 2020
Revised 16 July 2020
Accepted 21 August 2020

For numbered affiliations see end of article.

Correspondence to
Alessia Alunno; alessia.alunno82@gmail.com

ABSTRACT

Objective To summarise the literature on the assessment of competences in postgraduate medical training.

Methods A systematic literature review was performed within a EULAR taskforce on the assessment of competences in rheumatology training and other related specialities (July 2019). Two searches were performed: one search for rheumatology and one for related medical specialities. Two reviewers independently identified eligible studies and extracted data on assessment methods. Risk of bias was assessed using the medical education research study quality instrument.

Results Of 7335 articles in rheumatology and 2324 reviews in other specialities, 5 and 31 original studies were included, respectively. Studies in rheumatology were at variable risk of bias and explored only direct observation of practical skills (DOPS) and objective structured clinical examinations (OSCEs). OSCEs, including clinical, laboratory and imaging stations, performed best, with a good to very good internal consistency (Cronbach's α=0.83–0.92), and intrarater reliability (r=0.80–0.95). OSCEs moderately correlated with other assessment tools: r=0.48 vs rating by programme directors; r=0.2–0.44 vs multiple-choice questionnaires; r=0.48 vs DOPS. In other specialities, OSCEs on clinical skills had a good to very good inter-rater reliability and OSCEs on communication skills demonstrated a good to very good internal consistency. Multisource feedback and the mini-clinical evaluation exercise showed good feasibility and internal consistency (reliability), but other data on validity and reliability were conflicting.

Conclusion Despite consistent data on competence assessment in other specialities, evidence in rheumatology is scarce and conflicting. Overall, OSCEs seem an appropriate tool to assess the competence of clinical skills and correlate well with other assessment strategies. DOPS, multisource feedback and the mini-clinical evaluation exercise are feasible alternatives.

Key messages

What is already known about this subject?

- Assessment of competences in postgraduate training is highly heterogeneous across Europe with different times, methods and overarching strategies, and an overview of available evidence is lacking.

What does this study add?

- Evidence on assessment of competences in rheumatology training is scarce, but the available studies agree that objective structured clinical examination with clinical, laboratory and imaging stations may be an appropriate tool for this purpose.
- Data from other medical specialities point out that direct observation of practical skills, multisource feedback and the mini-clinical evaluation exercise may be feasible alternatives in rheumatology.

How might this impact on clinical practice?

- Evidence-based recommendations to harmonise assessment of competences in rheumatology training are needed.

INTRODUCTION

To date, a wide range of assessment tools are available to evaluate different educational domains in medical training. Knowledge is the priority in early years of medical training, and students are mainly trained and assessed in theoretical concepts. As training progresses, focus is shifted to the acquisition of complex medical competences, integrating knowledge with skills and attitudes to produce a positive, observable behaviour. Careful planning of assessment strategies is crucial in order to assess, not only knowledge but also competence or performance, particularly at a speciality training level.[1 2]

Both trainers and trainees benefit from assessments. Feedback motivates learners and identifies areas for improvement (formative assessment). Alternatively, validated methods of assessment of competence can measure the effectiveness of a teaching programme in achieving its objectives.[2 3] In

addition, the outcomes of assessments may be used to compare performances across centres, and to ensure the attainment of an agreed standard of a trainee (summative assessment).[3] Assessment tools used both formative and summative assessments need to be valid (ie, measure what they are supposed to measure), reliable (ie, consistent and applicable in different contexts) and feasible (ie, easy to be carried out with the available resources).[4]

Assessment can be performed in a real-life setting while trainees are in their working environment (workplace-based assessment) or in a dedicated setting (simulation-based assessment). Workplace-based assessment includes, among others, the direct observation of practical skills (DOPS) and the mini-clinical exercise. Conversely, simulation-based assessment may include oral exams, written tests and the objective structured clinical examination (OSCE), which reproduces a patient encounter.[5]

However, assessment tools are used in heterogeneous ways across European countries with different times, methods and overarching strategies.[6 7] Furthermore, strategies are planned to assess curricular competencies, and some competences or skills are considered core in one European country, but optional in others, and hence are not assessed. The lack of a harmonised strategy across Europe is a major unmet need in this field. In the era of large-scale movement of specialists across European countries, a pan-European approach to training and, eventually, to assessment of competences is advisable. This would ensure that when one country certifies a doctor as a rheumatology specialist, he/she has acquired the same core of knowledge, practical skills and other, previously agreed-upon, competences, at a given standard, independently of the country in which he/she has trained. However, specific recommendations for such an approach are currently lacking.

This systematic literature review (SLR) was developed to inform the EULAR taskforce responsible for developing the points to consider (PtC) for the assessment of competences in rheumatology training. Specifically, the SLR aims to summarise the available information on competence assessment methods and strategies within postgraduate medical training in both rheumatology and other related specialities.

METHODS

Search methodology

An SLR was conducted. The steering group of the EULAR task force to develop PtC for the assessment of competences in rheumatology training outlined the scope of the literature search, according to the Population, Instrument of interest and Measurement properties of interest approach following the Outcome in Rheumatology Trials (OMERACT) methodology.[8] The population consisted of medical doctors in speciality training (also referred to as trainees or fellows) in rheumatology or other related specialities. Instruments of interest included any assessment strategy or method, while the measurement properties of interest were validity, discrimination (including both reliability and sensitivity to change) and feasibility; at least one of them was required to be reported.

Two separate searches (online supplemental text S1 and S2) were performed, one for studies in rheumatology and another for SLRs in other medical specialities (online supplemental text S3). The searches were performed in MEDLINE, Embase, The Cochrane Database of Systematic Reviews, CENTRAL, DARE, HTA Database, NHS EED, CINAHL, Eric, Web of science, PsycINFO, PubMed Health (discontinued on 31 October 2018), Epistemonikos and Index to theses to July 2019. The PubMed Similar Articles tool was also used, and a crosscheck of the top 12 scientific journals in medical education was performed. Individual original research studies were selected in the rheumatology search. Conversely, in the medical specialities search, SLRs were retrieved and, subsequently, original studies extracted from the retrieved SLRs, using the same inclusion criteria as those in rheumatology.

Study selection, data collection and assessment of risk of bias

Two reviewers (AA and AN) independently assessed titles and abstracts according to predetermined inclusion/exclusion criteria following the OMERACT methodology (online supplemental text S4), followed by full-text review. The agreement between reviewers, calculated with the Cohen's kappa, was 0.93. Discrepancies were resolved by consensus. Data on study characteristics, investigated assessment method and the included measurement(s) of interest were extracted. Risk of bias was assessed using the medical education research study quality instrument.[9] This instrument has been developed to measure the quality of experimental and observational studies in medical education (eg, sampling strategy, validity of the assessment instrument and appropriateness of data analysis) (online supplemental table S1). In the absence of a validated cut-off value, we classified individual studies based on the medical education research study quality instrument scores as low (≥12), unclear (≥10 but <12) or high (<10) risk of bias. Similarly, we classified the validity/discrimination (including reliability and sensitivity to change)/feasibility as fair to moderate or good to very good.[10 11] Studies were too heterogeneous to allow any form of pooling; therefore, descriptive results are presented. Table 1 shows the abbreviation, full name, short description and setting of all the assessment tools mentioned in the manuscript.

RESULTS

The search in rheumatology yielded 7335 articles, of which 20 were selected for detailed review; 5 met the inclusion criteria. The search in other specialities yielded 2324 SLRs; 46 were selected for detailed review, of which 36 met the inclusion criteria. Individual studies included

Table 1 Instruments used in medical education for the assessment of competence and mentioned in the article (in alphabetical order)

Abbreviation	Full name	Description	Setting
CbD	Case-based discussion	CbD involves a comprehensive review of clinical case(s) between the trainee and an assessor.	Workplace based
DOPS	Direct observation of practical skills	DOPS requires direct observation of the trainee during a procedure, followed by a discussion and record assessing their current skill level. The number of DOPS required from a trainee varies depending on number of skills required for learning.	Workplace based
MCQ	Written multiple-choice questionnaire	MCQ is an assessment method in which trainees are asked to select the best answer(s) from the choices offered as a list. The list includes the correct answer and several plausible but incorrect answers (distractors).	Simulation based
Mini-ACE	Mini assessed clinical encounter	The mini-ACE has been developed from the mini-CEX and involves a single assessor observing the trainee (usually in an early career stage) while they conduct a patient assessment in any of a variety of settings. It enables a structured observation of an aspect of clinical practice.	Workplace based
Mini-CEX	Mini-clinical evaluation exercise	The mini-CEX was derived from the long case assessment and is a 10–20 min comprehensive assessment of the trainee–patient interaction by direct observation.	Workplace based
MSF	Multisource feedback	MSF is the is a comprehensive assessment of the trainee by different evaluators (eg, peers, patients, nurses, supervisors) using a questionnaire. It is most commonly used to evaluate professional behaviours.	Workplace based
OSCE	Objective structured clinical examinations	An OSCE usually comprises a circuit of short (5–15 min) stations, in which each trainee is examined on a one-to-one basis with one or more objective examiner(s) and either real or simulated patients (actors or electronic patient simulators). Trainees complete all the stations in a circuit.	Simulation based

in these SLRs totalled 2211, of which 347 were selected. After deduplication, 278 underwent full-text evaluation. Ultimately, 31 studies met the inclusion criteria in related medical specialities.

Assessment of competences in rheumatology

The five eligible studies on the assessment of competences in rheumatology were at variable risk of bias (two low, two unclear and one high) and explored only two methods: DOPS (one study) and OSCE (four studies)[12–16] (table 2). Rheumatology OSCEs have been used to assess clinical skills,[12–14] communication,[13] professionalism[13] and practical skills on musculoskeletal ultrasound.[15] In particular, the latter included stations with healthy subjects and patients with rheumatic diseases and trainees were blinded on whether the joint to be examined was abnormal or normal. In the abnormal stations, gouty arthritis, synovitis and erosive arthritis were represented. With regard to the other OSCEs, they encompassed different clinical scenarios such as rheumatoid arthritis and systemic lupus erythematosus along with laboratory (eg, synovial fluid analysis) and imaging (eg, synovial bone radiography) stations. Conflicting evidence on internal consistency, intermethod and inter-rater reliability was reported in these five studies. One study at low risk of bias demonstrated that a rheumatology OSCE including a mix of clinical, laboratory and imaging stations showed a good to very good internal consistency (Cronbach's α=0.83–0.92), intrarater reliability (correlation coefficient=0.80–0.95) and construct validity.[12] A fair to moderate correlation (r=0.44–0.52) between OSCEs and other assessment tools, including DOPS, rating by programme directors and written exam was also found.[12–14 16] In the specific ultrasound OSCE, the assessment of normal joint more reliably discriminated examinees from the ultrasonography experts (control population) than the evaluation of pathologic joints; inter-rater reliability was also better for normal joint assessment stations.[15] The OSCE scores of normal joint stations correlated with the scores of a written multiple-choice questionnaire exam; both the overall scores of the OSCE and the multiple-choice questionnaire exam showed a poor discrimination of performing examinees from the faculty. The study on DOPS provided evidence only for feasibility[16] reporting that 14 forms per resident over the time frame of a month provide a reliable estimate. None of the studies on OSCE provided evidence on feasibility.

Assessment of competences in other medical specialities

Studies in related specialities were more heterogeneous in terms of assessment tools and both type and comprehensiveness of the analysis. The 31 eligible studies were at variable risk of bias (15 low, 14 unclear and 2 high) and

Table 2 Assessment of competences in rheumatology

Author, country, year	Berman JR, USA, 2009[13]	Kissin EY, USA, 2014[15]	Pascual-Ramos V, Mexico, 2015[12]	Pascual-Ramos V, Mexico, 2018[14]	Humphrey-Murto S, Canada, 2009[16]
Competence assessed	Clinical skills, communication skills, professionalism	Practical skills (MSUS)	Clinical skills	Clinical skills	CanMEDS roles
Tool*	OSCE	OSCE	OSCE	OSCE	DOPS†
Comparator tool	Rating by programme directors	MCQ (76-question)	Theory Test Board Exam (300 MCQ)	Theory Test Board Exam (200–222 MCQ)	OSCE (national 10-stations)
Sample size	70 (number of participants to year 4 NA)	35	68	80	73
Workplace based or simulation based	Simulation based	Simulation based	Simulation based	Simulation based	Workplace based
Current practice (CP) or research purpose (RP)	CP	RP	CP	CP	RP
Control population	None	3 Faculty members	3 Certified rheumatologists	≥6 Certified rheumatologists	None
Study duration	5 years	NA	2 years	2 years	18 months
Internal consistency (Cronbach's α unless otherwise stated)‡	–	r=0.15	0.83–0.92	–	–
Inter-rater reliability	Correlation between raters not significant	ICC 0.3 between live assessors and assessors on videotapes ICC 0.7 between assessors on videotapes	–	–	–
Test–retest reliability/intrarater reliability	–	–	OSCE total and partial scores in Year 1 vs Year 2 r=0.80–0.95	–	–
Intermethod (or parallel forms) reliability/ concurrent validity	OSCE versus rating by programme directors r=0.48	OSCE versus MCQ r=0.52	OSCE total score versus MCQ Year 1: r=0.203; Year 2: r=0.436	OSCE total score versus MCQ Year 1: r=0.277; Year 2: r=0.445	DOPS versuss OSCE r=0.48
Feasibility	–	–	–	–	14 Forms per resident to achieve a g coefficient of 0.8
Construct validity	–	OSCE scores significantly discriminated fellows and fellows versus faculty	Significantly higher scores in higher levels of training	–	–
Predictive validity	–	–	–	–	–
Risk of bias (MERSQI score)	Unclear (11)	High (9.5)	Low (14.5)	Unclear (10)	Low (12.5)

*The number of OSCE stations was 6–8 of 12 min each[13], 9[15], 12 or 15 of 8 min each.[12 14.]
†Rated with the ambulatory clinic evaluation form.
‡The range of the study[12] indicates the values across the stations repeated in the 2 academic years.
DOPS, direct observation of practical skills; ICC, intraclass correlation coefficient; MCQ, multiple-choice questionnaire; MERSQI, Medical Education Research Study Quality Instrument; MSUS, musculoskeletal ultrasound; NA, not available; OSCE, objective structured clinical examination.

explored different methods, including OSCE, DOPS, multisource feedback, mini-clinical evaluation exercise and patient satisfaction questionnaires. Online supple mental tables S2 and S3 show the information of individual studies.

Simulation-based assessment (OSCE)

As far as simulation-based assessment is concerned, evidence on internal consistency of an OSCE assessing clinical skills was conflicting. The majority of studies at low risk of bias reported a fair to moderate internal consistency (Cronbach's α=0.12–0.69),[17–20] while most studies at high risk of bias reported a good to very good internal consistency (Cronbach's α=0.8–0.98)[21–24] (table 3). Nevertheless, the majority of studies exploring inter-rater reliability agreed that OSCEs assessing clinical skills have a good to very good inter-rater reliability (r=0.60–0.95).[17 18 22 23] Conversely, OSCEs assessing communication skills consistently demonstrated a good to very good internal consistency (Cronbach's α=0.7–0.98),[23 25 26] while evidence on inter-rater reliability was conflicting.[23 25–27] However, OSCE scores poorly correlated with those of other assessment tools such as oral exams,[17 23] written exams,[28 29] in-training examinations,[18–20] assessment by staff and peers[25] or the American Board of Internal Medicine evaluation form.[19] Finally, with regard to feasibility, 10–14 OSCE stations would provide a reliable estimate of both clinical[30] and communication skills.[27] Simulation-based assessment can also rely on the use of standardised patient encounters to evaluate clinical and communication skills. One study at unclear risk of bias demonstrated that, in standardised patient encounters, nonverbal communication was most closely associated with patient satisfaction, with a good to very good internal consistency.[31] Although scores on clinical skills obtained in the setting of standardised patient encounters poorly correlated with those of the American Board of Internal Medicine,[32] one study at low risk of bias exploring patient satisfaction as an indicator of the trainee's clinical skills demonstrated that this assessment is feasible and has a good internal consistency but a poor inter-rater reliability.[33]

Workplace-based assessment (DOPS, mini-assessed clinical encounter, case-based discussion, mini-clinical evaluation exercise, multisource feedback)

With regard to workplace-based assessment, one study at low risk of bias reported that DOPS showed a good to very good internal consistency (Cronbach's α≥0.8), inter-rater reliability (r=0.83–0.87) and a good prediction of the American Board of Internal Medicine certifying examination scores in internal medicine.[34] Conversely, in the field of psychiatry, DOPS did not correlate with any other assessment tool investigated such as the mini-assessed clinical encounter or the case-based discussion and was also less feasible (table 3).[35]

Three studies at low risk of bias provided evidence of a good to very good internal consistency (Cronbach's α=0.65–0.90) and feasibility of the mini-clinical evaluation exercise,[36–38] which also showed a good correlation with other assessment tools such as the American Board of Internal Medicine monthly evaluation form,[37] or the Royal College of Physicians and Surgeons of Canada Comprehensive Examination in Internal Medicine[38] (table 4). Likewise, most studies on multisource feedback reported a good to very good internal consistency (Cronbach's α=0.65–0.90), feasibility and inter-rater reliability.[39–42] However, results on validity and reliability of these two tools were conflicting.

Online supplemental table S4 displays the remaining studies on written exams and script concordance test.

DISCUSSION

Our SLR has shown a large heterogeneity in the strategies and methods used for assessment of competences in the training of rheumatology and other medical specialities. Specifically, evidence in rheumatology is scarce with all studies published over the last 10 years, while in other specialities including internal medicine and paediatrics, research on such educational matters has been ongoing for at least 25 years.[21 43] This study attempted to overcome the lack of data in the field of rheumatology by exploring other related medical specialities. However, many of the investigated tools were speciality specific (eg, OSCEs with intubation simulators for anaesthesiology trainees[44]) and therefore non-relevant to the field of rheumatology; studies exploring these tools were excluded. Furthermore, a crucial aspect emerging from the SLR is the difficulty of comparing studies. It is challenging to explore the same tools even within the same speciality due to the heterogeneity of the specific instruments and the context of their application, the measurement properties evaluated and the data analysis. A wide variety of statistical methods have been employed to determine the properties of interest of the investigated assessment strategies, and in some cases, the analysis provided is insufficient or not robust. For example, despite being rejected as an adequate measure of inter-rater reliability, some studies continue to report the percentage of agreement between raters instead of an intraclass correlation coefficient or kappa.[45 46]

Overall, with regards to rheumatology training, the SLR provides enough evidence only for OSCEs, as other assessment tools were not sufficiently investigated. Initially developed to address the unreliability of traditional strategies of clinical assessment, such as the long case discussion,[47] OSCEs are extensively used in undergraduate medical training.[48] The key concepts behind the OSCEs are standardisation and generalisability, as all candidates deal with the same clinical tasks to be completed in the same time frame, in the same environment and are scored through a structured checklist. OSCE stations can be designed and tailored for specific

Table 3 Characteristics of the studies on objective structured clinical examination (OSCE) and direct observation of practical skills (DOPS) in other specialities

Tool	OSCE	OSCE	OSCE	DOPS[34]	DOPS[35]
Competence assessed	**Clinical skills**	**Communication skills**	**Knowledge**	**Clinical skills**	**Overall competence**
Number of studies	11	4	2	1	1
Current practice or research purpose	Current practice (N=5)[17–19 28 29] Research purpose (N=6)[20–24 30]	Current practice (N=1)[25] Research purpose (N=3)[23 26 27]	Current practice (N=1)[28] Research purpose (N=1)[24]	Current practice	Research purposes
Comparator tool*	Intraining examination (ITE) (N=4)[18–21] Oral exam (N=2)[17 23] Written exam (N=2)[28 29] National board exam (N=1)[19] Resident performance ratings (RPR) (N=1)[21] Faculty global evaluation (N=1)[18] None (N=3)[22 24 30]	Oral exam (N=1)[23] OSCE clinical scores (N=1)[25] None (N=2)[26 27]	None (N=2)[24 28]	None	Case-based discussion, (mini) assessment of clinical expertise, peer assessment tool; patient satisfaction questionnaire, case conference, journal club presentation
Sample size, median (IQR)†	43 (21–74)	29 (12–69)	58[24] and 147[28]	110	600
Specialties	General practice (GP) (N=3)[18 28 29] Paediatrics (N=3)[20 21 24] Anesthesiology (N=3)[17 23 30] Internal medicine (N=1)[19] Emergency medicine (N=1)[22]	GP (N=2)[25 26] Respiratory medicine (N=1)[27] Anaesthesiology (N=1)[23]	GP (N=1)[28] Paediatrics (N=1)[24]	Internal medicine	Psychiatry
External control population	Medical students (N=1)[21] GP specialists (N=1)[29] None (N=9)[17–20 22–24 28 30]	Internal medicine residents and pulmonology attending physicians (N=1)[27] None (N=3)[23 25 26]	None (N=2)[24 28]	None	None
Study duration, median (IQR)	Reported (N=8)[17–20 22 24 28 29] 80 (4.5–156) weeks Not reported (N=3)[21 23 30]	Reported (N=1)[25] 4 weeks Not reported (N=3)[23 26 27]	4[24] and 6[28] weeks	156 weeks	36 weeks
Internal consistency	Cronbach's α=0.12–0.99 (N=8)[18–24 28] Intraclass correlation coefficient=0.40 (N=1)[30] r correlation coefficient=0.27–0.32 (N=1)[17]	Cronbach's α=0–0.98 (N=4)[23 25–27]	Cronbach's α=0.70–0.99 (N=2)[24 28]	≥0.8	Not performed
Inter-rater reliability	r=0.26–0.95 (N=4)[17 18 23 30]	r=0.26–0.88 (N=4)[23 25–27]	N=0	0.83–0.87‡	Not performed

Continued

Table 3 Continued

Tool	OSCE	OSCE	OSCE	DOPS[34]	DOPS[35]
Competence assessed	**Clinical skills**	**Communication skills**	**Knowledge**	**Clinical skills**	**Overall competence**
Test–retest reliability/ intrarater reliability	r=0.62–0.72 (N=1)[24]	N=0	r=0.62–0.72 (N=1)[24]	Not performed	Not performed
Intermethod (or parallel forms) reliability/ concurrent validity	OSCE versus ITE: no correlation (N=1)[19] OSCE versus ITE: r=0.30–0.71 (N=3)[18 20 21] OSCE versus oral exam: r=0.14–0.54 (N=2)[17 23] OSCE score versus written exam: r=0.54 (N=1)[29] OSCE versus National board exam: no correlation (N=1)[19] OSCE versus RPR: r=0.41 (N=1)[21] OSCE versus faculty global evaluation: r=0.03–0.51 (N=1)[18]	OSCE versus oral exam: r=0.52–0.53 (N=1)[23] No correlation between clinical and communication scores of OSCE (N=1)[25]	OSCE versus written exam: r=0.22 (N=1)[28]	Not performed	None of the tools correlate with each other (all r values <0.65)
Feasibility	Approximately 100 case presentations through OSCE to reach a g coefficient of 0.8 (N=1)[19] 10–12 OSCE case presentations through OSCE with three to four judges each rating all the cases to achieve a g coefficient of 0.8 (N=1)[30]	A g coefficient of 0.8 can be feasibly achieved with 14 case presentations through OSCE (N=1)[27]	N=0	Not performed	It is possible to achieve a 0.8 reliability in a feasible way only with case based discussion and (mini) assessment of clinical expertise
Predictive validity	N=0	N=0	N=0	The ratings predicted performance on the national certifying examination	Not performed
Risk of bias	Low (N=6)[17–20 28 29] Unclear (N=4)[21–23 30] High (N=1)[24]	Low (N=3)[25–27] Unclear (N=1)[23] High (N=0)	Low (N=1)[28] Unclear N=0 High (N=1)[24]	Low	Low

*Studies may use more than one comparator tool. For details, see online supplemental table S2 displaying individual studies.
†If less than three studies, individual values are shown.
‡Calculation method developed in James LR *et al. J Appl Psychol*. 1984;69:85–98.
Numbers (N) indicate the number of studies.

Table 4 Characteristics of the studies on the mini-clinical exercise (mini-CEX), multisource feedback (MSF) and standardised patient encounters (SPE) in other specialities

Tool	Mini-CEX	MSF	MSF	MSF	SPE[32]	SPE
Competence assessed	**Clinical skills**	**Clinical skills**	**Interpersonal/ communication skills**	**Overall competence**	**Clinical skills**	**Interpersonal/ communication skills**
Number of studies	3	2	2	2	1	2
Current practice or research purpose	Current practice (N=2)[36 37] Research purpose (N=1)[38]	Current practice (N=1)[33] Research purpose (N=1)[42]	Current practice (N=0) Research purpose (N=2)[41 42]	Current practice (N=0) Research purpose (N=2)[39 40]	Research purpose	Research purposes (N=2)[31 32]
Comparator tool*	National board exam (N=2)[37 38] In-training examination (ITE) (N=1)[37] None (N=1)[36]	Criterion audit, patient satisfaction questionnaires, assessment of referral letters, significant event analysis (N=1)[33] None (N=1)[42]	None (N=2)[41 42]	None (N=2)[39 40]	National board exam	National board exam (N=1)[32] Patient satisfaction checklist (PSC) (N=1)[31]
Sample size, median (IQR)†	23 (22–88)	16[42] and 171[42]	70[41] and 16[42]	18[40] and 56[39]	48	48[32] and 59[31]
Specialties	Internal Medicine (N=3)[36–38]	Obstetrics and gynecology (N=1)[42] General practice (N=1)[33]	Internal medicine and pulmonary medicine (N=1)[41] Obstetrics and gynecology (N=1)[42]	Physical medicine and rehabilitation (N=2)[39 40]	Internal medicine	Internal medicine
External control population	None (N=3)[36–38]	None (N=2)[33 42]	None (N=2)[41 42]	None (N=2)[39 40]	None	None (N=2)[31 32]
Study duration, median (IQR)	52 (52–156) weeks	52 weeks[42] Not reported (N=1)[33]	52[42] and 72[41] weeks	156[39] and 208[40] weeks	72 weeks	72 weeks (N=1)[32] Not reported (N=1)[31]
Internal consistency	Cronbach's α=0.65–0.90 (N=3)[36–38]	Cronbach's α=0.80–0.81 (N=1)[33] Intraclass correlation coefficient 0.23–0.90 (N=1)[42]	Cronbach's α=0.96 (N=1)[41] Intraclass correlation coefficient=0.23–0.90 (N=1)[42]	Cronbach's α=0.77–0.99 (N=2)[39 40]	Not performed	Cronbach's α=0.7–0.83 (N=1)[31]
Inter-rater reliability	N=0	r=0–0.86 (N=2)[33 42]	r=0.2–0.86 (N=2)[41 42]	Significant difference if medical students are among raters (N=1)[39] No difference among raters (N=1)[40]	Kendall's W 0.3–0.84	Kendall's W 0.3–0.84 (N=1)[32]
Test–retest reliability/ intrarater reliability	N=0	r=0.20–0.80 (N=1)[33]	None (N=2)[41 42]	None (N=2)[39 40]	r=0.32–0.8	r=0.32–0.8 (N=1)[32]

Continued

Table 4 Continued

Tool	Mini-CEX	MSF	MSF	MSF	SPE[32]	SPE
Competence assessed	**Clinical skills**	**Clinical skills**	**Interpersonal/ communication skills**	**Overall competence**	**Clinical skills**	**Interpersonal/ communication skills**
Intermethod (or parallel forms) reliability/ concurrent validity	MiniCEX versus national board exam r=0.67–0.81 (N=2)[37 38] MiniCEX versus ITE r=0.5–0.6 (N=1)[37]	N=0	None (N=2)[41 42]	None (N=2)[39 40]	r=0.16–0.5	SPE versus national board exam r=0.16–0.5 (N=1)[32] SPE versus PSC r=0.6–0.76 (N=1)[31]
Feasibility	12–14 encounters	Yes (N=1)[33]	Up to 20 ratings only by attending physicians and nurses (N=1)[41]	5 Ratings from research nurses, 4 from other staff and 23 from medical students. With other raters not possible to feasibly achieve the target. (N=1)[39]	Not performed	Not performed
Predictive validity	N=0	N=0	N=0	N=0	Not performed	Non-verbal communication is the best predictor of patient satisfaction (N=1)[31]
Risk of bias	Low (N=2)[36 37] Unclear (N=1) High (N=0)[38]	Low (N=1)[33] Unclear (N=1)[42] High (N=0)	Low (N=0) Unclear (N=2)[41 42] High (N=0)	Low (N=0) Unclear (N=2)[39 40] High (N=0)	Unclear	Unclear

*Studies may use more than one comparator tool. For details, see online supplemental table S3 displaying individual studies.
†If less than three studies, individual values are shown.
Numbers (N) indicate the number of studies.

purposes and should always be closely aligned to the relevant training curriculum, in order to demonstrate construct validity. In rheumatology training, current evidence suggests that OSCEs including clinical, laboratory and imaging stations are the most reliable and demonstrate good content validity.[12 14] In most of the studies, not only in rheumatology but also in other specialities, clinical OSCE scores moderately correlated with those of other assessment tools such as rating by programme directors and written exams.[12–20 23 25 28 29] This probably reflects that methods measure different dimensions of competence; hence, a combination of different (complementary) tools is advisable to obtain an overall perspective on the trainee. The lack of correlation between an imaging-specific OSCE and a written exam suggests that the latter may assess whether physicians can identify abnormalities on a static image, but who might not yet be skilled enough to obtain the relevant images independently.[15] OSCEs can be applied beyond assessment of clinical skills, as they allow assessment of non-clinical competences, such as communication or professionalism. Medical speciality curricula highlight that, upon training completion, specialists should have acquired skills beyond clinical knowledge, and these competences can be difficult to assess. Communication skills for appropriate interaction with patients, caregivers and colleagues are included, among others, in the European, American and Canadian frameworks for the competences that doctors should have. In Europe, the European Union of Medical Specialists defines professional competence as 'the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served'.[49]

With regard to workplace-based assessment, the only study exploring DOPS in rheumatology provided evidence of feasibility but did not explore any other relevant features of this assessment method.[16] Furthermore, authors acknowledged that only 10% of the evaluations reflected a true, direct observation, with 80% resulting from case discussion with the trainee and 10% from a review of the written notes. Therefore, the resulting feasibility should be considered with caution, especially because two additional studies in other specialities failed to prove its feasibility.[34 35] Evidence from other specialities such as internal medicine and physical medicine/rehabilitation underpins the reliability and feasibility of mini-clinical evaluation exercise and multisource feedback,[36–42] but no data on these methods are available in rheumatology. The similarities between these two specialities and rheumatology suggest that both mini-clinical evaluation exercise and multisource feedback might be successfully employed in rheumatology to assess clinical and broader generic competences, respectively. The multisource feedback has the potential to contribute to the professional development of trainees. It provides a comprehensive trainee overview resulting from observation over a long period, under natural circumstances, and may include people not responsible for formal judgements about trainees. The lack of correlation between raters involved in multisource feedback is considered a strength of the assessment method as different raters (eg, nurses, patients and programme directors) focus on different skills and attitudes.[41]

Although the need to tackle the assessment of competences in rheumatology postgraduate training has been highlighted for the past 20 years,[5] it still remains an unmet need. Despite a consistent body of evidence on assessment of competences in other specialities, data in rheumatology are scarce and conflicting. Owing to its good to very good internal consistency, intrarater reliability, construct validity and moderate correlation with other assessment strategies, OSCEs with clinical, laboratory and imaging stations appear an appropriate tool to assess clinical competences in rheumatology. Based on evidence from other specialities, DOPS, multisource feedback and mini-clinical evaluation exercise are feasible alternatives.

Considering the increasing use of technology in the medical field, one could envisage that at least for some assessment tools and strategies, online-based platforms may represent a good alternative and become the routine. In several universities across Europe, digital portfolios are already implemented in rheumatology training,[50] and the scoring of recordings may replace direct observation of trainees performing a certain procedure, either in the workplace or in an OSCE station, as already shown in emergency medicine and paediatrics.[22 51] Furthermore, the availability of imaging software may facilitate the assessment of MRI or radiography readings at distance. Finally, yet importantly, the recent coronavirus pandemic has dramatically increased the use of remote teaching and assessment and probably set the stage for major changes in the way education will be delivered in the future.

In conclusion, the results of the present SLR further underscore this gap with other medical specialities and highlight the need to develop recommendations to harmonise strategies and methods for the assessment of competences across Europe. This SLR informs the ongoing initiative to formulate EULAR PtC for the assessment of competences in rheumatology training.

Author affiliations
[1]Rheumatology Unit, University of Perugia Department of Medicine, Perugia, Italy
[2]University Hospital, Inserm Umr 1238, Nantes, France
[3]Department of Rheumatology, Hospital General Universitario Elda, Elda, Spain
[4]Department of Medicine, Miguel Hernandez University of Elche, Elche, Spain
[5]King's College London, London, UK
[6]Center for Personalized Health, Northwell Health Feinstein Institutes for Medical Research, Manhasset, New York, USA
[7]Department of Rheumatology, Leiden University Medical Center, Leiden, Netherlands
[8]Department of Rheumatology, Zuyderland Medical Centre Heerlen, Heerlen, Netherlands

Twitter Aurélie Najm @AurelieRheumo and Alessia Alunno @alessia_alunno.

Contributors All authors contributed and finally approved the current manuscript.

Funding This work was funded by European League Against Rheumatism.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information.

ORCID iDs

Alessia Alunno http://orcid.org/0000-0003-1105-5640
Aurélie Najm http://orcid.org/0000-0002-6008-503X

REFERENCES

1 Wass V, Van Der Vleuten C, Shatzer J, *et al.* Assessment of clinical competence. *Lancet* 2001;357:945–9.
2 Aaron S. Moving up the pyramid: assessing performance in the clinic. *J Rheumatol* 2009;36:1101–3.
3 Doherty M. Specialist training in rheumatology-the need for a curriculum and assessment. *Ann Rheum Dis* 1994;53:215–7.
4 Dacre J, Haq I. Assessing competencies in rheumatology. *Ann Rheum Dis* 2005;64:3–6.
5 Epstein RM. Assessment in medical education. *N Engl J Med* 2007;356:387–96.
6 Sivera F, Ramiro S, Cikes N, *et al.* Working Group on Training in Rheumatology across Europe. Differences and similarities in rheumatology specialty training programmes across European countries. *Ann Rheum Dis* 2015;74:1183–7.
7 Sivera F, Ramiro S, Cikes N, *et al.* Working Group on Training in Rheumatology across Europe. Rheumatology training experience across Europe: analysis of core competences. *Arthritis Res Ther* 2016;18:213.
8 Available https://omeracthandbook.org/handbook
9 Reed DA, Cook DA, Beckman TJ, *et al.* Association between funding and quality of published medical education research. *JAMA* 2007;298:1002–9.
10 Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res Sci Educ* 2018;48:1273–96.
11 Taylor R. Interpretation of the correlation coefficient: a basic review. *J Diagn Med Sonogr* 1990;6:35–9.
12 Pascual-Ramos V, Medrano Ramírez G, Solís Vallejo E, *et al.* Performance of an objective structured clinical examination in a national certification process of trainees in rheumatology. *Rheumatol Clin* 2015;11:215–20.
13 Berman JR, Lazaro D, Fields T, *et al.* The New York City rheumatology objective structured clinical examination: five-year data demonstrates its validity, usefulness as a unique rating tool, objectivity, and sensitivity to change. *Arthritis Rheum* 2009;61:1686–93.
14 Pascual-Ramos V, Guilaisne Bernard-Medina A, Flores-Alvarado DE, *et al.* The method used to set the pass mark in an objective structured clinical examination defines the performance of candidates for certification as rheumatologists. *Rheumatol Clin* 2018;14:137–41.
15 Kissin EY, Grayson PC, Cannella AC, *et al.* Musculoskeletal ultrasound objective structured clinical examination: an assessment of the test. *Arthritis Care Res (Hoboken)* 2014;66:2–6.
16 Humphrey-Murto S, Khalidi N, Smith CD, *et al.* Resident evaluations: the use of daily evaluation forms in rheumatology ambulatory care. *J Rheumatol* 2009;36:1298–303.
17 Berkenstadt H, Kantor GS, Yusim Y, *et al.* The feasibility of sharing simulation-based evaluation scenarios in anesthesiology. *Anesth Analg* 2005;101:1068–74.
18 Skinner BD, Newton WP, Curtis P. The educational value of an OSCE in a family practice residency. *Acad Med* 1997;72:722–4.
19 Petrusa ER, Blackwell TA, Ainsworth MA. Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. *Arch Intern Med* 1990;150:573–7.
20 Hilliard RI, Tallett SE. The use of an objective structured clinical examination with postgraduate residents in pediatrics. *Arch Pediatr Adolesc Med* 1998;152:74–8.
21 Joorabchi B. Objective structured clinical examination in a pediatric residency program. *Am J Dis Child* 1991;145(7):750–54.
22 Williams JB, McDonough MA, Hilliard MW, *et al.* Intermethod reliability of real-time versus delayed videotaped evaluation of a high-fidelity medical simulation septic shock scenario. *Acad Emerg Med* 2009;16:887–93.
23 Savoldelli GL, Naik VN, Joo HS, *et al.* Evaluation of patient simulator performance as an adjunct to the oral examination for senior anesthesia residents. *Anesthesiology* 2006;104:475–81.
24 Hergenroeder AC, Laufman L, Chorley JN, *et al.* Development and evaluation of a method for evaluating pediatric residents' knowledge and skill in performing physical examinations of the ankle and knee. *Pediatrics* 2001;107:E51.
25 Leung KK, Wang WD, Chen YY. Multi-source evaluation of interpersonal and communication skills of family medicine residents. *Adv Health Sci Educ Theory Pract* 2012;17:717–26.
26 Van Nuland M, Van Den Noortgate W, Degryse J, *et al.* Comparison of two instruments for assessing communication skills in a general practice objective structured clinical examination. *Med Educ* 2007;41:676–83.
27 Boudreau D, Tamblyn R, Dufresne L. Evaluation of consultative skills in respiratory medicine using a structured medical consultation. *Am J Respir Crit Care Med* 1994;150:1298–304.
28 Kramer AW, Jansen JJ, Zuithoff P, *et al.* Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice. *Med Educ* 2002;36:812–9.
29 Kramer AW, Jansen KJ, Düsman H, *et al.* Acquisition of clinical skills in postgraduate training for general practice. *Br J Gen Pract* 2003;53:677–82.
30 Weller JM, Robinson BJ, Jolly B, *et al.* Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. *Anaesthesia* 2005;60:245–50.
31 Griffith CH, Wilson JF, Langer S, *et al.* House staff nonverbal communication skills and standardized patient satisfaction. *J Gen Intern Med* 2003;18:170–4.
32 Day RP, Hewson MG, P K Jr, *et al.* Evaluation of resident performance in an outpatient internal medicine clinic using standardized patients. *J Gen Intern Med* 1993;8:193–8.
33 Murphy DJ, Bruce DA, Mercer SW, *et al.* The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. *Adv Health Sci Educ Theory Pract* 2009;14:219–32.
34 Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine resident evaluation form detect differences in clinical competence? *J Gen Intern Med* 1994;9:140–5.
35 Brittlebank A, Archer J, Longson D, *et al.* Workplace-based assessments in psychiatry: evaluation of a whole assessment system. *Acad Psychiatry* 2013;37:301–7.
36 Norcini JJ, Blank LL, Arnold GK, *et al.* The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;123:795–9.
37 Durning SJ, Cation LJ, Markert RJ, *et al.* Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Acad Med* 2002;77:900–4.
38 Hatala R, Ainslie M, Kassen BO, *et al.* Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. *Med Educ* 2006;40:950–6.
39 Massagli TL, Carline JD. Reliability of a 360-degree evaluation to assess resident competence. *Am J Phys Med Rehabil* 2007;86:845–52.
40 Musick DW, McDowell SM, Clark N, *et al.* Pilot study of a 360-degree assessment instrument for physical medicine & rehabilitation residency programs. *Am J Phys Med Rehabil* 2003;82:394–402.
41 Woolliscroft JO, Howell JD, Patel BP, *et al.* Resident-patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. *Acad Med* 1994;69:216–24.
42 Davis JD. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstet Gynecol* 2002;99:647–51.

43 Norcini JJ, Swanson DB, Grosso LJ, *et al.* A comparison of knowledge, synthesis, and clinical judgment. Multiple-choice questions in the assessment of physician competence. *Eval Health Prof* 1984;7:485–99.
44 Mudumbai SC, Gaba DM, Boulet JR, *et al.* External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. *Simul Healthc* 2012;7:73–80.
45 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
46 Krippendorff K. *Content analysis: an introduction to its methodology*. Beverly Hills, CA: Sage Publications, 1980.
47 Gormley G. Summative OSCEs in undergraduate medical education. *Ulster Med J* 2011;80:127–32.
48 Khan KZ, Ramachandran S, Gaunt K, *et al.* The Objective Structured Clinical Examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. *Med Teach* 2013;35:e1437–1446.
49 Available https://www.uems.eu/__data/assets/pdf_file/0005/44438/UEMS-2014.21-European-Training-Requirements-Rheumatology-.pdf
50 Van Onna M, Ramiro S, Haines C, *et al.* On behalf of working group on development of a EULAR portfolio of rheumatology. Establishing the key components of a EULAR portfolio for training in rheumatology: a EULAR school of rheumatology initiative. *Ann Rheum Dis* 2020;79:531.
51 Grant EC, Grant VJ, Bhanji F, *et al.* The development and assessment of an evaluation tool for pediatric resident competence in leading simulated pediatric resuscitations. *Resuscitation* 2012;83:887–93.