Rater Training in Medical Education: A Scoping Review

Ashley Vergis 1, Caleb Leung 1, Reagan Roberston 1

1. Surgery, St. Boniface Hospital, University of Manitoba, Winnipeg, CAN

Corresponding author: Ashley Vergis, umerg99@gmail.com

Abstract

There is an increasing focus in medical education on trainee evaluation. Often, reliability and other psychometric properties of evaluations fall below expected standards. Rater training, a process whereby raters undergo instruction on how to consistently evaluate trainees and produce reliable and accurate scores, has been suggested to improve rater performance within behavioral sciences. A scoping literature review was undertaken to examine the effect of rater training in medical education and address the question: "Does rater training improve performance attending physician evaluations of medical trainees?" Two independent reviewers searched PubMed®, MEDLINE®, EMBASE™, the Cochrane Library, CINAHL®, ERIC™, and PsycInfo® databases and identified all prospective studies examining the effect of rater training on physician evaluations of medical trainees. Consolidated Standards of Reporting Trials (CONSORT) and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklists were used to assess quality. Fourteen prospective studies met the inclusion criteria. All had heterogeneity in design, type of rater training, and measured outcomes. Pooled analysis was not performed. Four studies examined rater training used to assess technical skills; none identified a positive effect. Ten studies assessed its use to evaluate non-technical skills: six demonstrated no effect, while four showed a positive effect. The overall quality of studies was poor to moderate. Rater training in medical education literature is heterogeneous, limited, and describes minimal improvement on the psychometric properties of trainee evaluations when implemented. Further research is required to assess rater training’s efficacy in medical education.

Introduction And Background

In many fields, including medicine, measuring performance is limited to subjective observational judgments. Recent changes to traditional medical education present new challenges for training physicians. Initiatives towards competency-based training have caused many programs to introduce the use of standardized, outcomes-based clinical assessment tools. However, the psychometric properties of these tools remain insufficient for high-stakes testing, with reliability often below desired benchmarks. Although several means to improve reliability exist, many studies fail to suggest or examine these options. One method to improve the reliability of assessments is to attempt to improve the objectivity of raters [1].

Rater training (RT) is a process whereby raters undergo instruction on how to evaluate trainees best and produce reliable and accurate scores. RT was developed in an effort to address the natural bias introduced by subjective performance assessments. There is compelling evidence in the behavioral and social sciences literature to suggest that RT can improve rater performance [2]. The process is thought to work by focusing on optimizing the standardized use of a tool and limiting the effect of individual preconceived notions [3]. RT is commonly used in these disciplines to improve the psychometric properties of a variety of observational assessment tools [1-6]. In a landmark study on RT methods, Woehr and Huffcutt classified RT into four different types, including (1) rater error training, (2) behavioral observation training, (3) performance dimension training, and (4) frame-of-reference training [2].

Rater error training educates raters regarding common rating errors such as halo, central tendency, and leniency. Evidence of rating errors is generally considered to reflect a considerable inaccuracy degree within an evaluation [2]. Generally, specific errors are defined, and the raters are then given strategies on how to avoid them [7]. For example, raters may be informed to look for both good and bad performance features and avoid forming overall impressions to prevent halo [4]. Behavioral observation training instructs raters to observe and record behavior as opposed to forming global judgments. Raters are taught to anticipate specific behaviors within a dimension and make a careful record of these observations to improve recall of particular events [2, 7]. An example would be classifying a subject based on the exact number of times a specific behavior was present, or action was performed. Performance dimension training educates raters on the specific dimensions used to evaluate trainees before the observation is begun. Understanding each dimension can then guide rater observation and subsequent evaluation. Each dimension is clearly defined, and examples of actions and behaviors associated with the dimension are given. Having raters participate in assessment tool development or familiarizing raters with the tool prior to observation are examples of
performance dimension training [2]. The frame of reference (FOR) training builds a common construct between raters, which they use to observe and evaluate subjects. Raters are instructed on performance standards for each dimension. Rating tool dimensions and their associated behaviors are defined, leading FOR training to often include aspects of behavioral observation and performance dimension training. The desired level of performance to each specific rating is explained in order to create a shared definition between raters of an appropriate ranking for an observed performance. In some instances, an element of rater practice, discussion, and feedback is incorporated into the training to further develop the shared criteria [2, 7].

Despite RT’s proven effectiveness in other fields, RT has not been widely studied in medical and surgical education. Many studies of commonly used clinical assessment tools have commented on their insufficient psychometric properties, particularly poor reliability [8-10]. Even so, many studies fail to suggest or examine the available options to improve reliability. Existing options include increasing the number of assessments, modifying the tools themselves, or improving rater objectivity, such as through the use of RT. In medicine, significant time constraints for both trainees and evaluators make increasing the number of assessments extremely challenging [11]. Changing existing tools creates problems pertaining to having multiple versions of a similar tool, requiring re-validation of the modified instrument [12]. Therefore, we performed a scoping review to examine what is currently known about the effect of RT on trainee assessments in medical education.

This review is composed based on the master thesis (Maniar R, The Effect of Rater Training on the Reliability and Validity of Technical Skill Assessments: A Randomized Control Trial. Faculty of Graduate Studies of The University of Manitoba, Department of Surgery; 2016).

**Review**

**Methods**

A search for original publications until January 2020 was performed using PubMed®, MEDLINE®, EMBASE™, the Cochrane Library, CINAHL®, ERIC™, and PsycInfo®. The inclusion criteria of the search were prospective studies with RT for physicians as a primary intervention, where some formal description of the training was given. Additionally, some form of the control group was needed, although it was not limited to any specific type so long as it was present. Studies using pre- and post-training comparisons were eligible, as were studies comparing trained raters to an untrained group. Specific psychometric properties such as reliability or validity had to be specified as an outcome variable. Articles were excluded if the RT intervention was not described, if the subjects undergoing training were not attending physicians, if there was no comparison group, or if an outcome variable was not specified. Review articles were also excluded. The specific search terms were ‘rater training’ AND ‘medical’ OR ‘surgical education’. A manual review of the selected articles’ references was also performed to ensure search completion. Two authors assessed articles for eligibility for inclusion. Any disagreements were resolved by a third author at each stage.

**Results**

The initial search strategy found 529 papers for abstract review, with an additional 30 papers found by the manual search. Forty papers were selected for full-text review. Fourteen papers met the criteria and were selected for final inclusion (Figure 1) [10, 11, 15-24].
FIGURE 1: Search strategy for scoping review on rater training in medical education

The included studies with a description of their methods and findings can be found in the Appendix. The quality of included studies was assessed using modifications of the Consolidated Standards of Reporting Trials (CONSORT) and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklists (see Appendix).

Included studies had marked heterogeneity in terms of their design, methods, and type of RT. Ten of the studies were randomized trials, and four were cohort studies. Seven of the studies specifically included surgeons; however, the majority of the studies looked at rating non-technical skills. Only four studies measured assessments of technical skill [13-16]. The majority of the studies had raters evaluating a videotaped clinical encounter. Of these, five used a standardized, scripted encounter as opposed to real patient interactions. Two studies assessed performance in training evaluation reports (ITERS), and a final study assessed a surgical oral exam.

The training was in a workshop format for eight of the studies and ranged from an hour-long training session to a four-day course. The remainder of the studies utilized a training video. Although all studies described their RT intervention, only eight studies specified the intervention using well-defined terminology such as "rater error training" or "FOR training". Types of training ranged from using a single format to incorporating all four types of RT into a workshop. Most studies compared the training group to a group of untrained raters. Two studies compared raters pre- and post-training and George, et al. compared an intensive RT program to an accelerated version [14]. Seven of the studies looked at interrater reliability (IRR), although a variety of statistics were used, including interclass correlations, Cronbach’s alpha, and Kendall’s coefficient. The outcomes assessed in the remaining studies included accuracy (four studies), correlation coefficients (one study), and assessment quality (two studies). One study examined the effect of RT on the reliability and validity of the psychometric outcomes.

Rater Training for Technical Skill Assessments

Of the four studies that assessed RT for assessments of technical skill, all were unable to show an effect of training. In the first study by Robertson et al., a FOR training video showed no statistically significant effect on the IRR of forty-seven attending surgeons randomized to RT versus no training and assessing simple suturing and instrument knot tying videos of ten trainees [15]. The performance was assessed using a procedure-specific checklist, visual analog scale, and a modified Objective Structured Assessment of
Technical Skills (OSATS) Global Rating Scale (GRS) [23]. Interclass correlation coefficients (ICC) were measured to assess IRR. Although there was a trend towards improved ICC with RT, this was not statistically significant.

Using the same study population, Robertson et al. also examined the effect of FOR training on the reliability and validity of multiple psychometric assessment tools for surgical suturing and knot-tying, including a pass-fail assessment, a visual analogue scale, a modified OSATS GRS, and a task-specific checklist [16]. Rates were randomized to RT versus no RT. Assessments of trainee suturing and knot-tying videos were made at the initial start of the study and then after a delay of two weeks. Internal consistency and reliability were measured with Cronbach’s alpha and IRR scores. Validity was assessed using univariate and multivariate analyses. Although there was a trend towards improvement of all three domains, there were no statistically significant differences after RT when compared to no RT.

Similarly, in a study by Rogers et al., a rater error training video had no effect on the IRR of eight surgeons evaluating a simple two-handed knot-tying video using a standardized checklist [13]. Cronbach’s alpha was used to assess IRR. Scores were high for both groups regardless of training (0.80 for the untrained group vs. 0.71 for the trained group), limiting the ability to show a difference between groups. Additionally, rater error training was used, which has been shown to not be the most effective training method, especially for improving reliability. The training group was noted to give more specific comments in their feedback, which the authors suggested may be indicative of some unmeasured effect of training on rater behavior.

The other study assessed technical skills, randomizing surgeon raters to either an accelerated or immersive FOR training session using the Zwisch OR performance GRS [14]. The accuracy of the two groups was measured by comparing scores to an expert consensus score. Although the immersive group had a slightly higher accuracy of 88% as compared to 80% for the accelerated group, this was not statistically significant. There was no difference in the overall Zwisch GRS scores between groups. Although this study was not significant, the difference in accuracy scores was close to achieving significance, suggesting a possible effect of training. The study may have been underpowered to detect a difference as forty-four surgeons were randomized, but only ten were in the non-training group. Accuracy was also assessed using expert consensus scores, which may not be the best measure of a tool’s psychometric properties. Reliability and validity are generally more critical, especially if the tool is to be used for high-stakes purposes.

Rater Training for Non-Technical Skill Assessments

Of the remaining ten studies assessing the evaluation of non-technical skills, four studies had a positive outcome. Holmboe et al. randomized internal medicine attending physicians to a four-day performance dimension and FOR training workshop on using the mini Clinical Evaluation Exercise (CEX) tool [10]. The mini-CEX is an observational instrument that uses GRS to assess a trainee’s clinical interaction with a patient [26]. After training, the tool was used to assess scripted, videotaped clinical encounters. After adjusting for baseline rating and program, the trained group had significant improvement in IRR that persisted for eight months after the workshop [10]. Van der Vleuten, et al. showed a one-hour training session improving the accuracy of medicine attendings using a checklist to evaluate videotaped history and physical skills as compared to those randomized to no training [17]. In two papers by Dudek et al., a home training session for medicine attendings was shown to improve the quality of ITER assessments post-training [18, 19]. The remaining six studies were unable to show an effect of RT [11, 20-24].

Quality of Evidence

The overall quality of the included studies was poor to moderate (see Appendix). Only four of the randomized studies adequately described their randomization techniques, and one was minimally described. Three studies did not explicitly state their eligibility criteria. Only seven studies included a clear and detailed explanation of their training intervention, four had a moderate description, while three studies had only a brief description. Three of the randomized studies did not state if there were any baseline differences between training groups. Three studies failed to list any significant limitations, and in two, this was only very superficially discussed. Finally, only two studies included any type of power calculation, and one of these failed to achieve their desired recruitment. Many of the studies were small and thus may have been underpowered to detect a difference between groups. Notably, the only high-quality study in the group by Holmboe et al. was able to show a significant and prolonged effect of RT [10].

Discussion

Rater training within the social sciences has been demonstrated to have a positive effect on common rating measures. A meta-analysis of 29 comparative studies found moderate effectiveness in all four rater training domains: i.e., rater error training resulted in reduced halo effect or leniency errors, performance dimension training reduced halo error, FOR training improved increased rating accuracy and behavioral observation training results in improved observational accuracy. Among all four domains, FOR training appeared to be the most effective intervention to improve rating accuracy. Although limited in number, several studies also examined the role of combined RT strategies and demonstrated positive effects on various rater errors.
among small sample sizes [2].

With the continued shift in medical training towards a competency-based medical education and increased evaluation, there is a need to ensure reliable and accurate trainer assessments. Our review found that within the current medical education literature, RT has not been demonstrated to significantly improve various psychometric properties of medical trainee assessments. However, there was a trend towards a positive effect in specific outcomes such as IRR, with the most studied intervention involving FOR training.

One response to the lack of success in improving the psychometric properties of assessments in medical education has been to rethink the approach to assessment. Newer theories advocate moving away from the traditional viewpoint of objective, quantifiable measures as the gold standard in favor of more subjective assessments. Proponents of this model argue that the traditional approach is limited because it reduces the evaluation of complex aptitudes of medical trainees into individual quantifiable skills. This results in a loss of overall “gestalt” and limits how to address the evaluation of certain key aspects of modern healthcare delivery, such as team-based work and collaboration [25]. Others have sought to address the concept of RT within contemporary frameworks by seeking to understand rater error from the perspective of the psychological sciences, specifically impression formation literature. Such theories suggest if rater-based assessments are understood as a psychological and social judgment phenomenon, educators may be better equipped to address and correct the issue of rater error [27].

The decision to focus this review on the traditional psychometric properties of standardized assessments was for two reasons. Firstly, medical education remains focused on quantitative assessment. The advent of competency-based medical education demonstrates this, as it, by definition, seeks to identify and evaluate individual domains required by medical trainees to succeed. Therefore, the continued study and improvement of how evaluations are made and their quality measured are needed, particularly for high-stakes testing. Secondly, technical skills assessment may be most appropriately evaluated by traditional standardized assessments. Other skills, such as clinical decision making or working within team-based healthcare systems, may be less amenable to standardized evaluation forms. However, a technical skill assessment is well suited to this type of measurement. This is reflected by the development, ongoing application, and widespread use of standardized tools such as OSATS in surgical education [25]. Improving the use of these assessment tools within technical fields of medical education may be necessary. For example, by investigating more optimal forms of rater training.

Limitations of our review include marked heterogeneity among studies in terms of the study population, rater training intervention, and measured outcomes. The overall quality of studies was poor to moderate, and no pooled analysis was possible. It is clear that although RT may represent a means to improve the reliability of skill assessments, further high-quality studies are needed to determine the role of RT within medical and surgical education.

Conclusions

Future research also should investigate the optimal format and duration of rater training for each individual setting or tool. Within the literature examining RT in medical education, there is a lack of evidence on the ideal training format, as there is nearly unlimited variability in the way training can be administered. This makes it difficult to know the best starting point when developing a new training intervention. A variety of training formats have been described, varying from in-person tutorials to training videos, the use of single or multiple RT types, and sessions ranging from less than an hour to multiple day workshops.

Appendices
<table>
<thead>
<tr>
<th>Study</th>
<th>Type of intervention</th>
<th>Intervention details</th>
<th>Primary Outcomes</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rogers et al. 2002 [13]</td>
<td>Does RT improve IRR</td>
<td>Randomized RT or none; not described how randomized</td>
<td>No training</td>
<td>Small; simple task; short training, high IRR regardless of teaching</td>
</tr>
<tr>
<td>Canada Single centre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanager et al. 2004 [14]</td>
<td>Does RT improve reliability rating</td>
<td>Four-hour training workshop (FOR, RET, IRT, PDT)</td>
<td>Nine scripted videos of OR encounters, rated before and immediately after RT</td>
<td>Non-technical skills; voluntary/assigned participation; non-randomized; no effect</td>
</tr>
<tr>
<td>Denmark Single centre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cook et al. 2008 [22]</td>
<td>Effect RT on IRR and accuracy of mini-CEX scores (internal medicine clinical exam) (five-point GRS)</td>
<td>Randomized controlled (21/54 declined randomizing)</td>
<td>16 scripted videos; four weeks after training</td>
<td>Randomized, the high number declined randomization one-month delay between training and rating; no effect</td>
</tr>
<tr>
<td>Germany Single centre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weitz et al. 2014 [24]</td>
<td>Does RT improve the accuracy of assessment of physical examination skills? (five-point German grading code)</td>
<td>Randomized controlled to RT or none</td>
<td>242 students undergoing 10-minute physical exam skills assessment with a standardized patient</td>
<td>Reference rating using video-based reassessment of all 242 participants; no effect on accuracy</td>
</tr>
<tr>
<td>US Single centre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>George et al. 2013 [16]</td>
<td>Determine type of FOR training for reliable and accurate use of assessment of surgeons using the Zwisch scale (four-point GRS)</td>
<td>Quasi-experimental (randomized)</td>
<td>Immersive – FOR with videos and testing, discussion</td>
<td>Proportion correct response for accuracy (80.2% immersive vs 88% accelerated); Spearman coefficient for correlation accuracy (r,0.90 immersive vs 0.93 accelerated); Cronbach’s alpha for rater bias (0.045 immersive vs 0.049 accelerated).</td>
</tr>
<tr>
<td>US Multi centre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noel et al. 1992 [11]</td>
<td>Determine the accuracy of faculty evaluations of residents clinical skills, how structured form and RT improve evaluations (structured form included four-point GRS)</td>
<td>Quasi-experimental; open form vs structured form</td>
<td>Two scripted videos of resident history and physical examination on standardized patient; rated immediately post RT</td>
<td>Accuracy scores (% correct) = no difference between RT and no training with structured form (64 vs 65% RT for case 1); and 63% vs 54% RT for case 2). Regression showed</td>
</tr>
<tr>
<td>US Multi centre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heimbach et al. 2004 [15]</td>
<td>Evaluate the efficacy of direct observation of competence training to change rating behavior (nine-point GRS)</td>
<td>Cluster randomized controlled trial; stratified, sealed envelopes; rated pre- and post-training</td>
<td>Nine scripted tapes (three cases at three levels of performance), rated eight</td>
<td>High quality; positive effect; faculty had to be active in teaching for nomination; no differences at baselines</td>
</tr>
<tr>
<td>US Multi centre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Value of RT on the reliability of scores &amp; examiner selection for a clinical exam</td>
<td>Physicians (surgeon and family doctor, prior examiners (physicians n=22))</td>
<td>FOR training with practice 1.5 hours (n=11)</td>
<td>Physicians who supervise medical trainees (n=69; only 37 returned at all required time points)</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------</td>
<td>-------------------------------------------------</td>
<td>------------------------------------------</td>
<td>-----------------------------------------------------------------</td>
</tr>
<tr>
<td>Dudek et al. 2013 Canada Multi-centre (four schools)</td>
<td>Does training improve the quality of ITERS for medicine residents (CCERR form used to assess ITERS quality)</td>
<td>Randomized trial of five types of training – varying stages of feedback guide</td>
<td>Surgical faculty (n=16)</td>
<td>2.5 hour FOR training with videos, practice (n=15)</td>
</tr>
<tr>
<td>Dudek et al. 2012 Canada Multicentre (three sites)</td>
<td>Effectiveness of workshop at improving ITERS</td>
<td>Uncontrolled pre/post-training design</td>
<td>Physicians who supervise trainees and complete ITERS (n=23)</td>
<td>Three-hours workshop (explain good FITER, recognize challenges)</td>
</tr>
<tr>
<td>Ludbrook et al. 1971 Australia Single centre</td>
<td>Does examiner training decrease inter-examiner variability for the clinical skills exam</td>
<td>Randomized trial – not stated how or if for sure random; marked in pair (both trained or untrained)</td>
<td>Surgical faculty (n=16)</td>
<td>2.5 hour FOR training with videos, practice (n=15)</td>
</tr>
<tr>
<td>Van der Vlaeten et al. 1989 US Single centre</td>
<td>Does training increase the accuracy of assessments of clinical skills (history &amp; physical standardized checklist)</td>
<td>Randomized trial (three groups – doctors, med. students and lay people)</td>
<td>Physicians (surgeons and family doctor), prior examiners (physicians n=22)</td>
<td>FOR training with practice 1.5 hours (n=11)</td>
</tr>
<tr>
<td>Vleuten et al. 1989 Canada Single centre</td>
<td>Randomized controlled trial, rated pre- and post-training</td>
<td>Limited 30- minute training, individual, PDT (n=10), extensive two-hour group training, PDT &amp; FOR with practice (n=6)</td>
<td>No training (rated two months later) (n=6)</td>
<td>Five videos of real students with standardized patient (students not aware standardized); rated a few days after RT</td>
</tr>
<tr>
<td>Dudek et al. 2013 Canada Multi-centre (four schools)</td>
<td>Randomized controlled trial, rated pre- and post-training</td>
<td>Limited 30- minute training, individual, PDT (n=10), extensive two-hour group training, PDT &amp; FOR with practice (n=6)</td>
<td>No training (rated two months later) (n=6)</td>
<td>Five videos of real students with standardized patient (students not aware standardized); rated a few days after RT</td>
</tr>
<tr>
<td>Ludbrook et al. 1971 Australia Single centre</td>
<td>Randomized controlled trial, rated pre- and post-training</td>
<td>Limited 30- minute training, individual, PDT (n=10), extensive two-hour group training, PDT &amp; FOR with practice (n=6)</td>
<td>No training (rated two months later) (n=6)</td>
<td>Five videos of real students with standardized patient (students not aware standardized); rated a few days after RT</td>
</tr>
<tr>
<td>Dudek et al. 2013 Canada Multi-centre (four schools)</td>
<td>Randomized controlled trial, rated pre- and post-training</td>
<td>Limited 30- minute training, individual, PDT (n=10), extensive two-hour group training, PDT &amp; FOR with practice (n=6)</td>
<td>No training (rated two months later) (n=6)</td>
<td>Five videos of real students with standardized patient (students not aware standardized); rated a few days after RT</td>
</tr>
</tbody>
</table>

TABLE 1: Summary of studies included in the rater training scoping review

BOT - behavioral observation training; FOR - frame of reference training; PDT - Performance dimension training; RET - rater error training; GRS - Global Rating Scale; OSATS - Objective Structured Assessment of Technical Skills; RT - rater training; NOTSS - non-technical skills for surgeons; CEX - Clinical Evaluation Exercise; CCERR - completed clinical evaluation report rating; ITER - in training evaluation reports; FITER - final in-training evaluation report; RE - rater error; OR - operating room; ICC - interclass correlation coefficients

<table>
<thead>
<tr>
<th>Reference</th>
<th>Abstract or summary described</th>
<th>Randomization described</th>
<th>Blinding or concealment described</th>
<th>Sample size calculation</th>
<th>N lost to follow up</th>
<th>Eligibility criteria given</th>
<th>Intervention described</th>
<th>Ethics approval</th>
<th>Baseline characteristics listed</th>
<th>Limits discussed</th>
<th>Quality and generalizability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robertson et al. [15]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Well described</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Moderate</td>
</tr>
<tr>
<td>Robertson et al. [16]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Well described</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Moderate</td>
</tr>
<tr>
<td>Rogers et al. [17]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Briefly described</td>
<td>Not known</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Poor; not generalizable</td>
</tr>
<tr>
<td>Sporangar et al. [18]</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Well described</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Moderate</td>
</tr>
<tr>
<td>Cook et al. [19]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Briefly described</td>
<td>No</td>
<td>Yes</td>
<td>Yes, brief</td>
<td>Moderate</td>
<td>Poor; not generalizable</td>
</tr>
<tr>
<td>Witzl et al. [20]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Well described</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Moderate</td>
</tr>
<tr>
<td>George et al. [21]</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Well described</td>
<td>Not known</td>
<td>Yes, brief</td>
<td>No</td>
<td>Moderate</td>
</tr>
<tr>
<td>Noel et al. [22]</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Briefly described</td>
<td>No</td>
<td>Yes, brief</td>
<td>No</td>
<td>Moderate</td>
</tr>
<tr>
<td>Holmboe et al. [23]</td>
<td>Yes</td>
<td>Yes (minimal)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Well described</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>High; generalizable</td>
</tr>
<tr>
<td>Neale et al. [24]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Moderately described</td>
<td>Not known</td>
<td>Yes, brief</td>
<td>No</td>
<td>Poor; not generalizable</td>
</tr>
<tr>
<td>Van der Vaalen et al. [25]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Moderately described</td>
<td>Not known</td>
<td>No</td>
<td>Yes</td>
<td>Poor; not generalizable</td>
</tr>
<tr>
<td>Ludbrook et al. [26]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Moderately described</td>
<td>Not known</td>
<td>Yes, brief</td>
<td>Yes</td>
<td>Poor; not generalizable</td>
</tr>
<tr>
<td>Daukis et al. 2013 [27]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes (not met)</td>
<td>Yes</td>
<td>Yes</td>
<td>Moderately described</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Moderate generalizable</td>
</tr>
<tr>
<td>Daukis et al. 2013 [28]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Well described</td>
<td>Not known</td>
<td>Yes</td>
<td>Yes</td>
<td>Moderate generalizable</td>
</tr>
</tbody>
</table>

TABLE 2: Quality assessment of included studies in the scoping review

Additional Information

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: Payment/services info: All authors have declared that no financial support was received from any organization for the submitted work. Financial relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. Other relationships: All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References
