The Cognitive Behaviour Therapist (2024), vol. 17, e23, page 1 of 22
doi:10.1017/S1754470X24000321

BABCP
www.babcp.com

ORIGINAL RESEARCH

The Quality of Supervision Questionnaire – associations between quality, person and context variables

Ulrike Maaß[1], Franziska Kühne[1], Alexandra Schöttler[2] and Florian Weck[1]

[1]Department of Clinical Psychology and Psychotherapy, University of Potsdam, Potsdam, Germany and [2]Psychologisch-Psychotherapeutisches Institut (PPI), UP Transfer GmbH at University of Potsdam, Potsdam, Germany
Corresponding author: Ulrike Maaß; Email: ulrikemaass@uni-potsdam.de

(Received 31 October 2023; revised 8 July 2024; accepted 9 July 2024)

Abstract

Only a few instruments can monitor the quality of individual supervision sessions. Therefore, the first objective was to develop a brief Quality of Supervision Questionnaire (QSQ). The second objective was to examine person and context variables associated with more effective supervision sessions. Two online samples of $n = 374$ psychotherapy trainees and $n = 136$ supervisors were used to develop the QSQ using exploratory factor analysis, validity and reliability analyses, and tests for measurement invariance. In addition, correlations between the QSQ and person and context variables were examined. The final QSQ included 12 items and three factors (Effectiveness, Procedural Knowledge, Relationship). The supervisee version had good reliability ($\alpha = .83$ to .88) and correlated moderately to strongly with convergent measures ($r = .37$ to .68). The supervisor version was partially invariant to the supervisee version, displayed weak to good convergent validity ($r = .27$ to .51) and mixed reliabilities ($\alpha = .67$ to .81). Regarding person variables, higher session quality was positively associated with supervisee self-efficacy ($r = .16$) and being a supervisor (*vs* supervisee, $d = 0.33$ to 0.56). Regarding context variables, there were significant effects for supervisors in cognitive behaviour therapy (*vs* psychodynamic therapy; in terms of Procedural Knowledge, $d = 0.86$) and for competence feedback (*vs* no feedback; $d = 0.47$ to 0.68), but not for individual (*vs* group-based) sessions. Overall, the QSQ is a valid and reliable self-report questionnaire. We discuss the conceptual overlap between supervision scales.

Key learning aims

As a result of reading this paper, readers will:

(1) Be aware of the Quality of Supervision Questionnaire (QSQ), which is a brief self-report scale assessing the quality of individual supervision sessions with 12 items and three subscales: Effectiveness, Procedural Knowledge, and Relationship.
(2) Learn that there are no significant differences in the quality of supervision between sessions in individual and group formats. Compared with psychodynamic supervisors, supervisors in cognitive behaviour therapy report more procedural knowledge (i.e. what exactly to do and how to do it) in their sessions.
(3) Understand that supervisees evaluate sessions that included competence feedback as qualitatively better than supervisees who did not receive competence feedback.

Keywords: group supervision; measurement; self-efficacy; supervision process; supervisory alliance; therapist competence

Introduction

Clinical supervision is considered one of the most important training methods for psychotherapist development (Rønnestad *et al.*, 2018). Within supervision research, there is variance regarding the quality of supervision. For example, some studies report that up to one-third of clinical

and counselling psychologists are indeed somewhat or very dissatisfied with their supervision (e.g. McMahon and Errity, 2013). There is also evidence that marginalized supervisees in particular report emotional difficulties, for example, due to microagressions in supervision (Jendrusina and Martinez, 2019). In turn, systematic reviews with a focus on cognitive behaviour therapy and related approaches report that supervisees are highly satisfied with supervision, experience positive relationships with their supervisors, indicate higher job satisfaction, and gain feelings of self-efficacy and self-awareness, whereas the effects on supervisees' skill development remain unclear (Alfonsson *et al.*, 2018; Kühne *et al.*, 2019; Milne and James, 2000). It is therefore important to regularly monitor the quality of supervision sessions and adjust them if necessary.

Conceptual overlap

When evaluating supervision, researchers use terms such as 'quality', 'effectiveness', 'satisfaction' and 'supervisory relationship', whereby satisfaction and relationship are the most frequent variables of interest (Milne, 2018). These variables are not always clearly defined and reciprocally influence one another. For example, the terms *quality* and *effectiveness* are defined as the successful promotion of skills and techniques and securing quality of treatment but also include the quality of the *supervisory relationship* (Winstanley and White, 2011; Zarbock *et al.*, 2009). Corresponding measurements (e.g. the MCSS-26 and STEP-SV) often include relationship subscales regarding feeling understood and supported by supervisors. The *supervisory relationship*, in turn, often refers to the dyadic relationship, including broader cultural, educational and evaluative aspects (Beinart, 2014). Consequently, some relationship instruments also include subscales regarding skill improvement (e.g. Supervisory Relationship Questionnaire; Palomo *et al.*, 2010). Similarly, *satisfaction with supervision* is regarded as 'the supervisee's subjective perception of how well or badly supervision progressed' (Gonsalvez *et al.*, 2017: p. 95). In this case, it is up to the supervisee to decide on what he or she bases his or her assessment, for example, on skills, methodological factors, trust and respect (e.g. the Supervisee Satisfaction Questionnaire; Ladany *et al.*, 1996). Given these intertwined terms and definitions, it is not surprising that the *supervisory relationship* and supervision *effectiveness* are strong predictors of *supervisee satisfaction* (i.e. $r = .81$ across 27 studies; Park *et al.*, 2019).

For these reasons, many researchers advocate for distinguishing between *satisfaction* and *effectiveness* (e.g. Binder 1993; Gonsalvez *et al.*, 2017; Milne and James, 2000; O'Donovan and Kavanagh, 2014). Satisfying supervision might not automatically be effective, just as satisfying doughnuts are not automatically nutritious (Goodyear and Bernard, 1998). In line with this idea is the finding that psychotherapy trainees who did or did not use a feedback system in supervision did not report any significant differences in satisfaction with supervision, although trainees who used that system achieved better client outcomes (Reese *et al.*, 2009). Thus, supervision evaluation should be clearly defined and include aspects such as learning aside from *satisfaction* (Kühne *et al.*, 2019; Lambert and Ogles, 1997). However, only a few questionnaires address the question of whether supervision is effective. In addition, researchers have repeatedly criticized the low validity of measurements and reliance on cross-sectional designs (Alfonsson *et al.*, 2018; Kühne *et al.*, 2019; Watkins *et al.*, 2021). Furthermore, most of the questionnaires are rather long (e.g. Supervisory Relationship Questionnaire, 67 items; Palomo *et al.*, 2010) and are not accessible free of charge (e.g. the Manchester Clinical Supervision Scale-26, MCSS-26; Winstanley and White, 2011). Finally, they are not suitable for capturing information about the effectiveness of single sessions in particular (e.g. the Supervisory Working Alliance, SWAI; Efstation *et al.*, 1990; the Evaluation Process within Supervision Inventory, EPSI; Lehrman-Waterman and Ladany, 2001).

All in all, considering the previous literature (Gonsalvez *et al.*, 2017; Maiwald *et al.*, 2019; Milne, 2007; O'Donovan and Kavanagh, 2014; Winstanley and White, 2011; Zarbock *et al.*, 2009), supervision quality can be defined as a combination of *satisfaction*, *relationship* and *effectiveness*. *Satisfaction* is the subjective evaluation of whether supervisees and supervisors experience a

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

positive emotional response to the session. *Relationship* is the positive interaction between supervisee and supervisor, including an emotional bond. *Effectiveness* is the extent to which supervisees and supervisors believe that their expectations are met and the session's benefits are greater than its costs by achieving intended outcomes such as professional development and improvement of supervisees' therapeutic competencies.

Supervision measurements that focus on effectiveness and single supervision sessions

Evaluating supervision session-by-session would improve the investigation of variables that are associated with person and context variables contributing to effective or ineffective supervision, supervisory relationship rupture and repair processes, or the transfer of tasks developed in supervision to the therapy room. In addition, it enables an examination of the variation of the relevant variables over time. For example, Ybrandt *et al.* (2016) showed that the supervisory relationship fluctuated from the perspectives of novice therapists in their group supervision. In the first phase of therapy, the relationship with their supervisors worsened and then improved but decreased again towards the end of therapy. The authors emphasize that especially for a good relationship with new therapists, it is important to respond to their fears and lack of self-confidence in supervision.

To our knowledge, there are only two promising instruments for examining single supervision sessions, the Short Scale to Evaluate Supervision and Supervisor Competence (SE-SC8; Gonsalvez, 2021) and the Questionnaire to Evaluate Supervision (STEP-SV; Zarbock *et al.*, 2009): Gonsalvez and colleagues (Gonsalvez, 2021; Gonsalvez *et al.*, 2017) developed several self-report scales to monitor (group) supervisor competence based on the competence framework of supervision (American Psychological Association, 2014), including normative (e.g. demonstrating expertise), formative (enhancing reflective skills), and restorative (e.g. being supportive) skills. The 8-item short scale (SE-SC8; Gonsalvez, 2021) demonstrated adequate convergent and discriminant validity and good internal reliability in a sample of $n = 122$ supervisees. However, the scale primarily assesses supervisor competence and is therefore of limited use for a general assessment of supervision quality from the perspectives of supervisees and supervisors.

The STEP-SV is a 12-item questionnaire for supervision effectiveness, with both supervisee and supervisor versions. It showed good internal consistency in a sample of $n = 90$ trainees and $n = 37$ supervisors. Although convergent or discriminant validities are still unknown, they demonstrated concurrent validity insofar as the Relationship and Clarifying subscales positively predicted satisfaction with supervision. However, the authors indicated that some items from the supervisor version did not load on the expected factors in an exploratory factor analysis (EFA). Together with the low or non-significant correlations between supervisee and supervisor perspectives, this could mean that the items were understood differently by each group. Thus, an examination of the equivalence of 'supervision effectiveness' across both groups, for example, by testing measurement invariance (Putnick and Bornstein, 2016), has yet to be conducted. Finally, the authors also state that their items might be too generic and 'confounded with consumer and provider satisfaction and does not only assess the quality of supervision' (Zarbock *et al.*, 2009: p. 202). For these reasons, psychometrically sound self-report measures are needed to study the processes of effective supervision sessions in more detail (Knox and Hill, 2021; Watkins *et al.*, 2021).

Variables contributing to effective supervision sessions

Regardless of the exact definition of the constructs, satisfaction with and the effectiveness of supervision depend on many variables that are associated with either the supervisees/supervisors or the context of supervision. For example, *at the person level*, fear of negative evaluation can lead to non-disclosure of errors in clinical supervision (Gray *et al.*, 2001), which, in turn, is associated with lower satisfaction with supervision. In contrast, there is a positive correlation between supervision satisfaction and supervisee self-efficacy (Kühne *et al.*, 2019). At the *contextual level*,

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

the applied techniques and the setting may also influence supervision quality. For example, the most frequently used technique in supervision is case discussion, followed by information brokering and recommending literature (Weck *et al.*, 2017). However, providing competence feedback in particular and using role-playing are positively correlated with an improved supervisory relationship and satisfaction with supervision (Lehrman-Waterman and Ladany, 2001; Maiwald *et al.*, 2019; Reiser and Milne, 2016; Weck *et al.*, 2017). Furthermore, group supervision is one of the most common supervision modes and has numerous advantages over individual sessions (Ögren *et al.*, 2014). These include the opportunity to learn from peers and receive feedback. However, group supervision also poses particular challenges in terms of group dynamics, time management, and supervisory relationships (Hedegaard, 2020; Ögren *et al.*, 2014). The extent to which the quality of supervision differs in group or individual sessions is still an open question. In addition, concepts of supervision differ between theoretical orientations (Nelson, 2014; Weck *et al.*, 2017). Two of the most discussed orientations are psychodynamic therapy (PDT) and cognitive behaviour therapy (CBT). While PDT focuses mainly on the transference and counter-transference of supervisees, CBT supervision is more goal-oriented and addresses supervisees' assumptions that hinder the therapeutic process. However, Weck *et al.* (2017) found rather small differences between CBT and PDT. As only a few studies have investigated the relationships between contextual and personal variables and the effectiveness of single supervision sessions, this study aims to contribute to answering these research questions.

Objectives

The present study has two goals. First, we aimed to develop a brief self-report questionnaire (i.e. the Quality of Supervision Questionnaire; QSQ) that (a) focuses on effectiveness instead of satisfaction, (b) is suitable for evaluating the quality of single supervision sessions, and (c) can be used by both supervisees and supervisors to closely monitor the supervision process from both perspectives. The second goal of this study was to better understand which variables are associated with more effective supervision sessions. On the one hand, we examined the relationships between the QSQ and *person variables*, such as being a supervisee or supervisor, and supervisee fear of negative evaluation and self-efficacy. On the other hand, we studied *context variables*, such as the setting (e.g. individual *vs* group supervision, CBT *vs* PDT) and specific techniques (e.g. case discussion, competence feedback). We expected positive associations between the QSQ and self-efficacy and competence feedback but negative correlations with fear of negative evaluation. However, the analyses with respect to the supervisee–supervisor differences and the setting were exploratory without directed hypotheses.

Method

Participants, recruitment, and cligibility criteria

Between March and August 2022, psychotherapy trainees (supervisees) and supervisors were recruited by contacting 254 training institutes across Germany via email. All institutes offer state-approved training in one of four treatments (i.e. CBT, psychodynamic therapy, psychoanalysis, and systemic therapy) that are recognized under social law. The survey was viewed by 724 people, 526 of whom responded to the questions. The inclusion criteria were (a) participation in psychotherapy training or work as a licensed supervisor and (b) provided informed consent ($n = 16$ did not provide consent and were excluded). The final sample consisted of $N = 510$ participants ($n = 374$ supervisees, $n = 136$ supervisors). Table 1 shows the sample characteristics for a subset of the data ($n = 354$), namely, for those participants who also provided information on their demographic variables at the end of the survey. On average, the *supervisees* based their answers on a supervision session that occurred primarily in an individual ($n = 150$, 60.0%) rather than a group-based setting ($n = 100$, 40.0%). It was scheduled on average 9.46 days after the actual

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

Table 1. Sample characteristics

	Supervisees		Supervisors	
	n	%	*n*	%
Gender identity				
Women	224	88.5	69	68.3
Men	26	10.3	32	31.7
Diverse	1	0.4	—	—
No indication	2	0.8	—	—
Age, *M* (*SD*), range	34.14 (7.84)	24–67	55.83 (11.14)	33–80
Theoretical orientation				
Cognitive behaviour therapy	159	62.8	56	55.4
Psychodynamic therapy	81	32.0	21	20.8
Other/multiple	13	5.2	24	23.8
Target group				
Adults	165	65.2	62	61.4
Children/youth	61	24.1	17	16.8
Both	27	10.7	22	21.8
No. of supervisors/supervisees[a], *M* (*SD*), range	2.92 (1.32)	1–8	11.37 (9.75)	1–50

The sample size was $n = 101$ for supervisors and $n = 250$–253 for supervisees. [a]Either the number of current supervisors with whom a supervisee is working, or the number of supervisees a supervisor is currently supervising.

treatment session ($SD = 11.63$; range 0–115) and lasted approximately 79.99 minutes ($SD = 44.31$, range: 22–240; note that group-based supervision must be conducted every fourth therapy session). At the time of recruitment, the supervisees were currently treating 7.36 patients ($SD = 3.82$, range: 0–27) in in-patient and/or out-patient settings on average, but most of whom were in out-patient settings ($n = 218$, 86.5%; in-patient: $n = 34$, 13.5%). On average, the *supervisors* had been working as supervisors for 12.4 years ($SD = 10.34$, < 1 year to 49 years). They based their answers on individual supervision rather than group supervision ($n = 86$, 85.1% *vs* $n = 15$ group supervision, 14.9%), and sessions lasted approximately 68.76 minutes ($SD = 38.37$, range: 45–240).

Sample size and power

The sample size was guided by the analysis requiring the largest sample (i.e. EFA). Recommendations vary between authors and reach from at least $n = 300$ to 400 (Goretzko *et al.*, 2021). Thus, we aimed for a sample size of at least $n = 300$.

Item generation

Items were developed using top-down and bottom-up procedures: (1) reviewing the literature and deriving a definition for *quality of supervision*, (2) considering existing supervision measures, (3) interviewing six supervisees and supervisors about their experiences, and (4) revising the first items. In accordance with the definitions of 'supervision quality' described above, including its elements satisfaction, relationship, and effectiveness, we formulated items (e.g. satisfaction: 'All my concerns were addressed satisfactorily'; relationship: 'We were a good team'; and effectiveness: 'The session was effective'). Items of existing supervision measures (Efstation *et al.*, 1990; Gonsalvez *et al.*, 2017; Ladany *et al.*, 1996; Lehrman-Waterman and Ladany, 2001; Palomo *et al.*, 2010; Winstanley and White, 2011; Zarbock *et al.*, 2009) were considered when they matched our definitions and were adapted for assessing the quality of single supervision sessions (e.g. 'Now it is easier for me to understand my patient(s)', based on the Supervisory Working Alliance Inventory (SWAI); Efstation *et al.*, 1990).

In addition, the third author (A.S.) conducted six independent expert interviews. Three supervisees (women, age: $M = 30$, $SD = 2.2$, two CBT, one psychodynamic; all treating

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

out-patients) and three supervisors (one woman, two men, age: $M = 46$, $SD = 7.7$; supervision experience: $M = 9$ years, all CBT) were asked to describe details of an individual supervision session that they perceived as particularly effective or ineffective. All interviews were audio-recorded, transcribed, and categorized. The content analysis yielded four main themes of effective supervision sessions (productive structure, positive relationship, focus on specific situations rather than global topics, and reflection on and connection between theory and practice), which were used to create additional items (e.g. 'We linked theoretical knowledge to therapeutic practice'). A first pool of 81 items was developed. All co-authors provided feedback regarding content validity, redundancy, and formulations, and the pool was reduced to 38 items.

Measures

Measures for validity analyses

Successful supervision. The questionnaire to evaluate supervision (STEP-SV; Zarbock *et al.*, 2009) assesses successful supervision with three subscales on a rating scale from 1 (not at all true) to 7 (totally true): Clarifying (5 items), Relationship (3 items), and Problem Coping (4 items). The internal consistencies were $\alpha = .94$ (total scale), $\alpha = .88$ (clarifying), $\alpha = .83$ (problem solving) and $\alpha = .85$ (relationships) for the supervisees, and $\alpha = .87$ (total scale), $\alpha = .78$ (clarifying), $\alpha = .76$ (problem solving) and $\alpha = .60$ (relationships) for the supervisors.

Supervision alliance. We translated the Supervisory Working Alliance Inventory (SWAI; Efstation *et al.*, 1990) into German using forward and back translation through a native English speaker. It is one of the most widely used instruments for the assessment of supervisory relationships and consists of 19 items for supervisees (e.g. 'My supervisor makes the effort to understand me') and 23 items for supervisors (e.g. 'I help my trainee stay on track during our meetings'), and it uses a rating scale from 1 (almost never) to 7 (almost always). The internal consistencies were $\alpha = .93$ (supervisees) and $\alpha = .72$ (supervisors).

Measures of person and context variables

Fear of negative evaluation. The German version of the revised Brief Fear of Negative Evaluation Scale (Reichenberger *et al.*, 2016) assessed supervisees' anxiety of being negatively evaluated by others in social situations with 12 items (e.g. 'I worry about what other people will think of me even when I know it doesn't make any difference'; Cronbach $\alpha = .94$) and a rating scale from 1 (not at all characteristic for me) to 5 (absolutely characteristic for me).

Counselling self-efficacy. Supervisee self-efficacy was measured with two subscales of the German Counselor Activity Self-Efficacy Scales (CASES-R; Hahn *et al.*, 2021): Confidence in Applying Basic Therapy Skills (15 items) and Managing a Session (9 items). The items used a 10-point Likert scale (0 = 'no confidence' to 9 = 'complete confidence'; Cronbach's $\alpha = .89$).

Supervision techniques. We asked the participants to what extent 11 methods (Maiwald *et al.*, 2019; Weck *et al.*, 2017) were applied in their last supervision session (e.g case discussion, role play). They described the extent on a rating scale from 1 (not at all) to 4 (very much).

Competence feedback and adherence to feedback rules. We asked participants if they received explicit feedback on their therapeutic competences (yes/no). There are general guidelines for providing feedback to learners to increase its effectiveness. For example, feedback should be behaviour-based and concrete, provided as soon after the exercise as possible, and given with permission from the learner (Freeman, 1985; e.g. Hattie and Timperley, 2007; Heckman-Stone, 2004). Thus, we also assessed the extent to which those typical 'feedback rules' were met (e.g. 'The feedback referred to a specific behaviour or task', 'I had the chance to decide whether I would like to receive feedback', 'The feedback was personally hurtful'; Cronbach's $\alpha = .82$ and

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

$\alpha = .67$ for supervisees and supervisors, respectively). The rating scale ranged from 1 (do not agree at all) to 5 (agree completely).

Data analysis

The analyses were performed using RStudio 1.1.456 (RStudio Team, 2015).

Handling of missing data

The data included missing values (supervisees: 0.57%, supervisors: 0.65%). Visual inspection of the response pattern indicated that missing values most likely arose because participants dropped out of the survey and not for any other reason related to the observed data. Thus, the percentage of missing values in the QSQ items (supervisees: 8.78%; supervisors: 5.63%) was less than in the following questionnaires (supervisees: 32.20% for CASES-R; supervisors: 25.27% for SWAI), and Little's test per questionnaire indicated that most data were missing at random ($p \geq .132$; see ESM2 in the Supplementary material for details). Thus, we conducted multiple imputation by chained equations (MICE; imputation was conducted per questionnaire in combination with some demographic variables with 20 iterations). All results from the following analyses (i.e. descriptive analyses, EFA, group comparisons, correlations) were pooled across all 20 datasets according to Rubin's rules (Heymans and Eekhout, 2019), including Fishers' Z transformation of correlations. In the measurement invariance, missing values were addressed using full-information maximum likelihood estimation (FIML), as it produces results comparable to those of multiple imputation (e.g. Liu and Sriutaisuk, 2021) and is a straightforward method built into R as a standard procedure.

Exploratory factor analysis and item selection

The EFA was performed on the supervisee data, and the obtained solution was applied to the supervisor data as part of the measurement invariance analysis (see below). Consistent with our definition of supervision quality (see 'Item generation' section above), we included the entire item pool in the EFA, including items on satisfaction, relationship, and effectiveness. With the help of EFA, we wanted to gain information about the differentiability between these aspects. Pre-requisites for the EFA were tested with Bartlett's test and Shapiro's test of multivariate normality. To determine the number of factors, we combined parallel analysis and the empirical Kaiser criterion. In line with current recommendations in the literature (Goretzko *et al.*, 2021), we calculated and compared multiple EFAs with different rotation methods (i.e. promax *vs* oblimin as oblique rotations), different estimations depending on the data distribution (i.e. maximum likelihood (ML) estimation for normal distributions or weighted least squares (WLS) estimation with polychoric correlations for non-normal data), and different factor solutions (e.g. three *vs* four factors). In addition, we added bootstrapping with 2000 iterations. The best fitting factor solution was chosen in accordance with the following factor retention criteria: (1) at least four items per factor (Goretzko *et al.*, 2021), (2) high explained variance, (3) sufficient content validity, (4) good item communalities ($h^2 > .60$), and (5) good item statistics, preferring items with moderate item difficulties, normal distribution, good item discrimination ($r > .30$), and at least moderate correlations with all other items ($r > .30$). Finally, before the factor structure was further examined using measurement invariance (see next section), we conducted an additional EFA on the combined dataset (supervisees and supervisors) with the final 12 items to ensure that the factor solution did not change with fewer items.

Measurement invariance

Before comparing and interpreting the mean differences between supervisees and supervisors, we tested for measurement invariance (MI), which assesses the equivalence of a construct (i.e. interpretation of items) across groups (Byrne and Watkins, 2003; Putnick and Bornstein, 2016).

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

Using multi-group confirmatory analysis (lavaan; Rosseel, 2012), we tested for configural (equivalence of factorial structure), metric (equivalence of factor loadings), scalar (equivalence of item intercepts), and strict (equivalence of item residuals) MIs. Configural model fit was regarded as acceptable when the root mean square error of approximation was RMSEA $< .08$ and the standardized root mean square residual was SRMR $< .08$, and the comparative fit index was CFI $> .90$ (Byrne and Watkins, 2003). Three additional models with increasing restrictions (metric, scalar and strict) were specified and compared with each other. A significant worsening of model fit (i.e. non-invariance) was indicated by a ΔCFI $\geq -.005$, ΔRMSEA $\geq .010$, and ΔSRMR $\geq .025$ (metric MI) or ΔSRMR $\geq .005$ (scalar MI; Chen, 2007). These cut-off criteria are suggested when sample sizes are unequal.

Reliability and validity

Internal consistency was determined using Cronbach's alpha. We determined the convergent validity by calculating the correlations between the QSQ and the STEP-SV or SWAI. We expected the total scale of the QSQ to have a high correlation with the STEP-SV and the SWAI (with $r \geq .50$).

Associations with person and context variables

To examine *person variables*, we compared all the outcome scores of supervisees and supervisors (i.e. the means of the QSQ and its subscales, the STEP-SV and its subscales, the SWAI, and the adherence to feedback rules). To this end, we used a series of independent *t*-tests with Bonferroni–Holm adjusted significance levels for 10 tests and calculated Cohen's *d* as an effect size measure. In addition, we calculated the correlations between the QSQ and supervisees' fear of negative evaluation and their self-efficacy.

To examine *context variables* in both datasets (i.e. supervisee and supervisor), we compared differences in the QSQ for the following variables: (a) supervision mode (individual *vs* group supervision), (b) theoretical orientation (CBT *vs* PDT), and (c) competence feedback (feedback *vs* no feedback). To this end, we used a series of independent *t*-tests for each variable, applying Bonferroni–Holm adjusted significance levels for four tests (total scale and three subscales) each and calculating Cohen's *d* as an effect size measure. Furthermore, we calculated correlations in each dataset (i.e. supervisee and supervisor) between the QSQ and the variables: (a) techniques in supervision and (b) adherence to feedback rules.

Results

Item statistics

Overall, item difficulty was rather high (most items $P \geq .75$ and $M \geq 3.73$; Table 2), and item discrimination was good ($r_{i(t\text{-}i)} \geq .38$; Table 2). Based on inter-item correlations, we excluded two items (24, 27) with low associations with all other items ($r < .30$) from all further analyses because they targeted quite situation-specific aspects that did not apply to all supervision sessions (i.e. feedback based on video recordings, discussion of therapy documents).

Exploratory factor analysis and item selection

Pre-requisites

Shapiro's test of normality was significant ($p < .001$); thus, WLS estimation with polychoric correlations was used in the EFA. Bartlett's test of sphericity was significant (combined χ^2, $F(630, 302.59) = 18.086$, $p < .001$), indicating that the data were suitable for EFA. Parallel analyses and the empirical Kaiser criterion suggested 5 and 3 factors, respectively.

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

Table 2. Item statistics and results of exploratory factor analysis with three Factors and oblimin rotation (imputed data set with $n = 374$ supervisees

		EFA				Item statistics			
No.	Item	F1	F2	F3	h^2	*M*	*SD*	*P*	$r_{i(t-i)}$
6	***The session was effective***	***0.92***	−0.16	0.03	***0.73***	***4.04***	***1.04***	***.81***	***.78***
5	The session was helpful	0.88	0.02	−0.03	0.77	4.23	0.98	.85	.82
7	The session was a waste of time *(reversed coded)*	0.84	−0.08	0.02	0.65	4.39	1.02	.88	.75
1	I am satisfied with the session	0.82	0.18	−0.11	0.77	4.20	0.91	.84	.82
3	***My expectations on the session were met***	***0.82***	0.00	−0.02	***0.64***	***4.12***	***0.94***	***.82***	***.76***
38	All my concerns were addressed satisfactorily	0.77	0.12	0.05	0.78	4.11	1.00	.82	.85
29	We have made progress in terms of content	0.69	0.09	0.10	0.67	4.10	0.94	.82	.80
19	I received feedback on the aspects that were important for me	0.68	0.01	0.15	0.62	4.11	0.91	.82	.76
4	We worked focused	0.67	−0.03	0.05	0.48	4.17	0.93	.83	.67
32	***I benefited from the therapeutic exchange***	***0.67***	0.22	0.03	***0.72***	***4.29***	***0.98***	***.86***	***.81***
26	We talked too much about aspects that were unimportant *(reversed coded)*	0.65	−0.15	0.12	0.42	4.09	1.10	.82	.62
2	I left the session with a good feeling	0.63	0.32	−0.08	0.67	4.17	1.00	.83	.76
28	We have worked on all the topics that we had planned to work on	0.57	−0.06	−0.03	0.26	3.96	1.14	.79	.50
25	We had enough time to address my concerns	0.56	−0.08	0.02	0.27	3.92	1.13	.78	.51
30	***I perceived a learning progress***	***0.54***	0.12	0.20	***0.58***	***3.75***	***1.07***	***.75***	***.76***
18	I received useful feedback	0.45	0.25	0.26	0.67	4.04	1.07	.81	.82
16	I was supported in dealing with my current difficulties in therapy	0.42	0.22	0.25	0.58	4.02	1.02	.80	.76
14	Now it is easier for me to understand my patient(s)	0.42	0.17	0.13	0.40	3.73	1.09	.75	.64
17	I was supported in reflecting on the therapeutic approach	0.42	0.19	0.25	0.54	4.02	1.01	.80	.74
31	We followed a similar therapeutic approach	0.42	0.33	0.07	0.51	4.14	0.94	.83	.69
13	We linked theoretical knowledge with therapeutic practice	0.39	−0.11	0.35	0.38	3.52	1.24	.70	.59
15	Everybody could picture the patient(s) and his or her uniqueness	0.34	0.17	0.14	0.32	3.95	0.96	.79	.59
34	***I had the feeling that mistakes were allowed***	0.02	***0.80***	−0.08	***0.62***	***4.30***	***0.95***	***.86***	***.53***
36	***The way we worked, we were even able to resolve dissent***	0.05	***0.78***	0.01	***0.67***	***3.98***	***1.07***	***.80***	***.64***
37	***We constructively handled situations that were embarrassing or uncomfortable***	0.01	***0.68***	0.10	***0.53***	***3.99***	***1.05***	***.80***	***.60***
35	***We were a good team***	0.29	***0.64***	−0.01	***0.72***	***4.14***	***1.02***	***.83***	***.74***
20	I received feedback what I did well	−0.11	0.56	0.31	0.44	3.43	1.32	.69	.59
22	The feedback was understandable for me	0.11	0.50	0.27	0.54	4.26	0.91	.85	.68
33	I understand why certain feedback was given to me	0.20	0.44	0.21	0.51	4.29	0.90	.86	.68
9	***I know exactly <u>how</u> to implement the recommendations in therapy***	0.07	0.02	***0.72***	***0.61***	***3.85***	***0.95***	***.77***	***.65***
10	***I received concrete ideas (e.g. phrases, implementation of interventions)***	0.07	0.01	***0.69***	***0.55***	***3.65***	***1.26***	***.73***	***.63***
8	***I know exactly <u>what</u> to do in the next therapy session***	0.08	−0.04	***0.65***	***0.47***	***3.93***	***0.94***	***.79***	***.56***
21	I received feedback what I can improve	−0.13	0.12	0.61	0.35	3.52	1.13	.70	.48
12	***I'm able to realize the recommendations***	0.18	0.08	***0.52***	***0.48***	***4.09***	***0.88***	***.82***	***.64***
11	It was explained to me how other people would act in my therapy situation	−0.06	−0.03	0.52	0.23	2.90	1.31	.58	.38
23	The things we talked about remained vague and unspecific *(reversed coded)*	0.36	−0.05	0.43	0.48	4.02	1.12	.80	.66
R^2		.55	.23	.22					
Correlations									
F1	Effectiveness	—							
F2	Relationship	.62	—						
F3	Procedural Knowledge	.64	.38	—					

F1, Factor 1 (Effectiveness); F2, Factor 2 (Relationship); F3, Factor 3 (Procedural Knowledge); h^2, communality; *P*, item difficulty; $r_{i(t-i)}$, item discrimination. Final items of QSQ are printed in bold italics.

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

EFA

We tested 3- to 5-factor solutions (see Table 2 for the 3-factor solution and ESM3 and 4 (see Supplementary material) for the 4- and 5-factor solutions). The solutions explained 55% (three factors), 58% (four factors) and 60% (five factors) of the variance. All solutions included three factors that targeted topics of effectiveness (Factor 1), knowledge (Factor 2), and relationship (Factor 3). However, the 4- and 5-factor solutions contained additional factors that were difficult to interpret. The fourth factor in both solutions was broadly related to time management but contained only three items, including one item with a low factor loading and a low item communality (Item 15, $\lambda = 0.30$, $h^2 = 0.35$). The fifth factor referred to receiving feedback but included only one item with only moderate factor loading and item communality (Item 21, $\lambda = 0.57$, $h^2 = 0.49$). The 4- to 5-factor solutions therefore did not meet our factor retention criteria, and we chose the 3-factor solution (Table 2). We labelled the three factors Effectiveness, Procedural Knowledge, and Relationship.

Item selection

In the next step, we eliminated additional items from the questionnaire. In selecting the items, we were guided by our factor retention criteria and selected the items based on the factor loadings, thereby ensuring that they sharpened the content validity of each factor and displayed good item statistics. In line with our study objectives, we also focused on effectiveness rather than satisfaction with supervision. For example, we excluded Item 7 ('The session was a waste of time'), which was strongly skewed ($M = 4.39$, $P = .88$) and did not add much information to the factor Effectiveness in comparison with, for example, Item 6 ('The session was effective'), which was somewhat less skewed but represented the construct better and yielded comparable statistics ($\lambda = 0.92$, $h^2 = .73$, $P = .81$, $r_{i(t\text{-}i)} = .78$). Furthermore, Item 30 ('I perceived a learning progress') displayed only a moderate factor loading on Effectiveness and communality ($\lambda = 0.54$, $h^2 = .58$). However, we included this item because its difficulty and discrimination were good ($P = .75$, $r_{i(t\text{-}i)} = .76$), and we considered it an essential part of supervision effectiveness (Kühne *et al.*, 2019). Finally, care was taken to ensure that at least four items with good item statistics formed one factor.

The final questionnaire contained three factors with four items each (Table 2). Factor 1, Effectiveness, assesses the supervisees' general perceptions that the supervision session met their expectations (Item 3), was effective (Item 6), was associated with learning progress (Item 30), and that they benefited from professional exchange (Item 32). Factor 2, Procedural Knowledge, measures the supervisees' knowledge of *what* to do in the next therapy session (Item 8) and *how* to do it (Item 9) and the impression that supervisory advice was specific (Item 10) and doable (Item 12). Factor 3, Relationship, represents the emotional bond and cooperation between supervisees and supervisors, including the feeling that errors are permitted (Item 34), the feeling of being a good team (Item 35), constructive handling of disagreements (Item 36), and handling of embarrassing situations (Item 37).

Before the factor structure was further examined using measurement invariance (see next paragraph), we conducted an additional EFA on the combined dataset (supervisees and supervisors) with the final 12 items to ensure that the factor solution did not change with fewer items (see ESM5 and 6 of the Supplementary material for detailed results). The corresponding parallel analyses and the empirical Kaiser criterion suggested three and two factors, respectively. While the 3-factor solution (ESM5 in the Supplementary material) matched the original solution, the 2-factor solution (ESM6 in the Supplementary material) explained less variance (56 *vs* 63% in the 3-factor solution), had more items with low communalities (eight *vs* five items), and did not sufficiently separate effectiveness and relationship. For this reason, we chose the 3-factor QSQ as described above. The item statistics for the final QSQ for both the supervisee and supervisor versions are displayed in ESM7 of the Supplementary material.

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

Table 3. Convergent validity (supervisees ($n = 374$) are in lower half, and supervisors ($n = 136$) are in upper half)

	QSQ	QSQ-E	QSQ-P	QSQ-R	STEP	STEP-C	STEP-P	STEP-R	SWAI
QSQ	—	.80	.84	.64	.51	.40	.47	.45	.43
QSQ-E	.89	—	.50	.44	.50	.45	.43	.40	.34
QSQ-P	.80	.62	—	.25	.32	.21	.37	.27	.37
QSQ-R	.81	.62	.41	—	.37	.33	.27	.41	.27
STEP	.68	.65	.50	.55	—	.92	.88	.78	.40
STEP-C	.59	.57	.40	.49	.94	—	.66	.66	.27
STEP-P	.65	.62	.51	.48	.94	.82	—	.54	.40
STEP-R	.68	.63	.49	.58	.87	.71	.78	—	.44
SWAI	.58	.49	.36	.59	.55	.46	.48	.60	—

All correlations are significant at least at $p < 0.05$. QSQ, Quality of Supervision Questionnaire; QSQ-E, Effectiveness (QSQ); QSQ-P, Procedural Knowledge (QSQ); QSQ-R, Relationship (QSQ); STEP, Questionnaire to Evaluate Supervision (STEP-SV); STEP-C, clarifying (STEP-SV); STEP-P, problem coping (STEP-SV); STEP-R, relationship (STEP-SV); SWAI, Supervisory Working Alliance Inventory.

Measurement invariance

The 12-item QSQ was tested for MI between the supervisee and supervisor datasets. The basic model without any restrictions showed a good model fit, CFI = .955, RMSEA = .074, SRMR = .054, $\chi^2(102) = 218.76$, $p < .001$. Both configural MI and metric MI were reached (configural MI: ΔCFI < .001, ΔRMSEA < .001, ΔSRMR < .001; metric MI: ΔCFI = –.003, ΔRMSEA = .001, ΔSRMR = .009). However, the scalar model showed worse model fit (ΔCFI = –.010, ΔRMSEA = .004, and ΔSRMR = .003). To analyse problematic items, we followed an exploratory approach by sequentially releasing the intercept constraints of all items (Putnick and Bornstein, 2016). Partial scalar MI was reached by releasing the intercept of Item 10 (proposing concrete ideas), with ΔCFI = –.004, ΔRMSEA = .00, and ΔSRMR = .003, indicating that the QSQ was partially equivalent in terms of the interpretation of item means in both groups (see ESM6 of the Supplementary material for detailed results).

Reliability and validity

The internal consistency of the *supervisee version* of the QSQ was satisfactory, with $\alpha = .91$ (total scale), $\alpha = .88$ (Effectiveness), $\alpha = .83$ (Procedural Knowledge), and $\alpha = .87$ (Relationship). As Table 3 shows, the convergent validity was satisfactory, with strong correlations between the QSQ and the STEP-SV ($r = .68$, $p < .001$) and between the QSQ and the SWAI ($r = .58$, $p < .001$). Notably, the correlations between the STEP-SV and the SWAI were similarly high ($r = .47$ to .61, $p < .001$).

The internal consistencies for the *supervisor version* of the QSQ were satisfactory for the total scale ($\alpha = .81$) and Procedural Knowledge ($\alpha = .83$) but weaker for Effectiveness ($\alpha = .66$) and Relationship ($\alpha = .60$). The convergent validity was satisfactory for the total scale for the STEP-SV ($r = .51$, $p < .001$) but lower than expected for the SWAI ($r = .43$, $p < .001$). Overall, the correlational patterns were comparable between supervisees and supervisors but lower in absolute values for supervisors.

Associations with person variables

Figure 1 shows the results of the independent *t*-tests for the differences between supervisees and supervisors. The results showed that supervisors evaluated supervision sessions generally better than supervisees, with small to moderate effect sizes ($d = 0.33$ to 0.58; QSQ total and subscales). The effect sizes for the differences between supervisees and supervisors regarding the STEP-SV and the SWAI were comparable to the results for the QSQ ($d = 0.21$ to 0.54). In addition, supervisors were more likely than supervisees to report that the feedback rules were largely

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

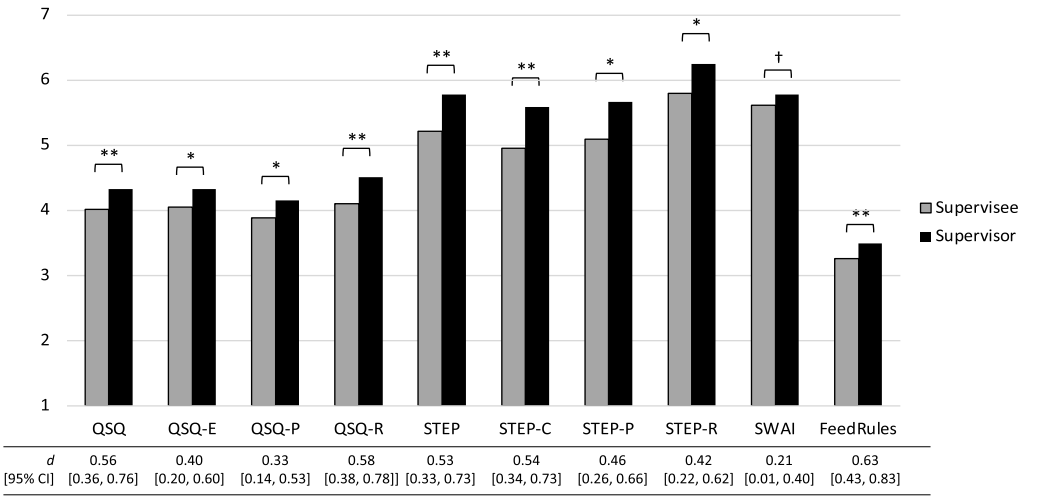

Figure 1. Differences between supervisees and supervisors. All effects are significant (Bonferroni-Holm adjusted significance levels, $^{**}p<.01$, $^{*}p<.05$, $\dagger p=.05$). QSQ, Quality of Supervision Questionnaire; QSQ-E, effectiveness (QSQ); QSQ-P, procedural knowledge (QSQ); QSQ-R, relationship (QSQ); STEP, Questionnaire to Evaluate Supervision (STEP-SV); STEP-C, clarifying (STEP-SV); STEP-P, problem coping (STEP-SV); STEP-R, relationship (STEP-SV); SWAI, Supervisory Working Alliance Inventory; FeedRules, adherence to feedback rules when providing competence feedback; *d*, Cohen's *d*.

adhered to when providing competency feedback ($d=.63$). Furthermore, supervisee self-efficacy was positively but marginally correlated with the QSQ (QSQ: $r=.16$, $p<.001$; Effectiveness: $r=.09$, $p<.001$; Procedural Knowledge: $r=.18$, $p<.001$; Relationship: $r=.14$, $p<.001$). The supervisees' level of fear of negative evaluation was negatively but only marginally associated with the QSQ (QSQ: $r=-.04$, $p<.001$; Effectiveness: $r=-.02$, $p=.051$, Procedural Knowledge: $r=-.07$, $p<.001$; and Relationship: $r=-.01$, $p=.203$).

Associations with context variables

Table 4 shows the results of the independent *t* tests for (a) supervision mode, (b) theoretical orientation, and (c) competence feedback. There were no significant differences between individual and group supervision sessions, either from the perspectives of the supervisees or the supervisors, although two effect sizes indicated a slight tendency for supervisees to regard individual sessions as more effective (QSQ-total scale, $d=-0.22$; QSQ-Effectiveness, $d=-0.29$). Furthermore, although theoretical orientation (CBT *vs* PDT) did not have a significant impact on supervisees' supervision quality, compared with PDT supervisors, CBT supervisors reported that Procedural Knowledge was much more prominent in their sessions ($d=-.86$). Finally, the supervisees found those sessions that included explicit competence feedback to be better (compared with sessions without competence feedback; all QSQ subscales, $d=0.43$ to 0.53). Similarly, the supervisors reported that more procedural knowledge was achieved if competence feedback was explicitly provided (*vs* no competence feedback; $d=-0.68$).

Table 5 shows the correlations between the QSQ and supervision techniques. Almost all associations were significant, although the effects were mostly marginal or small. Moderate correlations were found only for supervisors between personal goal setting and the QSQ total scale ($r=.33$, $p<.001$) or Procedural Knowledge ($r=.39$, $p<.001$). In addition, there were differential effects between supervisees and supervisors. For example, while case discussion was generally much more important for supervisees than for supervisors (e.g. $r=.29$ *vs* .02 for the QSQ total score), personal goal setting, in particular, was more strongly associated with session quality for supervisors than for supervisees (e.g. $r=.33$ *vs* .22 for the QSQ total score).

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

Table 4. Differences in QSQ depending on mode, theoretical orientation, and competence feedback

	Supervisees				Supervisors			
	M (*SD*)	*M* (*SD*)	*p*	*d* [95% CI]	*M* (*SD*)	*M* (*SD*)	*p*	*d* [95% CI]
				Mode				
	Individual ($n = 214$)	Group ($n = 160$)			Individual ($n = 105$)	Group ($n = 31$)		
QSQ	4.08 (0.76)	3.91 (0.82)	.053	−0.22 [−0.42, −0.01]	4.37 (0.45)	4.21 (0.49)	.169	−0.32 [−0.72, 0.08]
QSQ-E	4.17 (0.90)	3.89 (1.04)	.013	−0.29 [−0.49, −0.08]	4.36 (0.49)	4.24 (0.51)	.327	−0.22 [−0.62, 0.18]
QSQ-P	3.96 (0.88)	3.77 (0.99)	.078	-0.20 [−0.41, 0.01]	4.17 (0.80)	4.09 (0.82)	.703	-0.09 [−0.49, 0.31]
QSQ-R	4.12 (0.95)	4.08 (0.97)	.646	−0.05 [−0.26, 0.15]	4.58 (0.45)	4.30 (0.53)	.020	−0.56 [−0.96, −0.15]
				Theoretical orientation				
	CBT ($n = 224$)	PDT ($n = 118$)			CBT ($n = 70$)	PDT ($n = 28$)		
QSQ	4.03 (0.75)	3.99 (0.81)	.668	−0.05 [−0.27, 0.17]	4.47 (0.41)	4.27 (0.43)	.038	−0.48 [−0.92, −0.04]
QSQ-E	4.03 (0.92)	4.12 (0.94)	.444	0.09 [−0.13, 0.32]	4.37 (0.52)	4.36 (0.43)	.931	−0.02 [−0.46, 0.42]
QSQ-P	3.92 (0.88)	3.80 (0.88)	.256	−0.14 [−0.36, 0.09]	**4.49 (0.54)**	**3.92 (0.77)**	**<.001**	**−0.86 [−1.32, −0.41]**
QSQ-R	4.13 (0.85)	4.04 (1.04)	.426	−0.09 [−0.32, 0.13]	4.55 (0.46)	4.52 (0.45)	.811	−0.05 [−0.49, 0.39]
				Competence feedback				
	CF ($n = 117$)	noCF ($n = 257$)			CF ($n = 17$)	noCF ($n = 119$)		
QSQ	**4.14 (0.77)**	**3.73 (0.77)**	**<.001**	**−0.53 [−0.75, −0.31]**	4.37 (0.43)	4.07 (0.46)	.018	−0.67 [−1.18, −0.15]
QSQ-E	**4.17 (0.90)**	**3.77 (0.95)**	**<.001**	**−0.44 [−0.66, −0.21]**	4.35 (0.49)	4.23 (0.47)	.394	−0.25 [−0.76, 0.26]
QSQ-P	**4.02 (0.90)**	**3.58 (0.96)**	**<.001**	**−0.47 [−0.69, −0.25]**	**4.22 (0.70)**	**3.65 (0.95)**	**.009**	**−0.68 [−1.19, −0.16]**
QSQ-R	**4.23 (0.92)**	**3.83 (0.94)**	**<.001**	**−0.43 [−0.65, −0.21]**	4.54 (0.47)	4.34 (0.48)	.156	−0.42 [−0.93, 0.09]

Significant effects are in bold (Bonferroni–Holm adjusted significance levels). QSQ, Quality of Supervision Questionnaire; QSQ-E, Effectiveness (QSQ); QSQ-P, Procedural Knowledge (QSQ); QSQ-R, Relationship (QSQ); CBT, cognitive behaviour therapy; PDT, psychodynamic therapy and psychoanalysis; CF, competence feedback; noCF, no competence feedback.

Table 5. Correlations between QSQ and techniques used within the supervision session

	Supervisees ($n = 374$)						Supervisors ($n = 136$)					
	QSQ	QSQ-E	QSQ-P	QSQ-R	Occurrence %	Extent of use *M* (*SD*)[a]	QSQ	QSQ-E	QSQ-P	QSQ-R	Occurrence %	Extent of use *M* (*SD*)[a]
Case discussion	.29	.24	.24	.24	97.5%	3.40 (0.69)	.02	.13	−.03	−.02	99.0%	3.63 (0.59)
Information brokering	.24	.22	.23	.16	94.3%	2.77 (0.64)	.25	.15	.29	.08	97.0%	2.89 (0.65)
Recommending literature	.09	.08	.08	.08	43.2%	2.35 (0.57)	.17	.13	.15	.09	63.0%	2.36 (0.51)
Self-reflection	.20	.22	.08	.19	62.8%	2.45 (0.71)	.16	.19	.01	.27	87.4%	2.59 (0.73)
Videotapes	.07	.08	.06	.04	9.8%	3.04 (0.80)	.09	.04	.15	−.04	18.2%	2.74 (0.64)
Audiotapes	.08	.06	.08	.07	7.0%	3.01 (0.87)	.03	.02	.12	−.15	10.3%	2.65 (0.49)
Agenda	.15	.12	.14	.12	43.6%	2.53 (0.65)	.19	.01[b]	.32	.02	61.3%	2.75 (0.73)
Personal goal setting	.22	.20	.14	.19	71.1%	2.85 (0.73)	.33	.22	.39	.07	91.6%	3.02 (0.72)
Role play	.15	.15	.12	.12	11.8%	2.50 (0.69)	.23	.17	.24	.09	30.3%	2.56 (0.69)
Therapist homework tasks	.17	.17	.13	.13	25.9%	2.48 (0.68)	.17	.10	.23	−.00[b]	57.2%	2.59 (0.69)
Using supervision protocol	.10	.06	.10	.08	55.9%	2.99 (0.81)	.02	.03	−.05	.10	62.4%	3.01 (0.90)

All correlations are significant at least at $p<0.05$. QSQ. Quality of Supervision Questionnaire; QSQ-E, Effectiveness (QSQ); QSQ-P, Procedural Knowledge (QSQ); QSQ-R, Relationship (QSQ). [a]Values refer to those subjects who indicated that this method did occur in session (2 = 'little', 4 = 'very much'); [b]not significant with $p>0.05$.

In addition, information brokering and self-reflection had one of the strongest correlations with the QSQ for both supervisees and supervisors, whereby self-reflection was associated primarily with Relationship for supervisors ($r = .27$, $p < .001$). These four techniques (case discussion, information brokering, personal goal setting, and self-reflection) were also used most frequently overall (i.e. in $\geq 71.1\%$ of sessions). However, the supervisors indicated that they had used any of the supervision techniques to a greater extent than supervisees (e.g. homework: 57.2 *vs* 25.9%). In addition, the following correlations show that adhering to feedback rules was generally associated with better sessions in both the supervisee dataset (QSQ: $r = .23$, $p < .001$; Effectiveness: $r = .20$, $p < .001$; Procedural Knowledge: $r = .18$, $p < .001$; Relationship: $r = .20$, $p < .001$) and the supervisor dataset (QSQ: $r = .41$, $p < .001$; Effectiveness: $r = .39$, $p < .001$; Procedural Knowledge: $r = .28$, $p < .001$; Relationship: $r = .30$, $p < .001$).

Discussion

The first objective of this study was to develop a questionnaire to measure the quality of individual supervision sessions with a particular focus on more clearly separating the constructs of supervision effectiveness, satisfaction, and relationship. The final 12-item Quality of Supervision Questionnaire (QSQ) consists of three factors (Effectiveness, Procedural Knowledge, and Relationship). The internal consistency and convergent validity from the supervisee perspective were largely satisfactory. However, the correlations between all supervision measures require some thoughts on the construct 'supervision quality' and its measurement. In addition, the mixed internal consistencies and low convergent validities in the supervisor dataset indicate that the reliability of the supervisor version is somewhat limited.

Measurement of supervision quality

The QSQ is a valid method for monitoring the progress of supervision from session to session in a time-efficient manner. In contrast to other questionnaires (Ladany *et al.*, 1996; Palomo *et al.*, 2010; Winstanley and White, 2011; Zarbock *et al.*, 2009) and in line with the requirement of considering aspects other than satisfaction (Kühne *et al.*, 2019; Milne, 2018), the QSQ subscales do not contain any general satisfaction items and are defined briefly and succinctly. This made it possible for the Relationship subscale to focus exclusively on the constructive collaboration between supervisee and supervisor and their mutual problem solving, as described in the definition of supervision (Falender, 2014). Aspects of skills evaluation or education are not included here, in contrast to other procedures (e.g. SWAI). Due to its good convergent validity, one advantage of the Relationship subscale is that it can capture the supervisory relationship in a time-efficient manner. By using it from session to session, it is also possible to detect and react to relationship ruptures promptly.

The other two QSQ subscales relate primarily to concrete learning outcomes (i.e. general learning progress and specific knowledge), which is also a central element of competency-based supervision (Falender, 2014). Falender (2014) notes, among other things, that the acquisition of knowledge and the promotion of self-efficacy are central components of effective supervision. The QSQ is aimed precisely at these aspects, particularly the Procedural Knowledge subscale, and assesses the *what* and *how* of the therapeutic approach as well as the confidence to be able to implement it. Against this background, it is not surprising that the convergent validity of Procedural Knowledge was partly lower than that of the other subscales. For example, it was moderately correlated with the alliance questionnaire SWAI ($r = .36$) and only weakly correlated with the QSQ Relationship subscale ($r = .25$). This underscores that a good relationship contributes to a good session, but it does not automatically lead to knowledge of what and how to do it in the future (Goodyear and Bernard, 1998). Thus, this scale seems to be a particularly

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

interesting aspect of session quality, as knowledge is often considered a pre-requisite for demonstrated skills (Heinze *et al.*, in press; Muse and McManus, 2013).

Although the QSQ has several advantages, we believe that each supervision instrument investigated in this study has added value. For example, the SWAI measures supervision quality in terms of 'working alliance', including collaboration, understanding, feedback, structure, goals, technique, and patient focus. For this reason, it correlates approximately equally with all other instruments and their subscales in this sample. Therefore, the SWAI seems to be well suited for the assessment of a general picture of supervision quality. However, for a differentiated analysis and use at the session level, it seems somewhat too global. The STEP-SV, on the other hand, makes it possible to capture quality in a more differentiated way. In doing so, it follows an established theory in which Relationship measures understanding by the supervisor, Clarification measures (self-) reflection and patient focus, and Problem Coping measures productivity in future sessions. We advocate that researchers clearly describe which definition of quality they wish to examine (i.e. a combination of several aspects *vs* a specific focus) and that they be aware of potential conceptual overlap when using scales that assess quality rather globally (e.g. working alliance) as well as psychometric properties. Researchers who wish to specifically evaluate single CBT supervision sessions and are primarily interested in variables other than satisfaction, for example, general productivity, process knowledge, and relationship, can rely on the QSQ as a valid and reliable self-report measure.

The supervisor version of the QSQ

Even if the Effectiveness and Relationship subscales had lower internal consistencies in the supervisor version, this does not automatically call into question the reliability of short scales, especially when short scales assess a broad construct (Ziegler *et al.*, 2014). However, considering the lower correlations with the STEP-SV, the overall conclusion about the supervisor version must be formulated somewhat cautiously. On the one hand, the results might be due to a smaller sample size than in the supervisee sample ($n = 136$) or unclear factor structure of the supervisors' versions of the STEP-SV (Zarbock *et al.*, 2009). On the other hand, the results of the MI test (i.e. partial scalar MI) indicated that supervisors interpreted some items differently than supervisees did. It should be borne in mind that scalar MI is almost never achieved in practice and that some authors are convinced that a violation of MI does not substantially reduce the validity of the scale and that group differences can still be calculated and interpreted (Robitzsch and Lüdtke, 2023). In our case, the MI results indicate that supervisors and supervisees might use other benchmarks and quality standards, perhaps due to different levels of expertise and opportunities for comparison. This idea also fits well with the finding that supervisors use somewhat different techniques to be more effective than supervisees (e.g. find case discussions less effective than supervisees). Other authors also report psychometric problems in their supervisor versions or using partly different items (Efstation *et al.*, 1990; Zarbock *et al.*, 2009). To our knowledge, no other instrument has been previously tested for measurement invariance. Consequently, future researchers should replicate the findings in larger samples and perhaps develop items exclusively from the supervisor perspective rather than simply adapting the wording of the same items to fit both perspectives.

Associations with person and context variables

Although this study focuses on a randomly selected supervision session, many of the results are consistent with previous findings on the entire supervision process. These include, for example, the relationships between quality and self-efficacy (Kühne *et al.*, 2019), fear of negative evaluation (e.g. Gray *et al.*, 2001), giving competence feedback (Lehrman-Waterman and Ladany, 2001; Maiwald *et al.*, 2019), and differences between supervisee and supervisor perspectives (Mathieson

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

et al., 2009; O'Donovan and Kavanagh, 2014; Zarbock *et al.*, 2009). However, the effect sizes for self-efficacy and fear of negative evaluations were only marginal and may not reflect meaningful relationships. To study the relationships in more detail, future studies should examine the situational component of these variables (i.e. state self-efficacy, state fear of negative evaluations) when using the QSQ rather than focusing on trait measures, as in the present study. Furthermore, we found that video analysis and role-playing are still rarely used in sessions, while case discussions make up a large proportion of sessions (Milne and Dunkerley, 2010; Weck *et al.*, 2017). In addition, there are some differential findings worth discussing.

Person variables

The extent of session quality depends largely on whether you are a supervisor or a supervisee. Supervisors tend to rate session quality and supervisory alliance as significantly better than supervisees. This finding is similar to research findings on psychotherapy, in which therapists tend to over-estimate their effectiveness relative to their patients' outcomes (Constantino *et al.*, 2023). At worst, this over-estimation can lead to negative experiences for supervisees if the supervisor continues potential harmful behaviour, such as ignoring the supervisees' cultural background or behaving cold and distant (Ellis *et al.*, 2014). However, considering that quality was rated as very high in both groups, this likely did not occur in our sample.

Context variables

It is promising that the supervision mode and theoretical orientation did not have a major impact on supervision quality. The only specific finding that CBT supervisors (*vs* PDT supervisors) perceived that their sessions promoted significantly more procedural knowledge is generally in line with PDT sessions being less behaviour- or goal-oriented (Weck *et al.*, 2017). Nonetheless, one should keep in mind that the sample consisted of only 21 supervisors for PDT, which limits the generalizability of the results. Although the relationship between the QSQ and supervision mode (i.e. individual *vs* group sessions) was not significant, it might be worth examining potential differences regarding expectations and professional exchange group supervision in future studies, given the small effects for QSQ Effectiveness (Nelson, 2014).

It also seems relatively irrelevant which technique is used to a particularly large extent within a session as all methods showed positive but only small correlations with the QSQ. This relatively small influence of one specific method is comparable to the finding in the psychotherapy literature that specific interventions/techniques alone explain only a small proportion of the variance in therapy outcomes (Wampold and Imel, 2015). However, technique and interpersonal factors (i.e. how the technique is used) are often intertwined (De Felice *et al.*, 2019). This idea fits well with the result that adherence to feedback rules when providing competence feedback was associated with more effective sessions (Freeman, 1985). Adherence to feedback rules included: praise, value-free communication that does not hurt supervisees personally, feedback that is based on behaviours, tasks, or the last therapy session, addressing changeable aspects and making suggestions for improvement; and the opportunity to decide whether competence feedback is welcomed. Thus, the present study once again emphasizes the role of competence feedback (Freeman, 1985; Heckman-Stone, 2004; Weck *et al.*, 2021), which was one of the strongest factors influencing the quality of supervision sessions, with medium effect sizes, even though only approximately one-third of supervisees in our sample received such feedback.

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

Implications for supervision practice

Many researchers recommend that supervisors monitor their supervision quality regularly (cf. Milne and Reiser, 2017). Against this background, the QSQ might be particularly important for quickly and inexpensively observing supervision quality and identifying (perhaps unexpectedly) unhelpful supervision sessions, especially from the perspective of the supervisees (see ESM7 of the Supplementary material for the final questionnaires). We believe that the QSQ Relationship subscale could replace longer supervisory alliance questionnaires if time resources are limited. Although the content of the scale differs somewhat from the typical definition of the supervisory relationship (emotional bond, agreement on tasks and goals), it correlates sufficiently strongly with the SWAI to serve as an indicator of a cooperative relationship. The results can also serve as a basis for discussion in supervision, on the basis of which mutual expectations can be aligned. However, we advise combining the QSQ with objective measures of supervisor competence (e.g. Reiser *et al.*, 2018), supervisee learning outcomes or competency, and patient outcomes. These multiple perspectives (supervisors, supervisees, and raters) enable a more realistic picture of the quality of supervision (Knox and Hill, 2021; Muse and McManus, 2013; Watkins *et al.*, 2021).

Limitations and future research

First, it should be mentioned that this study validated the German version of the QSQ. The English translation must therefore also be investigated in an English-speaking sample. Although the QSQ has overall good psychometric properties, its factor structure needs to be replicated with confirmatory factor analyses in further studies, preferably with a longitudinal design. It is particularly important that supervisors and supervisees evaluate the same sessions to obtain a differentiated picture of the supervision process. There is also a lack of information on discriminant validity. However, the low correlations of the QSQ with self-efficacy (CASES) already provide initial indications that the QSQ captures supervision-related aspects and is not merely a measure of supervisee self-efficacy. Nevertheless, future studies should explicitly examine discriminant validity. In addition, the skewness of the data is striking. However, most supervision questionnaires also have skewed data (Gonsalvez *et al.*, 2017; Zarbock *et al.*, 2009), which indicates high satisfaction with supervision. Moreover, it is quite possible that in longitudinal studies, there is greater variance in the QSQ across supervisors and sessions. Nonetheless, future revisions of the instrument should consider ways to measure 'quality' in a more differentiated way, e.g. by items that make a high degree of agreement more difficult (e.g. 'My expectations of the session are met 100%') or by using a 7-point rating scale. Furthermore, while the analyses for supervisees were adequately powered, the analyses for supervisors were partly under-powered, should be interpreted with caution, and should be replicated in larger samples. It should also be noted that numerous moderator variables can have an additional influence on the results, such as the duration of the supervision session (which, for example, showed a high variance in our sample) or the general level of competence of the trainees (and therefore their experience with supervision sessions). Future studies could specifically control for such influences. Finally, it must be mentioned that the authors are cognitive behavioural therapists, and despite efforts to include other supervision concepts (cf. expert interviews with PDT therapists), the QSQ may not be equally appropriate for all theoretical orientations. In addition, aspects of culture and diversity were not sufficiently considered in either item development or sample recruitment, limiting its generalizability to marginalized groups.

https://doi.org/10.1017/S1754470X24000321 Published online by Cambridge University Press

In conclusion, the QSQ, especially its supervisee version, is a reliable and valid instrument for examining the quality of an individual supervision session, is potentially suitable for longitudinal study designs, and is applicable to individual and group sessions.

Key practice points

(1) The QSQ measures the quality of a supervision session from the supervisee's and supervisor's perspectives with 12 items. It measures how effective a supervision session was (e.g. in terms of learning progress), how much it promoted the acquisition of procedural knowledge (e.g. knowing exactly what to do in the next therapy session), and how cooperative and constructive the relationship between supervisor and supervisee was.
(2) Supervisors and training institutes can use the QSQ to record the quality of single supervision sessions and thus monitor the supervision process. This is relevant because in the present sample, supervisors generally rated the quality of their sessions higher than supervisees.
(3) Higher quality supervision sessions are associated with receiving competence feedback and adherence to 'feedback rules', but only marginally with generally higher self-efficacy levels. There were no significant differences between individual or group sessions. Cognitive behaviour supervisors perceive greater procedural knowledge in their sessions than psychodynamic supervisors.
(4) The QSQ Relationship subscale could replace longer supervisory alliance questionnaires if time resources are limited.
(5) The reliability of the subscales from the supervisor perspective is currently still limited. Therefore, the overall scale should be interpreted with caution.

Further reading

Knox, S., & Hill, C. E. (2021). Training and supervision in psychotherapy: what we know and where we need to go. In *Bergin's and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 327–349). John Wiley & Sons.

Kühne, F., Maas, J., Wiesenthal, S., & Weck, F. (2019). Empirical research in clinical supervision: a systematic review and suggestions for future studies. *BMC Psychology*, *7*, article 54. https://doi.org/10.1186/s40359-019-0327-7

Milne, D. L. (2018). *Evidence-Based CBT Supervision: Principles and Practice*. John Wiley & Sons.

O'Donovan, A., & Kavanagh, D. J. (2014). Measuring competence in supervisees and supervisors. Satisfaction and related reactions in supervision. In C. E. Watkins Jr & D. L. Milne (eds), *The Wiley International Handbook* (pp. 458–467). John Wiley & Sons.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1754470X24000321

Data availability statement. We pre-registered the methods and statistical analyses on the Open Science Framework where the original data sets and scripts of analysis will be available: https://doi.org/10.17605/OSF.IO/BF3RE

Acknowledgements. None.

Author contributions. **Ulrike Maaß:** Conceptualization (lead), Data curation (lead), Formal analysis (lead), Funding acquisition (equal), Investigation (equal), Methodology (lead), Project administration (lead), Software (lead), Validation (lead), Visualization (lead), Writing – original draft (lead), Writing – review & editing (lead); **Franziska Kühne:** Conceptualization (supporting), Writing – original draft (supporting); **Alexandra Schöttler:** Conceptualization (supporting), Investigation (lead), Methodology (equal), Project administration (equal); Florian Weck: Conceptualization (supporting), Resources (lead), Supervision (lead).

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The authors declare none.

Ethical standards. This study was pre-registered at the Open Science Framework (28 May 2022), and the study protocol was approved by the ethics committee of the University of Potsdam (reference number 8/2022). The study conformed to the Declaration of Helsinki, and informed consent was obtained from all participants. All data, analysis codes, and research materials are available at: https://doi.org/10.17605/OSF.IO/BF3RE. Deviations from the pre-registration are described in the electronic Supplementary material online (ESM1).

References

Alfonsson, S., Parling, T., Spännargård, Å., Andersson, G., & Lundgren, T. (2018). The effects of clinical supervision on supervisees and patients in cognitive behavioral therapy: a systematic review. *Cognitive Behaviour Therapy*, *47*, 206–28. doi: 10.1080/16506073.2017.1369559

American Psychological Association (2014). *Guidelines for Clinical Supervision in Health Service Psychology*. http://apa.org/about/policy/guidelines-supervision.pdf

Beinart, H. (2014). Building and Sustaining the Supervisory Relationship. In *The Wiley International Handbook of Clinical Supervision*, ed. C. E. Watkins Jr and D. L. Milne, pp. 257–81. West Sussex: John Wiley & Sons, Ltd.

Binder, J. L. (1993). Is it time to improve psychotherapy training? *Clinical Psychology Review*, *13*, 301–318. doi: 10.1016/0272-7358(93)90015-E

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, *34*, 155–75. doi: 10.1177/0022022102250225

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464–504. doi: 10.1080/10705510701301834

Constantino, M. J., Boswell, J. F., Coyne, A. E., Muir, H. J., Gaines, A. N., & Kraus, D. R. (2023). Therapist perceptions of their own measurement-based, problem-specific effectiveness. *Journal of Consulting and Clinical Psychology*. doi: 10.1037/ccp0000813

De Felice, G., Giuliani, A., Halfon, S., Andreassi, S., Paoloni, G., & Orsucci, F. F. (2019). The misleading dodo bird verdict. how much of the outcome variance is explained by common and specific factors? *New Ideas in Psychology*, *54*, 50–55. doi: 10.1016/j.newideapsych.2019.01.006

Efstation, J. F., Patton, M. J., & Kardash, C. M. (1990). Measuring the Working Alliance in Counsellor Supervision. *Journal of Counseling Psychology*, *37*, 322–329. doi: 10.1037/0022-0167.37.3.322

Ellis, M. V., Berger, L., Hanus, A. E., Ayala, E. E., Swords, B. A., & Siembor, M. (2014). Inadequate and harmful clinical supervision: testing a revised framework and assessing occurrence. *The Counseling Psychologist*, *42*, 434–472. doi: 10.1177/0011000013508656

Falender, C. A. (2014). Clinical supervision in a competency-based era. *South African Journal of Psychology*, *44*, 6–17. doi: 10.1177/0081246313516260

Freeman, E. M. (1985). The importance of feedback in clinical supervision: implications for direct practice. *The Clinical Supervisor*, *3*, 5–26. doi: 10.1300/J001v03n01_02

Gonsalvez, C. J. (2021). A short scale to evaluate supervision and supervisor competence – the SE-SC8. *Clinical Psychology & Psychotherapy*, *28*, 452–61. doi: 10.1002/cpp.2510

Gonsalvez, C. J., Hamid, G., Savage, N. M., & Livni, D. (2017). The Supervision Evaluation and Supervisory Competence Scale: psychometric validation. *Australian Psychologist*, *52*, 94–103. doi: 10.1111/ap.12269

Goodyear, R. K., & Bernard, J. M. (1998). Clinical supervision: lessons from the literature. *Counselor Education and Supervision*, *38*, 6–22. doi: 10.1002/j.1556-6978.1998.tb00553.x

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*, 3510–21. doi: 10.1007/s12144-019-00300-2

Gray, L. A., Ladany, N., Walker, J. A., & Ancis, J. R. (2001). Psychotherapy trainees' experience of counterproductive events in supervision. *Journal of Counseling Psychology*, *48*, 371–83. doi: 10.1037/0022-0167.48.4.371

Hahn, D., Weck, F., Witthöft, M., & Kühne, F. (2021). Assessment of counseling self-efficacy: validation of the German Counselor Activity Self-Efficacy Scales-Revised. *Frontiers in Psychology*, *12*, 780088. doi: 10.3389/fpsyg.2021.780088

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. doi: 10.3102/003465430298487

Heckman-Stone, C. (2004). Trainee preferences for feedback and evaluation in clinical supervision. *The Clinical Supervisor*, *22*, 21–33. doi: 10.1300/J001v22n01_03

Hedegaard, A. E. (2020). The Supervisory alliance in group supervision. *British Journal of Psychotherapy*, *36*, 45–60. doi: 10.1111/bjp.12495

Heinze, P. E., Weck, F., Maaß, U., & Kühne, F. (in press). The relation between knowledge and skills assessments in psychotherapy training: secondary analysis of a randomized controlled trial. *Training and Education in Professional Psychology*.

Heymans, M. W., & Eekhout, I. (2019). Applied Missing Data Analysis with SPSS and (R)Studio. https://bookdown.org/mwheymans/bookmi/ (accessed 5 January 2023).

Jendrusina, A. A., & Martinez, J. H. (2019). Hello from the other side: student of color perspectives in supervision. *Training and Education in Professional Psychology*, *13*, 116–60. http://dx.doi.org/10.1037/tep0000255

Knox, S., & Hill, C. E. (2021). Training and supervision in psychotherapy: what we know and where we need to go. In *Bergin's and Garfield's Handbook of Psychotherapy and Behavior Change*, pp. 327–349. John Wiley & Sons, Inc.

Kühne, F., Maas, J., Wiesenthal, S., & Weck, F. (2019). Empirical research in clinical supervision: a systematic review and suggestions for future studies. *BMC Psychology*, *7*, article 54. doi: 10.1186/s40359-019-0327-7.

Ladany, N., Hill, C. E., Corbett, M. M., & Nutt, E. A. (1996). Nature, extent, and importance of what psychotherapy trainees do not disclose to their supervisors. *Journal of Counseling Psychology*, *43*, 10–24. doi: 10.1037/0022-0167.43.1.10

Lambert, M. J., & Ogles, B. M. (1997). The effectiveness of psychotherapy supervision. *Handbook of Psychotherapy Supervision*, pp. 421–446. John Wiley & Sons, Inc.

Lehrman-Waterman, D., & Ladany, N. (2001). Development and validation of the Evaluation Process Within Supervision Inventory. *Journal of Counseling Psychology*, *48*, 168–77. doi: 10.1037/0022-0167.48.2.168.

Liu, Y., & Sriutaisuk, S. (2021). A comparison of FIML- versus multiple-imputation-based methods to test measurement invariance with incomplete ordinal variables. *Structural Equation Modeling*, *28*, 590–608. doi: 10.1080/10705511.2021.1876520

Maiwald, L. M., Kühne, F., Junga, Y. M., Rudolph, D., Witthöft, M., Lüthke, L., Heid, E., & Weck, F. (2019). Erfolgreiche Supervision in der Psychotherapieausbildung: Eine explorativ-qualitative Untersuchung der Supervisor_innen- und Supervisand_innenperspektive. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *48*, 228–236. doi: 10.1026/1616-3443/a000563

Mathieson, F. M., Barnfield, T., & Beaumont, G. (2009). Are we as good as we think we are? self-assessment *versus* other forms of assessment of competence in psychotherapy. *the Cognitive Behaviour Therapist, 2*, 43–50. doi: 10.1017/S1754470X08000081

McMahon, A., & Errity, D. (2013). From new vistas to life lines: psychologists' satisfaction with supervision and confidence in supervising. *Clinical Psychology & Psychotherapy*, *21*, 264–75.

Milne, D. L. (2007). An empirical definition of clinical supervision. *British Journal of Clinical Psychology*, *46*, 437–47. doi: 10.1348/014466507X197415

Milne, D. L. (2018). *Evidence-Based CBT Supervision: Principles and Practice*. John Wiley & Sons.

Milne, D. L., & Dunkerley, C. (2010). Towards evidence-based clinical supervision: the development and evaluation of four CBT guidelines. *the Cognitive Behaviour Therapist*, *3*, 43–57. doi: 10.1017/S1754470X10000048

Milne, D. L., & James, I. (2000). A systematic review of effective cognitive-behavioural supervision. *British Journal of Clinical Psychology*, *39*, 111–27. doi: 10.1348/014466500163149.

Milne, D. L., & Reiser, R. P. (2017). *A Manual for Evidence-Based CBT Supervision*. John Wiley & Sons.

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review*, *33*, 484–99. doi: 10.1016/j.cpr.2013.01.010.

Nelson, M. L. (2014). Using the major formats of clinical supervision. In *The Wiley International Handbook of Clinical Supervision*, pp. 308–328. John Wiley & Sons.

O'Donovan, A., & Kavanagh, D. J. (2014). Measuring competence in supervisees and supervisors. satisfaction and related reactions in supervision. In *The Wiley International Handbook*, ed. C. E. Watkins Jr & D. L. Milne, pp. 458–467. John Wiley & Sons.

Ögren, M.-L., Boalt Boëthius, S., & Sundin, E. (2014). Challenges and possibilities in group supervision. In *The Wiley International Handbook of Clinical Supervision*, ed. C. E. Watkins Jr & D. L. Milne, pp. 648–669. John Wiley & Sons.

Palomo, M., Beinart, H., & Cooper, M. (2010). Development and Validation of the Supervisory Relationship Questionnaire (SRQ) in UK Trainee Clinical Psychologists. *British Journal of Clinical Psychology*, *49*:131–49. doi: 10.1348/014466509X441033.

Park, E. H., Ha, G., Lee, S., Lee, Y. Y., & Lee, S. M. (2019). Relationship between the supervisory working alliance and outcomes: a meta-analysis. *Journal of Counseling & Development*, *97*, 437–46. doi: 10.1002/jcad.12292

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. doi: 10.1016/j.dr.2016.06.004

Reese, R. J., Usher, E. L., Bowman, D. C., Norsworthy, L. A., Halstead, J. L., Rowlands, S. R., & Chisholm, R. R. (2009). Using client feedback in psychotherapy training: an analysis of its influence on supervision and counselor self-efficacy. *Training and Education in Professional Psychology*, *3*, 157–68. doi: 10.1037/a0015673

Reichenberger, J., Schwarz, M., König, D., Wilhelm, F. H., Voderholzer, U., Hillert, A., & Blechert, J. (2016). Angst vor negativer sozialer Bewertung: Übersetzung und Validierung der Furcht vor negativer Evaluation-Kurzskala (FNE-K). *Diagnostica*, *62*, 169–81. doi: 10.1026/0012-1924/a000148

Reiser, R. P., Cliffe, T., & Milne, D. L. (2018). An improved competence rating scale for CBT supervision: Short-SAGE. *the Cognitive Behaviour Therapist*, *11*, e7. doi: 10.1017/S1754470X18000065

Reiser, R. P., & Milne, D. L. (2016). A survey of CBT supervision in the UK: methods, satisfaction and training, as viewed by a selected sample of CBT supervision leaders. *the Cognitive Behaviour Therapist*, *9*, e20. doi: 10.1017/S1754470X15000689

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling*, *30*, 859–70. doi: 10.1080/10705511.2023.2191292

Rønnestad, M. H., Orlinsky, D. E., Schröder, T. A., Skovholt, T. M., & Willutzki, U. (2018). The Professional Development of Counsellors and Psychotherapists: Implications of Empirical Studies for Supervision, Training and Practice. *Counselling and Psychotherapy Research*, *19*, 214–30. doi: 10.1002/capr.12198

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. doi: https://doi.org/10.18637/jss.v048.i02.

RStudio Team (2015). RStudio: Integrated Development for R.

Wampold, B. E., & Imel, Z. E. (2015). *The Great Psychotherapy Debate. The Evidence for What Makes Psychotherapy Work* (2nd edn). New York: Routledge.

Watkins, C. E., Vîşcu, L.-I., & Cadariu, I.-E. (2021). Psychotherapy Supervision Research: On Roadblocks, Remedies, and Recommendations. *European Journal of Psychotherapy & Counselling*, *23*, 8–25. doi: 10.1080/13642537.2021.1881139

Weck, F., Junga, Y. M., Kliegl, R., Hahn, D., Brucker, K., & Witthöft, M. (2021). Effects of competence feedback on therapist competence and patient outcome: a randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *89*, 885–897. doi: 10.1037/ccp0000686

Weck, F., Kaufmann, Y. M., & Witthöft, M. (2017). Topics and techniques in clinical supervision in psychotherapy training. *the Cognitive Behaviour Therapist*, *10*, E3. doi: 10.1017/S1754470X17000046

Winstanley, J., & White, E. (2011). The MCSS-26©: revision of the Manchester Clinical Supervision Scale© using the Rasch measurement model. *Journal of Nursing Measurement*, *19*, 160–78. doi: 10.1891/1061-3749.19.3.160

Ybrandt, H., Sundin, E. C., & Capone, G. (2016). Trainee therapists' views on the alliance in psychotherapy and supervision: a longitudinal study. *British Journal of Guidance & Counselling*, *44*, 530–39. doi: 10.1080/03069885.2016.1153037

Zarbock, G., Drews, M., Bodansky, A., & Dahme, B. (2009). The evaluation of supervision: construction of brief questionnaires for the supervisor and the supervisee. *Psychotherapy Research*, *19*, 194–204. doi: 10.1080/10503300802688478

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*, 185–89. doi: 10.1027/1614-0001/a000148

Cite this article: Maaß U, Kühne F, Schöttler A, and Weck F (2024). The Quality of Supervision Questionnaire – associations between quality, person and context variables. *The Cognitive Behaviour Therapist*. https://doi.org/10.1017/S1754470X24000321