



## Clinical supervision based on video vs. verbal report: a randomized controlled trial

Florian Weck, Ulrike Maaß, Tatjana Paunov, Peter E. Heinze & Franziska Kühne

**To cite this article:** Florian Weck, Ulrike Maaß, Tatjana Paunov, Peter E. Heinze & Franziska Kühne (28 Nov 2024): Clinical supervision based on video vs. verbal report: a randomized controlled trial, *Cognitive Behaviour Therapy*, DOI: [10.1080/16506073.2024.2434016](https://doi.org/10.1080/16506073.2024.2434016)

**To link to this article:** <https://doi.org/10.1080/16506073.2024.2434016>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 28 Nov 2024.



[Submit your article to this journal](#)



Article views: 1590



[View related articles](#)








[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

# Clinical supervision based on video vs. verbal report: a randomized controlled trial

Florian Weck , Ulrike Maaß , Tatjana Paunov , Peter E. Heinze   
and Franziska Kühne 

Department of Clinical Psychology and Psychotherapy, University of Potsdam, Potsdam, Germany

## ABSTRACT

Clinical supervision is considered important in psychotherapy training, but little is known about the efficacy of specific supervision methods. We investigate two such methods (video-based vs. verbal report-based supervision) in a randomized controlled trial. Seventy-three supervisees were trained in common cognitive-behavioral therapy methods (i.e. behavioral activation and cognitive restructuring) by means of written information and a modelling video demonstrating the techniques. Supervisees had to apply the techniques in role plays with standardized patients (presenting depressive patients). Subsequently, supervisees were randomized to supervision, based on the video, or supervision based on the verbal report of the supervisees. Subsequently and after a three-month follow-up period, supervisees had to demonstrate the therapeutic techniques again. Therapist competence, therapeutic alliance, empathy, and anxiety were assessed through various different perspectives (i.e. independent raters, standardized patients, and supervisees' self-evaluation). Both supervision conditions lead to a significant improvement of therapeutic competences, therapeutic alliance, and empathy. No significant differences were found between the two supervision conditions. At three-month follow-up, training effects decreased on all perspectives except standardized patients. A training condition without supervision would be necessary to demonstrate that improvements are specific effects of the supervision conditions. Moreover, further supervision seems necessary to maintain training effects over time.

## ARTICLE HISTORY

Received 12 March 2024


Accepted 4 November 2024

## KEYWORDS

Alliance; empathy; clinical supervision; psychotherapy process; psychotherapy training; therapeutic competence

Clinical supervision is considered internationally as high-quality clinical practice in the field of mental health (Milne & Watkins, 2014). In particular, for psychotherapy training, clinical supervision is regarded as important for the development of therapeutic competences of trainees (e.g. Frank et al., 2020). Empirical research reveals a positive estimation of clinical supervision by supervisees, positive effects on therapeutic competence, and on therapy outcome (Henrich et al., 2023; Keum & Wang, 2021; Kühne et al., 2019; Rakovshik & McManus, 2010). However, the quality of supervision research suffers

**CONTACT** Florian Weck  [fweck@uni-potsdam.de](mailto:fweck@uni-potsdam.de)  Department of Clinical Psychology and Psychotherapy, University of Potsdam, Karl-Liebknecht-Straße 24-25, Potsdam D-14476, Germany

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/16506073.2024.2434016>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

from methodological weaknesses, such that a positive effect on therapists' competence cannot always be supported empirically (Alfonsson et al., 2018; Kühne et al., 2021). For example, only a few studies have investigated the effect of clinical supervision in randomized controlled trials or assessed therapist competence using standardized instruments and independent raters. Thus, methodologically rigorous studies are necessary to evaluate the effects and mechanisms of clinical supervision.

Clinical supervision can be administered in very different ways, because different methods can be implemented in supervision sessions (e.g. role plays, case discussion, observation of videotapes, demonstration of methods). In clinical supervision practice, methods like case discussion or information brokering have been used very frequently, and methods like observation of videotapes and role plays only seldom (Weck et al., 2017). Based on empirical investigations, we know little about the effectiveness of different supervision strategies (Henrich et al., 2023). Alfonsson et al. (2018, p. 218) concluded, in their systematic supervision review, that "based on empirical findings, it is not possible to investigate the empirical support for specific supervision characteristics".

In their systematic review, Milne et al. (2010) analyzed various supervision methods and considered the use of audio- or videotapes (or direct observation of the supervision sessions) and feedback as the most important strategies for effective supervision. Those findings are in line with a methodologically sound randomized controlled trial which investigated the effect of competence-feedback on therapist competence and treatment outcome (Weck et al., 2021). Competence-feedback based on an established competence scale (namely the Cognitive Therapy Scale; CTS; Young & Beck, 1980), was conducted by independent raters (experienced clinicians), and was given both quantitatively (score of improvement in the CTS items) and qualitatively (written suggestions to improve the therapy). The feedback was given five times (during 20 therapy sessions) regarding videotaped treatment sessions of patients with major depression (treated with cognitive-behavioral therapy). The competence-feedback led to a significant improvement of therapists' competence in that treatment. Live-supervision, which also enables immediate feedback to the supervisee, based on the direct insight of the supervisor in the treatment, yielded better effects on therapist skills (e.g. Maaß et al., 2024b; Weck et al., 2016), the therapeutic alliance (e.g. Probst et al., 2018), and empathy (e.g. Hodorowicz et al., 2020; Smith et al., 2012), than less direct forms of supervision or no supervision at all (for an overview, see Maaß et al., 2022b).

Whether supervisors have direct insight into the therapeutic action of the supervisees might be an important difference for clinical supervision. If supervisors have direct insight, they can give specific feedback on the action of the supervisees. If supervisors have no direct insight, they are dependent on the report from the supervisees. Moreover, the report might be biased, due to a lack of introspective abilities or anxieties on the part of the supervisees. For example, nondisclosure in clinical supervision was frequently reported by supervisees (Reichelt et al., 2009; Weck et al., 2023) and correlated with supervisee fear of negative evaluation (Junga et al., 2019). Therefore, direct insight into supervisees' therapies by audio- or videotapes might be an important feature of effective supervision and may be able to improve supervisees' competence, alliance, and empathy. In contrast, supervisees' level of anxiety was found to be higher in supervision formats which allow direct insight into the therapy session. In the study of Mauzey and Erdman (1997), higher anxiety levels were found in live-supervision than in delayed supervision.

Therefore, higher levels of anxiety can also be expected for video-based supervision in comparison to supervision which draws only on the verbal report of the supervisees.

As stated above, audio- or video-tapes are rarely used in clinical supervision. For example, 46% of psychotherapy trainees reported that audio- or videotapes were never used in supervision (Weck et al., 2017). Accordingly, a study which investigates clinical supervision based on videotapes, in comparison to supervision based on supervisees' verbal reports, could clarify the impact of either method and would have both scientific and practical merit.

In the current study, we investigated whether clinical supervision, which enables direct insight into the therapeutic session (video-based supervision) is more effective than supervision based only on the report of the supervisees (verbal report-based supervision). We hypothesized that video-based supervision leads to a greater increase in competence and communication scores than verbal report-based supervision in the subsequent therapeutic role plays (Hypothesis 1). Moreover, we hypothesized that video-based supervision has better effects on the therapeutic alliance (Hypothesis 2) and the therapists' empathy (Hypothesis 3), in comparison to verbal report-based supervision. Furthermore, we hypothesize that the level of supervisees' anxiety in therapeutic role plays is higher in video-based supervision than in verbal report-based supervision (Hypothesis 4).

## Method

The current study is a randomized controlled trial which investigated the effect of clinical supervision. The study was preregistered with the International Standard Randomized Controlled Trial Number (ISRCTN) registry, on 10 December 2019 (ISRCTN19173895). The study protocol was published in 2020 (Kühne et al., 2020) and approved by the ethics review board of the University of Potsdam (No. 9/2018). All participants gave written informed consent for participation. The study was conducted between July 2021 and March 2023.

Participants were psychology students (bachelor or master of science). Further inclusion criteria were that participants gave verbal and written informed consent to the study and agreement to the video recording of the role plays. Exclusion criteria were insufficient proficiency in German and currently undergoing psychotherapeutic treatment.

## Participants

### Supervisees

Seventy-three supervisees participated in the current study. Sixty-two (84.9%) were female and eleven (15.1%) male. They were  $M = 25.86$  ( $SD = 7.05$ ; range: 19–52) years old. Fifty-seven (78.1%) were studying for a bachelor degree (28.1% in the first semester) and sixteen (21.9%) a master of science in psychology. Most (78.1%,  $n = 57$ ) of the supervisees had no practical experience in clinical psychology and psychotherapy. The clinical experience of the other 21.9% ( $n = 16$ ) most entailed experience in a clinical internship which lasts  $M = 97.63$  ( $SD = 99.79$ ; range: 3–330) hours.

### **Supervisors**

Six female clinical psychologists served as supervisors. On average, they were 38.00 years old ( $SD = 3.16$ ; range: 35–44) and were licensed as psychotherapists for  $M = 6.33$  years ( $SD = 2.07$ ; range: 4–10). In Germany, licensed psychotherapists have to pass a three-year full-time training period after completing their bachelor and master degrees in clinical psychology. The training comprises at least 4,200 hrs. of training including workshops (min. 600 hrs.), practice in a clinic (min. 1,800 hrs.), practice in an outpatient clinic (min. 600 hrs.), supervision (min. 150 hrs.) and self-reflection (min. 120 hrs.). Supervisors were trained in a one-day workshop based on a supervision manual (Falender & Shafranske, 2017) and the cognitive behavioral treatment manual for depression (Hautzinger, 2013).

### **Standardized patients (SPs)**

Four female students and one male served as SPs. They were  $M = 24.80$  years old ( $SD = 2.68$ ; range: 22–28). SPs were studying degrees other than psychology at the University of Potsdam (e.g. sociology) and were not familiar with the supervisees. SPs were trained by a licensed psychotherapist (F.K.) in a two-day workshop (12 hours) according to Kühne et al. (2021). During the first workshop day, SPs received information about the symptoms and treatment of depression, watched videotapes of patients suffering from depression, and discussed features of a high-quality standardized role play (i.e. authenticity and consistence). As homework, SPs had to read six role play scripts of depressive patients. During the second workshop-day, the role scripts were first discussed, after which role plays were conducted and videotaped. Based on the videotapes, feedback on the authenticity of the role plays was provided, and further role plays were practiced.

In the current study, authenticity of SPs in the role plays was checked by the independent raters (see below) by using the Authenticity of Patient Demonstrations Scale (APD; S. D. Ay-Bryson et al., 2022). The APD is a 10-item 4-point scale which evaluates the level of authenticity of SPs, and the response format ranges from 0 to 3 (0 = strongly disagree, 1 = disagree, 2 = agree, 3 = strongly agree). In the first  $n = 146$  role plays, the APD score was  $M = 2.96$  ( $SD = 0.08$ ), indicating a high level of authenticity.

### **Independent raters**

Two female clinical psychologists (both 31 years old) served as independent raters. Both had at least three years of clinical experience and had treated patients with depression. All videotaped role plays were judged by the raters in random order by the following measures.

### **Measures**

Table 1 shows an overview and the reliabilities of the measures for all assessment perspectives and for all measurement times.

#### **Clinical communication skills scale – short form (CCSS-S)**

The CCSS-S (Maaß et al., 2022a) evaluates basic counseling skills with 14 items. Example of items include “the therapist uses easily understandable language” and “the therapist does not judge the patient”. The response format of the CCSS-S ranges from 0 to 3 (0 = not at all

**Table 1.** Reliability of all measures, for all perspectives, and at all measurement times (the first score was evaluated in the behavioral activation and the second in the cognitive restructuring condition).

Measures	Supervisees (Cronbach's $\alpha$ )			Standardized Patients (Cronbach's $\alpha$ )			Raters (intrate reliability, ICC)		
	<i>pre</i>	<i>post</i>	<i>follow-up</i>	<i>pre</i>	<i>post</i>	<i>follow-up</i>	<i>pre</i>	<i>post</i>	<i>follow-up</i>
CCCS-S	.86/.83	.84/.89	.85/.87	.95/.95	.94/.94	.92/.94	.61/.78	.65/.67	.80/.85
CTS							.60/.71	.65/.65	.75/.82
HAQ	.90/.82	.86/.82	.88/.90	.97/.97	.97/.97	.96/.96	.63/.72	.58/.59	.75/.80
ES	.89/.85	.81/.84	.83/.84	.74/.69	.68/.62	.81/.77	.45/.47	.41/.55	.73/.70
STAI	.86/.88	.79/.81	.88/.86	.80/.79	.77/.80	.66/.76			

CCCS-S = Clinical Communication Skills Scale-Short Form; CTS = Cognitive Therapy Scale; HAQ = helping Alliance Scale; ES = Empathy Scale; STAI = Short Version of the State-Trait Anxiety Inventory.

appropriately, 1 = not particularly appropriately, 2 = generally appropriately, 3 = entirely appropriately). In previous studies, the CCSS-S demonstrated its sensitivity to change within supervision conditions (e.g. Maaß et al., 2024b).

### **Cognitive Therapy Scale (CTS)**

The CTS (Young & Beck, 1980; German version: Weck et al., 2010) measures specific competencies that are relevant for cognitive behavioral therapy. In the current study, we implemented 11 of the 14 items of the CTS: Dealing with problems/questions/objections (Item 2), clarity of communication (Item 3), pacing and efficient use of time (Item 4), interpersonal effectiveness (Item 5), resource activation (Item 6), using feedback and summaries (Item 8), guided discovery (Item 9), focusing on central cognitions and behavior (Item 10), rationale (Item 11), appropriate implementation of techniques (Item 14), and assigning homework (Item 14). We excluded three items, because they were not eligible for the role plays in the current study: Item 1 (agenda setting), Item 7 (reviewing previously set homework), and Items 12 (selecting appropriate strategies). The response format of the CTS ranges from 0 to 6 (0 = poor, 1 = barely adequate, 2 = mediocre, 3 = satisfactory, 4 = good, 5 = very good, 6 = excellent). In previous studies, the CTS was able to predict patients' treatment outcome (e.g. Weck, Grikscheit, Höfling, et al., 2014) and demonstrated sensitivity to changes in training interventions (e.g. Kühne et al., 2022).

### **Helping Alliance Questionnaire (HAQ)**

The HAQ (Luborsky, 1984; German: Eich et al., 2018) measures the quality of the therapeutic alliance. The questionnaire consists of 11 items. Example items are "I feel the therapist understands me" and "I feel I am working together with the therapist in a joint effort". The response format of the HAQ ranges from 1–6 (1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = slightly agree, 5 = agree, 6 = strongly agree). In previous studies, the HAQ was found to be a predictor for treatment failure (e.g. Weck, Grikscheit, Jakob, et al., 2014) and was sensitive to change within training conditions (Kühne et al., 2022).

### **Empathy Scale (ES)**

The ES (Persons & Burns, 1985; Partschefeld et al., 2013) measures the level of therapist empathy and consists of 10 items. Examples are "He usually understands what I say to

him” and “I feel that he really cares what happens to me”. The response format of the ES ranges from 1 to 4 (1 = weak feeling, 2 = moderate feeling, 3 = strong feeling, and 4 = extremely strong feeling). The ES demonstrate sensitivity to change in previous training studies (e.g. Maaß et al., 2024a).

### **Short Version of The State-Trait Anxiety Inventory (STAI-SKD)**

The STAI-SKD (Spielberger et al., 1970; German: Englert et al., 2011) measures state anxiety with five items. An example reads as follows “I feel nervous”. The response format of the STAI-SKD ranges from 1 to 4 (1 = not at all, 2 = slightly, 3 = moderately, 4 = very). In previous studies, the STAI-SKD was able to identify increased anxiety scores of participants in training conditions (e.g. Kühne et al., 2022)

### **Study design and procedure**

Supervisees were randomly assigned (1:1) to the intervention group (IG; supervision based on video tapes;  $n = 36$ ) and the control group (CG; supervision based on verbal report;  $n = 37$ ; see Figure 1). Random sampling and assignment were implemented using the R software (v4.0.2) with the *randomizr* package (Coppock, 2019; R Core Team, 2020) and were implemented by an independent researcher. Supervisees in both conditions (IG and CG) were trained in two cognitive behavioral techniques (behavioral activation [BA] and cognitive restructuring [CR]), conducted a role play (20 minutes) to demonstrate the therapeutic techniques, received supervision (IG or CG) and conducted a second role play (20 minutes) to demonstrate the therapeutic techniques after supervision. All supervisees conducted role plays in both cognitive behavioral techniques (BA and CR). Using random assignment, half of the supervisees started with BA and half with CR (see Figure 1).

### **Training of supervisees**

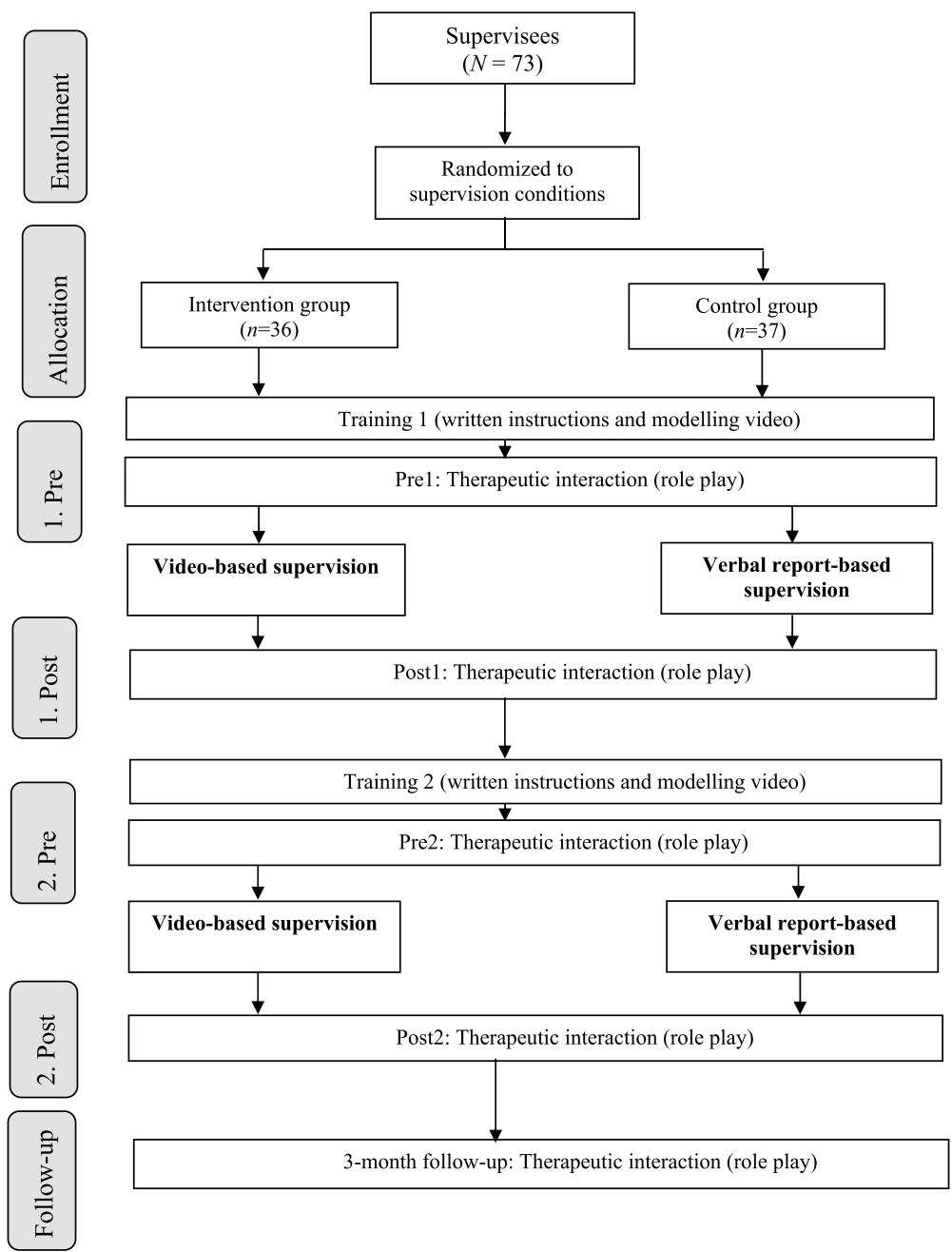
Training included written information regarding BA (12 pages) and CR (13 pages) based on a manual for cognitive-behavioral therapy for depression (Hautzinger, 2013). Supervisees were allowed 20 minutes for each technique to read the manualized information. Additionally, they watched a video tape of an experienced psychotherapist, who skillfully demonstrated BA (17:06 minutes) and CR (15:22 minutes).

### **Role plays**

Participants were instructed that role plays should last a maximum of 20 minutes and refer to BA or CR. The SPs roles were based on the previously discusses role scripts. These scripts described a depressed person with problems typical for patients with major depression (e.g. low rate of activities, cognitive biases). For one supervisee, the role plays (pre, post, and follow-up) were conducted by the same SP.

### **Supervision conditions**

In the IG, supervisors watched the role plays live through video-streaming to the supervisor’s PC (regarding BA and CR) and gave feedback immediately after each role play. Feedback was limited to 20 minutes, and based on the information presented in the role plays (i.e. in total 40 minutes duration of the interaction for BA and CR). In the CG, the



**Figure 1.** Study design and flowchart. The order of the task condition was randomized, that is, participants either completed the behavioral activation condition first and then the cognitive restructuring condition, or vice versa.



supervisors did not watch the role plays. Feedback was limited to 40 minutes and started immediately after each role play, and was based on the verbal report by the supervisees (i.e. in total 80 minutes duration of the interaction for BA und CR).

## Statistical analyses

### Reliability

Internal consistency of the measures was evaluated with Cronbach's  $\alpha$ . Coefficients between .70 and .80 were considered as reasonable, and coefficients between .80 and .90 as good (Moosbrugger & Kelava, 2020). Interrater reliability of the measures was evaluated by intraclass correlation coefficients (ICCs) using Model 2 (ICC<sub>(2, n)</sub>) (Shrout & Fleiss, 1979).; Coefficients below .40 were considered as poor, and between .40 and .59 as fair, between .60 and .74 as good, and higher coefficients as excellent (see Cicchetti, 1994).

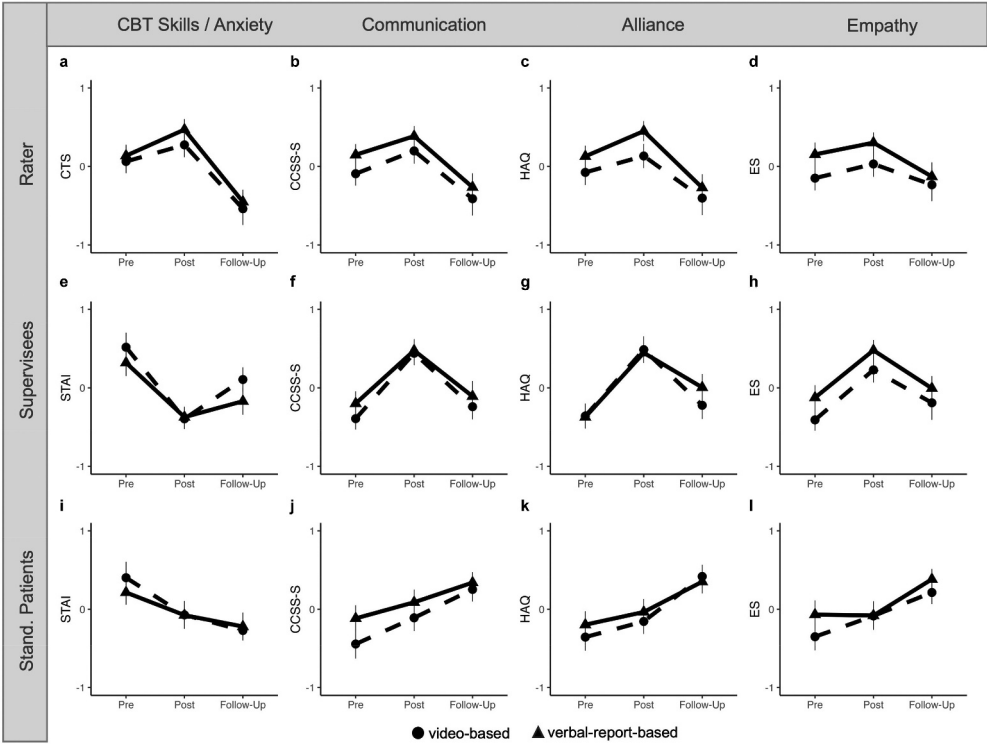
### Multilevel analysis

To answer the main research questions, we used a series of multilevel models (MLMs) for each perspective (i.e. rater, therapist, SPs), using R (v. 4.3.1, R Core Team, 2023) and the lme4 and lmer packages (Bates et al., 2015; Kuznetsova et al., 2017). The study group (IG vs. CG) was set at Level-2, time (pre, post, follow-up) was grand-mean centered and set at Level-1, and a cross-level interaction between group  $\times$  time was included. Based on power simulations (Arend & Schäfer, 2019), a sample size of  $N = 70$  is sufficient to detect at least moderate effect sizes of .50 for an interaction effect for each outcome (i.e. competence, alliance, empathy, and anxiety; Hypothesis 1–4) with a power  $\geq .80$ , assuming large ICCs, and three points in time. We specified the final models based on information from several preliminary analyses which tested for (a) differences between the two techniques BA and CR, (b) linear and polynomial trends, and (c) the inclusion of random intercepts and/or slopes.

*Test for differences between techniques.* Participants learned two techniques (BA and CR) per time point. In a first step, we therefore used repeated measures MANOVAs to test whether there were significant differences in outcomes between these techniques. If so, three-level MLMs with time points nested in supervisees and techniques were specified. As shown in the Supplementary Material (Supplement Tables 1 and 2), the technique did not significantly affect outcomes except for rater scores about communication skills (CCSS-S),  $F_{(1,427)} = 4.49, p = .035$ . However, the effect was very small ( $\eta_G^2 = .01$ ). In addition, the three-level MLM (Supplement Tables 3 and 4) was not substantially different from a model without nesting within techniques. For reasons of parsimony, the values (BA and CR) of all outcomes per time point were thus averaged across both tasks, and two-level MLMs were calculated (see above).

*Inclusion of linear and polynomial trends.* As can be seen in the plots of the outcome values across all time points (Figure 2), a curvilinear time trajectory (i.e. U-shaped curves) seemed appropriate. For this reason, we compared the model fits of MLMs with and without quadratic time effects (in addition to linear effects), based on the chi-square test of the deviances (Hox, 2013).

*Inclusion of random intercepts and/or slopes.* We compared four models with each other: an unconditional model (Model 0), a model with random intercepts and a linear time trend (Model 1a), a model with random intercepts and both



**Figure 2.** Time trajectories of all outcomes from the perspectives of raters, supervisees, and standardized patients. Scores were z-standardized across all time points and participants. CTS = German version of the Cognitive Therapy Scale. STAI=Short Version of the State-Trait-Anxiety Inventory. CCSS-S=Clinical Communication Skills Scale—Short Version. HAQ = helping Alliance Questionnaire. ES = Empathy Scale. Pre = pre-assessment. Post = post-assessment. Video-based = supervision based on verbal report of the supervisee. Verbal-report-based: supervision based on the video tape of the role play

a linear and quadratic time trend (Model 1b), and a model with random intercepts and random (linear) slopes (Model 2). We interpreted those MLMs with the best model fit. Supplement Table 3 displays all specified MLMs and their model fits. For all outcomes from the perspectives of rater and supervisees, MLMs with random intercepts, linear and quadratic time trends yielded the best model fits. For all outcomes from the perspective of SPs, MLMs with random intercepts and random slopes, and linear time trends, yielded the best model fits (only for empathy was a model without random slopes most appropriate).

## Results

### *Supervisions based on video tapes (IG) vs. supervision based on verbal reports (CG)*

Table 2 presents the descriptive data of all measures (mean scores of the total sample are reported in Supplement Table 5). Figure 2 presents the outcome of all measures from the perspective of raters, supervisees and SPs at pre, post, and follow-up. Table 3 displays the

**Table 2.** Descriptive statistics for all outcomes from the perspective of Rater, supervisees, and standardized patients ( $N = 73$ ).

Intervention Group										Control Group										Total Sample																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															
pre					post					FU					pre					post					pre vs. post					pre vs. FU					post vs. FU																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
M					SD					M					SD					M					SD					M					SD					M					SD					M					SD																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																							
d					d					d					d					d					d					d					d					d					d					d					d					d					d					d																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>					C <sub>I</sub>																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																								
C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>					C <sub>U</sub>									

**Table 3.** Fixed effects of multilevel models ( $N = 73$ ).

	CTS/STAI			CCSS-S			HAQ			ES		
	B [95% CI]	SE	p	B [95% CI]	SE	p	B [95% CI]	SE	p	B [95% CI]	SE	p
<i>Raters</i>												
Intercept	2.12 [1.94, 2.30]	0.09	<.001	1.74 [1.63, 1.85]	0.06	<.001	3.42 [3.23, 3.60]	0.09	<.001	3.31 [3.22, 3.40]	0.05	<.001
Time	<b>-2.53 [-3.47, -1.58]</b>	<b>0.49</b>	<b>&lt;.001</b>	<b>-0.91 [-1.54, -0.28]</b>	<b>0.32</b>	<b>.005</b>	<b>-1.49 [-2.48, -0.51]</b>	<b>0.51</b>	<b>.003</b>	-0.25 [-0.72, 0.22]	0.24	0.304
Time <sup>2</sup>	<b>-2.30 [-3.31, -1.45]</b>	<b>0.48</b>	<b>&lt;.001</b>	<b>-1.20 [-1.90, -0.67]</b>	<b>0.32</b>	<b>&lt;.001</b>	<b>-1.70 [-2.75, -0.82]</b>	<b>0.50</b>	<b>&lt;.001</b>	<b>-0.50 [-0.97, -0.05]</b>	<b>0.24</b>	<b>0.031</b>
Group	0.11 [-0.14, 0.36]	0.13	.395	0.10 [-0.06, 0.26]	0.08	.238	0.17 [-0.09, 0.44]	0.13	.200	0.08 [-0.05, 0.21]	0.06	0.215
Time x Group	0.23 [-1.10, 1.55]	0.68	.737	-0.09 [-0.98, 0.79]	0.45	.837	-0.12 [-1.50, 1.26]	0.71	.863	-0.29 [-0.94, 0.37]	0.34	0.395
Time <sup>2</sup> x Group	-0.40 [-1.76, 0.85]	0.67	.500	0.00 [-0.84, 0.90]	0.45	.945	-0.60 [-2.00, 0.72]	0.70	.365	-0.10 [-0.77, 0.52]	0.33	0.711
ICC	0.702			0.685			0.707			0.714		
<i>Supervisors</i>												
Intercept	<b>2.00 [1.84, 2.16]</b>	<b>0.08</b>	<b>&lt;.001</b>	<b>1.96 [1.86, 2.07]</b>	<b>0.06</b>	<b>&lt;.001</b>	<b>4.15 [3.98, 4.31]</b>	<b>0.08</b>	<b>&lt;.001</b>	<b>3.34 [3.24, 3.43]</b>	<b>0.05</b>	<b>&lt;.001</b>
Time	<b>-1.26 [-2.20, -0.31]</b>	<b>0.49</b>	<b>.011</b>	0.30 [-0.26, 0.85]	0.29	.300	0.49 [-0.55, 1.53]	0.54	.361	0.44 [-0.03, 0.91]	0.24	.068
Time <sup>2</sup>	<b>3.00 [2.16, 4.02]</b>	<b>0.48</b>	<b>&lt;.001</b>	<b>-2.0 [-2.62, -1.53]</b>	<b>0.28</b>	<b>&lt;.001</b>	<b>-3.50 [-4.58, -2.53]</b>	<b>0.53</b>	<b>&lt;.001</b>	<b>-1.20 [-1.69, -0.78]</b>	<b>0.23</b>	<b>&lt;.001</b>
Group	-0.10 [-0.33, 0.12]	0.11	.380	0.06 [-0.09, 0.21]	0.08	.434	0.04 [-0.19, 0.27]	0.12	.746	0.09 [-0.04, 0.22]	0.07	.192
Time x Group	-0.54 [-1.86, 0.79]	0.68	.432	-0.05 [-0.83, 0.72]	0.40	.896	0.91 [-0.55, 2.37]	0.75	.225	-0.17 [-0.82, 0.48]	-0.34	.618
Time <sup>2</sup> x Group	-1.20 [-2.55, 0.07]	0.67	.067	0.30 [-0.40, 1.13]	0.39	.358	0.70 [-0.70, 2.18]	0.74	.317	0.00 [-0.63, 0.66]	0.33	.959
ICC	0.649			0.723			0.613			0.724		
<i>Standardized Patients</i>												
Intercept	<b>1.28 [1.20, 1.36]</b>	<b>0.04</b>	<b>&lt;.001</b>	<b>2.58 [2.48, 2.68]</b>	<b>0.05</b>	<b>&lt;.001</b>	<b>4.98 [4.77, 5.19]</b>	<b>0.11</b>	<b>&lt;.001</b>	<b>3.62 [3.53, 3.71]</b>	<b>0.05</b>	<b>&lt;.001</b>
Time	<b>-0.11 [-0.17, -0.05]</b>	<b>0.03</b>	<b>&lt;.001</b>	<b>0.14 [0.06, 0.22]</b>	<b>0.04</b>	<b>&lt;.001</b>	<b>0.32 [0.16, 0.48]</b>	<b>0.08</b>	<b>&lt;.001</b>	<b>0.10 [0.04, 0.16]</b>	<b>0.03</b>	<b>.002</b>
Group	-0.03 [-0.14, 0.08]	0.06	.608	0.10 [-0.04, 0.25]	0.08	.176	0.09 [-0.21, 0.38]	0.15	.569	0.07 [-0.06, 0.19]	0.06	.306
Time x Group	0.05 [-0.04, 0.13]	0.04	.305	-0.05 [-0.16, 0.06]	0.05	.351	-0.10 [-0.32, 0.13]	0.11	.400	-0.02 [-0.11, 0.06]	0.04	.595
ICC	0.721			0.739			0.708			0.439		

Significant effects are in bold; CTS = German version of the Cognitive Therapy Scale; STAI-SKD = Short Version of the State-Trait-Anxiety Inventory; CCSS-S = Clinical Communication Skills Scale —Short Version; HAQ = helping Alliance Questionnaire; ES = Empathy Scale; Group = study group (1 = intervention group, video-based supervision; 2 = control group, supervision based on verbal report).

results of the MLMs from the perspectives of rater, trainee, and SPs. In neither model, were there significant group or group  $\times$  time effects, indicating that IG and CG did not yield significantly different scores in general and over time. However, all models showed significant time effects.

### *Perspective of independent raters*

There were significant negative quadratic time trends in addition to negative linear trends in all models (except for empathy, which included a significant quadratic trend only). This means that supervisees' skills, the therapeutic alliance, and empathy increased from pre to post, but decreased again at follow-up (inverted U-shaped relationship). The negative linear effect showed that the scores were significantly lower at follow-up than at pre-assessment (except for empathy, see Table 2).

### *Perspective of supervisees*

There were significant quadratic time trends in all models, indicating that supervisees perceived an increase in competence, therapeutic alliance, and empathy from pre- to post-assessment, followed by a decrease at follow-up. However, unlike the raters, supervisees did not evaluate their skills, therapeutic alliance, and empathy at follow-up as significantly lower than at pre-assessment (linear effect *n.s.*). Anxiety levels decreased from pre to post and increased again at follow-up—nonetheless, supervisees reported being less anxious at follow-up than at pre-assessment (see Table 2).

### *Perspective of standardized patients*

There were only linear time effects in all models, indicating that SPs perceived a significant increase in supervisees' skills, therapeutic alliance, and empathy from pre to follow-up (see Table 2). In addition, they found that supervisees seemed significantly less anxious over time.

## **Discussion**

We investigated the efficacy of two supervision methods (video-based vs. verbal report-based) on supervisee competence, therapeutic alliance, and supervisee empathy in role plays with SPs. In addition, the level of anxiety of supervisees was monitored. Both supervision methods demonstrated an improvement in rater-based competence, alliance, and empathy from pre to post assessment. However, a decrease from post- to three-month follow-up assessment was observed. Contrary to our hypotheses, video-based supervision did not lead to greater competence and communication scores than verbal report-based supervision (Hypothesis 1), therapeutic alliance (Hypothesis 2), and supervisees' empathy (Hypothesis 3). The level of anxiety was not higher in video-based supervision than in verbal report-based supervision (Hypothesis 4).

Altogether, supervisees' cognitive behavioral competencies and communication skills improved in standardized role plays from pre to post assessment. No differences were found between supervision conditions, such that both types of supervision seem to be comparably effective in improving therapeutic competencies (Hypothesis 1) in a controlled clinical training setting. Contrary to our hypothesis, the video-based insight

into the therapeutic role plays did not lead to greater increases in competence than the verbal report of the supervisees (with no direct insight into the therapeutic session). The current study provides evidence that supervisions based on a verbal report of the supervisees, are neither more nor less effective than those which use video tapes to obtain direct insight. This finding is important, given that most supervision in clinical practice does not use videotapes (e.g. Weck et al., 2017). Similarly, the video-based supervision also seems to be an effective alternative to verbal report-based supervision. On the one hand, in our study, both supervision formats were equally time-consuming: video-based supervision included 20-minutes of video watching and 20-minutes of supervision session, verbal report-based supervision included 40-minutes of supervision session. On the other hand, verbal report-based supervision allows a longer interaction time between the supervisor and the supervisee. Therefore, it is possible that both supervision formats display comparable effects on therapeutic competencies, but work differently and focus on different aspects of the therapeutic process. For example, it is possible that verbal report-based supervision leads to a better understanding of the needs and questions of the supervisees (e.g. due to longer discussion time between supervisor and supervisee), while video-based supervision leads to a better understanding of how supervisees implemented therapeutic techniques. In future studies, structured qualitative interviews conducted with the supervisors and supervisees might be a useful methodological approach to investigating mechanisms of change in the given supervision formats.

Generally, it is necessary to bear in mind that we had established only two active training conditions (including supervision) and no third training condition without supervision. Therefore, we cannot be sure that the positive improvements are caused by the given supervision only. For example, time effects, attention to training participants, or positive expectations might be also relevant factors for improvement. Therefore, findings should be replicated by using a further control group (without supervision).

Generally, the effects of supervision were detected after supervisees were trained by means of written information and modelling videos. Such training methods (without supervision) demonstrated their efficacy in previous studies (e.g. Kühne et al., 2022). We assume that the positive effect of the training (due to written information and modelling videos) disappeared in the follow-up assessment, which led to lower scores in comparison to the pre-assessment. This issue suggests that further training is necessary to obtain improvements.

As stated above, we did not establish a further control group without supervision; we therefore cannot be sure that improvements are caused purely by the implemented supervision methods. Alternatively, the repeated performance in the role plays might lead to improvements in supervisee competence. However, previous studies did not find competence improvements in role plays when therapists did not receive supervision (Sholomskas et al., 2005) or when trainees only performed role plays repeatedly (Kühne et al., 2022). Moreover, the effects on supervisee competencies decreased in the three-month follow-up assessment. Again, this suggests that additional supervision might be necessary to maintain or improve therapeutic competencies in the current study. Also, previous empirical studies have found that clinical supervision is the essential intervention for improving therapeutic competencies (Alfonsson et al., 2020).

For the therapeutic alliance and supervisee empathy, we found improvements after supervision, but no differences between the supervision conditions (Hypotheses 2 and 3).

This is important, because in psychotherapy research, the therapeutic alliance is considered an important variable for therapy outcome (Flückiger et al., 2018).

Supervisees reported that their level of anxiety decreased after supervision, and no significant differences between supervision methods were found (Hypothesis 3). At follow-up, the level of anxiety increased, but was still lower than at pre-assessment. Therefore, a positive trend can be noted, as anxiety was lower than in the first role plays. Because of the absence of an additional control group (without supervision), we cannot exclude the possibility that the repeated conduction of role plays and anxiety-reducing mechanisms (e.g. habituation) caused the reduction in anxiety level. In contrast to the supervisee perspective, SPs did not perceive an increase in supervisee anxiety level at follow-up, but a continuous decrease in the anxiety level. Considering this discrepancy, we believe that the self-evaluation of supervisees is more important than the evaluation of supervisees by the SPs, because a high level of supervisees' self-perceived anxiety probably influences the performance more powerfully than the perspective of the SPs.

Generally, the evaluations of independent raters and supervisees were comparable, but differed from the perspective of SPs, i.e. SPs did not report a decrease in competences, alliance, and empathy in the 3-month follow-up assessment. It would be interesting to determine whether real patients would evaluate this in a similar manner. However, we argue that the perspective of the independent raters is more important, because they did not know which video tape belongs to which assessment time. In contrast, evaluations of SPs might be biased, because they knew that they were conducting a follow-up role play and that supervisees were trained and had been supervised before. An interesting and useful approach would be to use different and blinded SPs for each assessment point, without the SPs having information about the training status.

The current study makes a variety of useful contributions. To achieve a high level of internal validity, we used a laboratory setting with standardized role plays and SPs. Participants were randomly assigned to the supervision conditions, and conducted two role plays including two important methods of cognitive behavioral therapy (i.e. behavioral activation and cognitive restructuring). Evaluations were conducted by means of different perspectives (i.e. independent raters, supervisees, and SPs), and several relevant outcome variables (i.e. competence, therapeutic alliance, empathy, and anxiety) were considered. SPs were found to be highly authentic, which enhanced the external validity of the study.

## **Limitations**

First, the external validity of the study is limited because of the use of role plays instead of real therapy sessions. Role plays were also used successfully in previous training studies (e.g. Kühne et al., 2021) and SPs were found to be indistinguishable from real patients (D. S. Ay-Bryson et al., 2023; S. D. Ay-Bryson et al., 2022). However, even though SPs in our study were trained and were evaluated as highly authentic, they were not real patients and the role plays were not real therapy sessions. Therefore, the results should be replicated in real therapy and supervision sessions. In addition, the supervision sessions were conducted immediately after the role plays, and the second role plays were conducted immediately after the supervision session. In clinical practice, temporal distances between therapy sessions and supervision sessions are usually longer (i.e. several days). Second, we considered only

cognitive-behavioral interventions methods, thus excluding other therapeutic approaches. Therefore, the findings cannot be generalized fully to other therapeutic approaches (e.g. psychodynamic therapy). Third, supervisees were in the first stage of their clinical training (i.e. in the BSc or MSc course for clinical psychology). Unfortunately, our sample size was too small to analyze training effects as a function of supervisees' university years of study. Furthermore, it is unclear whether a lack of differences between supervision methods would emerge for trainees in advanced training phases or for licensed psychotherapists as well. Future studies should also consider more advanced trainees and experienced psychotherapists, in order to investigate the effects of different supervision methods. Fourth, 85% of the supervisees were female, with other gender identities being underrepresented. However, most (about 90%) psychology students in Germany are in fact female. Therefore, the supervisor sample conformed to the German gender distribution of psychology students. Fifth, some of our measures had only "fair" reliability scores (see Table 1) which could have contributed to the lack of group differences. Finally, the sample size was too small to detect potential small effects.

## Conclusion

Our study has important implications for psychotherapy training. In the video-based and the verbal report-based supervision, improvements in therapeutic competencies, therapeutic alliance, and empathy were observed in a controlled clinical setting. Therefore, the current practice of using the verbal reports of supervisees for conducting supervision sessions may not be less effective than supervision based on video tapes. However, different mechanisms of action of the different supervision formats are possible, and should be considered in future studies. In addition, video-based supervision seems to be an effective alternative to verbal report-based supervision, without being more time-consuming and without producing any more negative side effects (in terms of supervisee anxiety). However, additional supervision seems to be necessary in the follow-up period in order to maintain training effects over time. To confirm that improvements in our study were indeed caused by the supervision methods used, an additional control group (without supervision) should be established in future studies. Furthermore, the findings of our laboratory study should be replicated in naturalistic clinical settings to ensure the external validity of our findings. A general open question in previous training research on psychotherapeutic competencies is the extent to which competence gains are also relevant in practice, especially with small to medium effect sizes. Even if the answer to this question is beyond the scope of the present study, future studies should address the aspects of clinical significance and reliable change.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

The work was supported by the Deutsche Forschungsgemeinschaft [KU 3790/2-1 and WE 4654/10-1].



## ORCID

Florian Weck  <https://orcid.org/0000-0001-9621-3227>  
 Ulrike Maaß  <https://orcid.org/0000-0001-7969-8250>  
 Tatjana Paunov  <https://orcid.org/0009-0007-7184-9473>  
 Peter E. Heinze  <https://orcid.org/0000-0002-1998-6485>  
 Franziska Kühne  <https://orcid.org/0000-0001-9636-5247>

## References

- Alfonsson, S., Lundgren, T., & Andersson, G. (2020). Clinical supervision in cognitive behavior therapy improves therapists' competence: A single-case experimental pilot study. *Cognitive Behaviour Therapy*, 49(5), 425–438. <https://doi.org/10.1080/16506073.2020.1737571>
- 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(3), 206–228. <https://doi.org/10.1080/16506073.2017.1369559>
- Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, 24(1), 1–19. <https://doi.org/10.1037/met0000195>
- Ay-Bryson, D. S., Weck, F., & Kühne, F. (2023). Can students in simulation portray a psychotherapy patient authentically with a detailed role-script? Results of a randomized-controlled study. *Training and Education in Professional Psychology*, 17(1), 89–97. <https://doi.org/10.1037/tep0000388>
- Ay-Bryson, S. D., Weck, F., & Kühne, F. (2022). Can simulated patient encounters appear authentic? Development and pilot results of a rating instrument based on the portrayal of depressive patients. *Training and Education in Professional Psychology*, 16(1), 20–27. <https://doi.org/10.1037/tep0000349>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <http://doi.org/10.18637/jss.v067.i01>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 282–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Coppock, A. (2019). *Randomizr: Easy-to-use tools for common forms of random assignment and sampling* [Software]. R-project. <https://CRAN.R-project.org/package=randomizr>
- Eich, H. S., Kriston, L., Schramm, E., & Bailer, J. (2018). The German version of the helping alliance questionnaire: Psychometric properties in patients with persistent depression disorder. *BMC Psychiatry*, 18(1), 107. <https://doi.org/10.1186/s12888-018-1697-8>
- Englert, C., Bertrams, A., & Dickhäuser, O. (2011). Entwicklung der Fünf-Item-Kurzskala STAI-SKD zur Messung von Zustandsangst [Development of a 5-Item-Short scale STAI-SKD to measure state fear]. *Zeitschrift für Gesundheitspsychologie*, 19(4), 173–180. <https://doi.org/10.1026/0943-8149/a000049>
- Falender, C. A., & Shafranske, E. P. (2017). *Supervision essentials for the practice of competency-based supervision*. American Psychological Association.
- Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy Theory, Research, Practice, Training*, 55(4), 316–340. <https://doi.org/10.1037/pst0000172>
- Frank, H. E., Becker-Haimes, E. M., & Kendall, P. C. (2020). Therapist training in evidence-based interventions for mental health: A systematic review of training approaches and outcomes. *Clinical Psychology Science & Practice*, 27(3), e12330. <https://doi.org/10.1111/cpsp.12330>
- Hautzinger, M. (2013). *Kognitive Verhaltenstherapie bei Depression* [Cognitive-behavioral therapy for depression] (7th ed.). Beltz.

- Henrich, D., Glombiewski, J. A., & Scholten, S. (2023). Systematic review of training in cognitive-behavioral therapy: Summarizing effects, costs and techniques. *Clinical Psychology Review*, 101, 102266. <https://doi.org/10.1016/j.cpr.2023.102266>
- Hodorowicz, M. T., Barth, R., Moyers, T., & Strieder, F. (2020). A randomized controlled trial of two methods to improve motivational interviewing training. *Research on Social Work Practice*, 30(4), 382–391. <https://doi.org/10.1177/1049731519887438>
- Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In T. D. Little (Ed.), *The oxford handbook of quantitative methods: Statistical analysis* (pp. 281–294). Oxford University Press.
- Junga, Y. M., Witthöft, M., & Weck, F. (2019). Assessing therapist development: Reliability and validity of the supervisee levels questionnaire (SLQ-R). *Journal of Clinical Psychology*, 75(9), 1658–1672. <https://doi.org/10.1002/jclp.22794>
- Keum, B. T., & Wang, L. (2021). Supervision and psychotherapy process and outcome: A meta-analytic review. *Translational Issues in Psychological Science*, 7(1), 89–108. <https://doi.org/10.1037/tps0000272>
- Kühne, F., Heinze, P. E., Maaß, U., & Weck, F. (2022). Modeling in psychotherapy training - a randomized controlled proof-of-concept trial. *Journal of Consulting & Clinical Psychology*, 90(12), 950–956. <https://doi.org/10.1037/ccp0000780>
- Kühne, F., Heinze, P. E., & Weck, F. (2020). Standardized patients in psychotherapy training and clinical supervision: Study protocol for a randomized controlled trial. *Trials*, 21(1), Article 276. <https://doi.org/10.1186/s13063-020-4172-z>
- Kühne, F., Maas, J., Wiesensthal, S., & Weck, F. (2019). Empirical research in clinical supervision: A systematic review and suggestions for future studies. *BMC Psychologie*, 7(1), 54. <https://doi.org/10.1186/s40359-019-0327-7>
- Kühne, F., Maaß, U., & Weck, F. (2021). Standardized patients in clinical psychology: From research to practice. *Verhaltenstherapie*, 31(2), 152–160. <https://doi.org/10.1159/000510049>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Luborsky, L. (1984). *Principles of psychoanalytic psychotherapy: A manual for supportive- expressive psychotherapy*. Basic Books.
- Maaß, U., Fehm, L., Kühne, F., Wenzel, H., & Weck, F. (2024a). *Aus Fehlern wird man klug? - Eine randomisierte kontrollierte Studie zum Erwerb von Gesprächsführungswissen anhand positiver vs. Gemischter Therapiemodelle* [Learning from mistakes? - A randomized controlled trial on the acquisition of communication knowledge using positive vs. mixed therapy models.]. PPmP - Psychotherapie Psychosomatik Medizinische Psychologie. <https://doi.org/10.1055/a-2359-7916>
- Maaß, U., Kühne, F., Ay-Bryson, D. S., Heinze, P. E., & Weck, F. (2024b). Efficacy of live-supervision regarding skills, anxiety and self-efficacy: A randomized controlled trial. *The Clinical Supervisor*, 43(1), 1–21. <https://doi.org/10.1080/07325223.2023.2267528>
- Maaß, U., Kühne, F., Heinze, P. E., Ay-Bryson, D. S., & Weck, F. (2022a). The concise measurement of clinical communication skills: Validation of a short scale. *Frontiers in Psychiatry*, 13, 977324. <https://doi.org/10.3389/fpsy.2022.977324>
- Maaß, U., Kühne Poltz, N., Ay-Bryson, A., Lorenz, D. S., Weck, F., & Weck, F. (2022b). Live supervision in psychotherapy training—A systematic review. *Training and Education in Professional Psychology*, 16(2), 130–142. <https://doi.org/10.1037/tep0000390>
- Mauzey, E., & Erdman, P. (1997). Trainee perceptions of live supervision phone-ins. *The Clinical Supervisor*, 15(2), 115–128. [https://doi.org/10.1300/J001v15n02\\_09](https://doi.org/10.1300/J001v15n02_09)
- Milne, D. L., Reiser, R., Aylott, H., Dunkerley, C., Fitzpatrick, H., & Wharton, S. (2010). The systematic review as an empirical approach to improving CBT supervision. *International Journal of Cognitive Therapy*, 3(3), 278–294. <https://doi.org/10.1521/ijct.2010.3.3.278>
- Milne, D. L., & Watkins, C. E. (2014). Defining and understanding clinical supervision: A functional approach. In C. E. Watkins & D. Milne (Eds.), *The Wiley international handbook of clinical supervision* (pp. 3–19). Wiley.

- Moosbrugger, H., & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion* [Test theory and test construction]. Springer. <https://doi.org/10.1007/978-3-662-61532-4>
- Partschefeld, E., Strauß, B., Geyer, M., & Philipp, S. (2013). Simulationspatienten in der Psychotherapieausbildung [Simulated patients in psychotherapy training]. *Psychotherapeut*, 58(5), 438–445. <https://doi.org/10.1007/s00278-013-1002-8>
- Persons, J. B., & Burns, D. D. (1985). Mechanisms of action of cognitive therapy: The relative contributions of technical and interpersonal interventions. *Cognitive Therapy and Research*, 9(5), 539–551. <https://doi.org/10.1007/BF01173007>
- Probst, T., Jakob, M., Kaufmann, Y. M., Müller-Neng, J. M. B., Bohus, M., & Weck, F. (2018). Patients' and therapists' experiences of general change mechanisms during bug-in-the-eye and delayed video-based supervised cognitive-behavioral therapy. A randomized controlled trial. *Journal of Clinical Psychology*, 74(4), 509–522. <https://doi.org/10.1002/jclp.22519>
- Rakovshik, S. G., & McManus, F. (2010). Establishing evidence-based training in cognitive therapy: A review of current empirical findings and theoretical guidance. *Clinical Psychology Review*, 30(5), 496–516. <https://doi.org/10.1016/j.cpr.2010.03.004>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reichelt, S., Gullestad, S. E., Hansen, B. R., Rønnestad, M. H., Torgersen, A. M., Jacobsen, C. H., Nielsen, G. H., & Skjerve, J. (2009). Nondisclosure in psychotherapy group supervision: The supervisee perspective. *Nordic Psychology*, 61(4), 5–27. <https://doi.org/10.1027/1901-2276.61.4.5>
- Sholomskas, D. E., Syracuse-Siewert, G., Rounsaville, B. J., Ball, S. A., Nuro, K. F., & Carroll, K. M. (2005). We Don't train in vain: A dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. *Journal of Consulting & Clinical Psychology*, 73(1), 106–115. <https://doi.org/10.1037/0022-006X.73.1.106>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Smith, J. L., Carpenter, K. M., Amrhein, P. C., Brooks, A. C., Levin, D., Schreiber, E. A., Travaglini, L. A., Hu, M.-C., & Nunes, E. V. (2012). Training substance abuse clinicians in motivational interviewing using live supervision via teleconferencing. *Journal of Consulting & Clinical Psychology*, 80(3), 450–464. <https://doi.org/10.1037/a0028176>
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the state-trait anxiety inventory*. Consulting Psychologists Press.
- Weck, F., Grikscheit, F., Höfling, V., & Stangier, U. (2014). Assessing treatment integrity in cognitive-behavioral therapy: Comparing session segments with entire sessions. *Behavior Therapy*, 45(4), 541–552. <https://doi.org/10.1016/j.beth.2014.03.003>
- Weck, F., Grikscheit, F., Jakob, M., Höfling, V., & Stangier, U. (2014). Treatment failure in cognitive-behavioural therapy: Therapeutic alliance as a precondition for an adherent and competent implementation of techniques. *British Journal of Clinical Psychology*, 54(1), 91–108. <https://doi.org/10.1111/bjc.12063>
- Weck, F., Hautzinger, M., Heidenreich, T., & Stangier, U. (2010). Erfassung psychotherapeutischer Kompetenz: Validierung einer deutschsprachigen Version der Cognitive Therapy Scale [Assessing psychotherapeutic competencies: Validation of a German version of the cognitive therapy scale]. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 39(4), 244–250. <https://doi.org/10.1026/1616-3443/a000055>
- Weck, F., Jakob, M., Neng, J. M. B., Höfling, V., Grikscheit, F., & Bohus, M. (2016). The effects of bug-in-the-eye supervision on therapeutic alliance and therapist competence in cognitive-behavioural therapy: A randomized controlled trial. *Clinical Psychology & Psychotherapy*, 23(5), 386–396. <https://doi.org/10.1002/cpp.1968>
- Weck, F., Junga, Y. M., Hahn, D., & Witthöft, M. (2023). Effects of competence-feedback on psychotherapy trainees' self-perceived competence, professional self-confidence, and self-disclosure: A secondary analysis of a randomized controlled trial. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 52(3), 142–152. <https://doi.org/10.1026/1616-3443/a000715>
- 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 & Clinical Psychology*, 89(11), 885–897. <https://doi.org/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. <https://doi.org/10.1017/S1754470X17000046>
- Young, J., & Beck, A. T. (1980). *Cognitive therapy scale rating manual* [Unpublished manuscript]. Center for Cognitive Therapy.