Cognitive Therapy and Research (2020) 44:241–257
https://doi.org/10.1007/s10608-019-10056-5

ORIGINAL ARTICLE

How Reliable Are Therapeutic Competence Ratings? Results of a Systematic Review and Meta-Analysis

Franziska Kühne[1] · Ramona Meister[2] · Ulrike Maaß[1] · Tatjana Paunov[1] · Florian Weck[1]

Published online: 16 November 2019

Abstract

Assessments of psychotherapeutic competencies play a crucial role in research and training. However, research on the reliability and validity of such assessments is sparse. This study aimed to provide an overview of the current evidence and to provide an average interrater reliability (IRR) of psychotherapeutic competence ratings. A systematic review was conducted, and 20 studies reported in 32 publications were collected. These 20 studies were included in a narrative synthesis, and 20 coefficients were entered into the meta-analysis. Most primary studies referred to cognitive-behavioral therapies and the treatment of depression, used the Cognitive Therapy Scale, based ratings on videos, and trained the raters. Our meta-analysis revealed a pooled ICC of 0.82, but at the same time severe heterogeneity. The evidence map highlighted a variety of variables related to competence assessments. Further aspects influencing the reliability of competence ratings and regarding the considerable heterogeneity are discussed in detail throughout the manuscript.

Keywords Competency · Therapist competence · Adherence · Psychotherapy · Assessment

Introduction

Psychotherapeutic competence is conceptualized as a therapist's general and treatment-specific knowledge level, skill level, values or attitudes while implementing therapeutic interventions (Muse and McManus 2016; Roth and Pilling 2007; Waltz et al. 1993). Barber et al. (2007) refer to psychotherapeutic competence more comprehensively as "the judicious application of communication, knowledge, technical skills, clinical reasoning, emotions, values, and contextual understanding for the benefit of the individual and community being served" (p. 494). Waltz et al. (1993) describe patient-specific aspects (such as symptoms, impairment or life situation) and treatment-specific variables (such as therapy stage, improvement or timing of interventions) to be considered for a broad perspective on competence.

The assessment of competencies not only plays a crucial role in treatment integrity in general but also may facilitate quality control during training, licensure and ongoing practice, may provide therapists with formative and summative feedback and may guide self-reflection (Muse and McManus 2013). However, meta-analyses on the association between therapeutic competence and patient outcomes yield results from no to small effects (Webb et al. 2010) or small to moderate effects (Zarafonitis-Müller et al. 2014). Further, the reviews report on a variety of competence measures—from the Cognitive Therapy Scale to the Collaborative Study Psychotherapy Rating Scale or study-specific developments, and they depict enormous variability in reliability—from no to nearly perfect agreement (Muse and McManus 2013; Zarafonitis-Müller et al. 2014). The Cognitive Therapy Scale (CTS; Young and Beck 1980), or Cognitive Therapy Rating Scale (CTRS; Beck Institute for Cognitive Behavior Therapy 2019), is a commonly used measure (Kazantzis et al. 2018). It was revised repeatedly, with the most prominent version being the Cognitive Therapy Scale-Revised (CTS-R; Blackburn et al. 2001; for a detailed description of different versions please see Muse and McManus 2013; Kazantzis et al. 2018).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10608-019-10056-5) contains supplementary material, which is available to authorized users.

✉ Franziska Kühne
dr.franziska.kuehne@uni-potsdam.de

[1] Department of Psychology, Clinical Psychology and Psychotherapy, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

[2] Department of Medical Psychology, University Medical Center Hamburg Eppendorf, Hamburg, Germany

Since therapeutic competence depends on the complexity of the patient's presentation, patient outcomes are not recommended unreservedly as a proxy for competence (Muse and McManus 2013). While ratings of in-session therapeutic skills performed by independent raters are highly recommended, Muse and McManus (2013) note that research on the reliability and validity of such competence assessments is sparse and that specifically regarding cognitive-behavioral therapy (CBT), "it is currently not possible to make evidence-based recommendations about how best to assess CBT competence" (p. 496).

Although the reliability of competence ratings is often considered improvable (Fairburn and Cooper 2011; Muse and McManus 2013), a number of variables are theorized to influence it. Rater training is consistently deemed central (Barber et al. 2007; Fairburn and Cooper 2011; Muse and McManus 2013, 2016). The same holds for rater expertise, but the definition and amount of expertise required is not always clear (Barber et al. 2007; Muse and McManus 2013, 2016). Other variables discussed in the literature are the number of raters, rater independence, the number of sessions rated per patient, the number of patients rated per therapist, the form of treatment, the stage of therapy, patient diagnosis and the competence scale used (Barber et al. 2007; Dennhag et al. 2012b; Fairburn and Cooper 2011; Muse and McManus 2013, 2016; Webb et al. 2010). Given these many variables, measurement quality (and thus reliability) is influenced by aspects related to rater, sample or instrument used (Kottner et al. 2011).

Moreover, there are various reliability measures, e.g. Cohen's κ for nominal data, Kendall's τ for ordinal data, or, depending on the model and number of raters, different forms of intra-class correlations for continuous data, to mention only some (Wirtz and Caspar 2002). Although a variety of primary studies examine psychotherapeutic competence ratings, to our knowledge, no evidence synthesis on their reliability has yet been published. Therefore, the first aim of the current study was to map the evidence regarding the interrater reliability (IRR) of psychotherapeutic competence ratings, and the second was to estimate the pooled IRR across methodologically sound studies. The third explorative aim of the study was to investigate moderators of the IRR of those competence ratings.

Method

We conducted our systematic review in accordance with the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) statement (Moher et al. 2009). The review protocol was pre-registered and published with the *International Prospective Register of Systematic Reviews* (PROSPERO; CRD42018111752).

Inclusion Criteria

Participants in the original studies had to be mental health patients diagnosed via a formal classification system (i.e., any edition of the International Statistical Classification of Diseases and Related Health Problems (ICD; WHO 1992) or the Diagnostic and Statistical Manual of Mental Disorders (DSM; APA 2013). Given that competence scales have been developed mainly for adults (Webb et al. 2010), we concentrated on studies with patients aged 18 and over. To address the performance of a therapist or mental health care provider within a real clinical encounter, we included any studies focusing on individual, face-to-face bona fide psychotherapy (APA 2017). To enable us to focus on psychotherapy and not exclusively on counseling, at least 50% of therapists in the studies were expected to be licensed and to have a minimum of 1 year of any clinical experience. Studies were included if at least two external judges performed the ratings. We allowed for any person to be an external rater (e.g., supervisor, peer, independent researcher) and for any competence scale to be included.

The outcome was the IRR of the total scores of therapeutic competence ratings. IRR refers to the variation between different raters measuring the same subjects under similar conditions (Koo and Li 2016; Kottner et al. 2011; Santelmann et al. 2016). We included IRR as measured by the intraclass correlation coefficient (ICC) since this score is used most frequently for continuous competence outcomes (Kottner et al. 2011), but we also included other IRR coefficients (e.g., Pearson's correlation coefficient). We only included studies reporting the size of the (sub-) sample for calculating the IRR (cf. Trajković et al. 2011) in order to enable proper interpretation.

Empirical original studies published during a peer-reviewed process (e.g., without commentaries or reviews) were considered. There were no restrictions regarding language or publication date.

Search Strategy

The *PubMed* (NCBI; 17th September 2018) and *PsycInfo* (EBSCOhost; 20th September 2018) databases were searched adapting the following search terms to the respective platforms: (mental* OR psych* OR therap*; TI/AB) AND (competenc*; TI/AB) AND (reliability OR ICC; all fields) AND (assessment* OR rater* OR rating*; all fields; humans). We did not exclude grey literature such as dissertations or conference abstracts. Further, we inspected the reference lists of relevant review papers (backward search; Barber et al. 2007; Kazantzis 2003; Muse and McManus

2013; Webb et al. 2010; Zarafonitis-Müller et al. 2014) and finished our search in November 2018.

Screening and Data Extraction

First, titles and abstracts were screened independently for inclusion (TP, RL). Then, full texts were retrieved and screened again independently by two researchers (TP, RL). Disagreements were resolved through discussion or through the inclusion of a third reviewer (FK). Interrater agreement was determined for all of the full texts and amounted to $\kappa = .65$, which reflects good agreement (Higgins and Green 2011). For data extraction, we used a structured form including study, patient, therapy, therapist, rater and rating aspects (e.g., rater training or rating material). The form was piloted by two reviewers (TP, RL) on five publications. After the form was finalized, two master's student reviewers (B.Sc. psych.; TP, RL) extracted all data first, and two licensed psychotherapists (FK, UM) doubled-checked all results independently.

Quality of Reporting

Referring to the *Guidelines for Reporting Reliability and Agreement Studies* (GRRAS; Kottner et al. 2011), in their review, Duffy et al. (2013) proposed a 7-item tool on the reliability of a specific measure of patients' activities of daily living. We adapted their tool to our research question and double-checked with the GRRAS. The final reporting checklist comprised the following aspects: (1) therapist sample (number, recruitment, qualification, experience, psychotherapy approach), (2) rater sample (number, recruitment, qualification, experience, psychotherapy approach), (3) administration of the ratings (rating material), (4) independence of the ratings, (5) rater training (form and amount of training), (6) patient sample (number, recruitment, diagnosis), and (7) blinding of the raters (availability of this information, no supervisors). Each of the seven aspects was rated as follows: 0 = insufficient, 1 = partly sufficient, and 2 = sufficient description in the primary study. Again, two independent raters (TP, RL) assessed the quality of reporting. For the sum scores, and before the resolution of disagreements, the IRR reached $ICC_{(1,2)} = .88$ [CI = .70 – .95], which is considered high (Wirtz and Caspar 2002).

Statistical Analysis

The outcome was the IRR, and different coefficients were reported. The ICC, Finn's *r* and the generalizability coefficient all refer to the same statistical model and were thus combined within one meta-analysis. As data may be combined statistically if at least two coefficients are available, meta-analysis could be conducted only on the ICC (and not on the kappa and Pearson coefficients presented in Table 1) after the unit of analysis was defined. To avoid dependent data, the study, and not the publication, was the unit of analysis (Higgins and Green 2011). If multiple outcomes refer to the same study, the most straightforward procedure to avoid statistical dependency is to include only one outcome per study using pre-defined criteria (Quintana 2015). If multiple publications were based on the same study, we chose the one with the most comprehensive data. The same was true for multiple ICCs reported within one study, then we chose the ICC based on the more comprehensive and purer data. Two studies reported two ICCs for the subscales of the instruments used (Brueck et al. 2009; Wittorf et al. 2013). In that case, we transformed the ICCs to Fisher's *z* values and then used the mean coefficient for further analyses.

If multiple data were available, we gave priority to video (instead of audio) data, as they enable more comprehensive judgments, and to Cognitive Therapy Scale (CTS) data (instead of CTS-R) data, since they were much more common and thus better to combine. As the terms CTS and CTRS are often used interchangeably (Muse and McManus 2013), we decided to use the abbreviation "CTS" for both, also to avoid confusion.

If the study authors did not explicitly report that whole sessions were rated, we documented that the sessions were "probably complete". However, we concluded from most descriptions that whole sessions were performed, so whole sessions comprised the data point then used for meta-analysis. If it was unclear how many sessions were rated per patient (e.g., 1–2), we chose the more conservative value (i.e., 1). Furthermore, if multiple options existed, we decided for expert raters, the higher number of ratings and entire sessions.

We performed a random effects meta-analysis using the restricted maximum likelihood estimator. Correlations were converted into Fisher's *z* values for all analyses and retransformed for interpretation. As the sample, we defined the number of tapes that were rated. We tested for statistical heterogeneity using Cochran's Q and the I^2 statistic (Higgins and Green 2011). A Baujat plot (Baujat et al. 2002) was used to examine potential outliers (Quintana 2015). To test for reporting bias, we used Egger's test (Egger et al. 1997) and visually examined the funnel plots. Following the script by Quintana (2015), we used the "metafor" (Viechtbauer 2010), "robumeta" (Fisher and Tipton 2015) and "dplyr" (Wickham et al. 2015) packages for R (R Core Team 2018).

We independently explored moderators using a series of meta-regression analyses (Quintana 2015). We derived the moderators from the previous literature (Muse and McManus 2013), specifically the number of raters, the quality of reporting, sessions rated per patient, the number of therapists, the number of patients and the study design [randomized-controlled trials (RCT) vs. other], rating material

Table 1 Complete evidence map of studies included in the qualitative summary

Publication	Original study design	Therapy	Stage of therapy	Main mental health problem	No. and method of therapist	Trainee	% of therapists with PhD	Clinical experience therapist years	Rating material	No. of patients
1 **Barber and Crits-Christoph** (1996)[1]	CSS	SE CT	3, 7, 11, 15	MDD	4 SE 4 CT	n/s	n/s	n/s	AUD	33 SE 7 CT
2 Barber et al. (1996)[1]	CSS	SE CT	3	MDD	4 SE n/s CT	Yes	n/s	n/s	AUD	29 SE 7 CT
3 Barber et al. (1997)[2]	RCT	SE(+GDC) CT(+GDC) IDC(+GDC) (GDC only)	Mdn 5 (2–24)	Cocaine Dep	17 SE 8 CT 8 IDC	Yes	Partly	9.9 years SE 10.6 years CT 2-3 years IDC	AUD	30 SE 9 CT 10 IDC
4 Barber et al. (2003)[2]	RCT	SE(+GDC) CT(+GDC) IDC(+GDC) (GDC only)	95% 2–10	Cocaine Dep	12 SE 18 CT 10 IDC	Partly	67 SE 66 CT 0 IDC	9.9 years SE 10.6 years CT > 3 years IDC	AUD	19 SE 88 CT 22 IDC
5 Barber et al. (2004)[2]	RCT	SE(+GDC) CT(+GDC) IDC(+GDC) (GDC only)	Mdn 5-6 (90% 2–11)	Cocaine Dep	13 SE 15 CT 12 IDC	No clinical phase	n/s	11.7 SE 13.4 CT 13.9 IDC	AUD	73 SE 185 CT 142 IDC
6 Dennhag et al. (2012a; 22:6)[2]	RCT	SE CT IDC	Superv. 2, 4 and every 4th Judges 2–11 versus 12-end	Cocaine Dep	12 SE 15 CT 12 IDC	No	67 SE 80CT 0 IDC	11.7 SE 13.4 CT 13.8 IDC	AUD	94 SE 103 CT 98 IDC
7 Dennhag et al. (2012b, 22:4)[2]	RCT	SE CT IDC	2–11 versus 12-end	Cocaine Dep	12 SE 15 CT 12 IDC	No	75 SE 80 CT 0 IDC	11.7 SE 13.4 CT 13.8 IDC	AUD	94 SE 103 CT 98 IDC
8 Blackburn et al. (2001)	OBS	CT	(1–4), (5–8), (9–12)	Depression Anxiety	20 CT	Yes	n/s	Trainees post-qualification course	AUD (13) VID (14)	34
9 Brueck et al. (2009)	RCT	MI	1, 2	Alcohol Dep	12	No	67%	n/s	AUD	16
10 Chevron and Rounsaville (1983)[3]	OBS	IPT	1, 6, 11	Recurrent MDD	9 IPT	Yes	n/s	min. 2	VID	13
11 Dobson et al. (1985)[3]	OBS	CT	1 or ongoing	Depression	21 CT	Yes	100%	min. 2	VID	21

Table 1 (continued)

Publication	Original study design	Therapy	Stage of therapy	Main mental health problem	No. and method of therapist	Trainee	% of therapists with PhD	Clinical experience therapist years	Rating material	No. of patients
12 Vallis et al. (1986)[3]	OBS	CT	n/s	Depression	9 CT	Yes	Partly	n/s	VID	n/s
13 Vallis et al. (1988)[3]	OBS	CT	n/s	Depression	8 CT	Yes	n/s	mean 8,5 years	VID	20
14 Dittmann et al. (2017)	RCT	CPT	5-end	PTSD BPS	5 licenced 2 CBT trainees	Partly	n/s	n/s	VID	8
15 Hoffart et al. (2005)	CCT	CT versus Schema	3	Panic agoraph. cluster C	2 licenced CPs	No	Min. 1	n/s	VID	12
16 Karterud et al. (2013)	CSS	Mentalization	n/s	BPS	9 clinicians, group analysis	No	n/s	n/s	VID	n/s
17 Kazantzis et al. (2018)	RCT	CBT	3, 15	MDD	4 CBT	No	n/s	9.5 years CBT	AUD	50
18 Kuyken and Tsivrikos (2009)	CSS	CBT	n/s	MDD	18 CBT	No	n/s	1–10 years	Individual evaluation	69
19 McGrath (2013)	RCT	ACT versus CBT	1–6, 7–12	Anxiety	32 CBT + ACT	Yes	Partly	1–2 years	AUD	88
20 Reichelt et al. (2003)	CSS	CT	n/s	n/s	24 CT	Yes	n/s	Post-qualification course trainees	VID	n/s
21 Schmidt et al. (2018) [4] Study 1	RCT	CT	2–24	MDD	6 CT	No	5/6	2–21	n/s	6
Study 2	CSS	CT	n/s	n/s	14 CT	Partly	7	0.6 years	AUD	14
22 Strunk et al. (2010)[4]	RCT	CT	1–4	MDD	6 CT	No	5/6	2–21	AUD + VID	60
23 Soygüt et al. (2008)	CSS	CT	Any	Anxiety mood	7 CT	Yes	0	CP graduate students	VID	10
24 Svartberg (1989)	CSS	PD	4-end (every 2nd/3rd)	SP, panic OCD, MDD	5 PD	Partly	n/s	4 years postgrad. experience	AUD + VID	8
25 Tadic et al. (2003)	CSS	PD	1	Anxiety mood personality disorder	7 PD	No	n/s	> 5 years	VID + transcript	16

Table 1 (continued)

Publication	Original study design	Therapy	Stage of therapy	Main mental health problem	No. and method of therapist	Trainee	% of therapists with PhD	Clinical experience therapist years	Rating material	No. of patients
26 Wittorf et al. (2013)	RCT	CBT	34×1–8 45×9–20	Psychosis	7 n/s	Partly	2/7	3.75	VID	79
27 von Consbruch et al. (2012)[5]	RCT	CT	1–3, 10–12, 20–22	SP	51 CT	n/s	n/s	n/s	VID	98
28 Weck, Bohn et al. (2011)[5]	RCT	CT	Any	SP	10 CT	n/s	n/s	4,8 years	VID	34
29 Weck, Hautzinger et al. (2011)[6]	RCT	MAPE	Any	Recurrent MDD	18 n/s	n/s	n/s	1.2 years	VID	30
30 Weck, Weigel et al. (2011)[6]					19 n/s			1.3 years		
31 Weck, Hilling et al. (2011)[6]	RCT	CBMT	Any	Recurrent MDD	19 n/s	n/s	n/s	2.3 years	VID	30
32 Weck et al. (2014)	RCT	CBT	Any	MDD, SP, hypochondriasis	50 CBT	Partly	n/s	2.9 years	VID	84

Publication	No. of tapes	No. of sessions rated per patient	Length rated tapes	No. of raters	Background rater experience, profession	Rater training	Independence raters	Reporting quality	Competence Scales	Outcome
1 **Barber and Crits-Christoph** (1996)[1]	84 SE 7 CT	1-4	Probably COMPL	2	PhD/CPs 1 SE expert	No	Yes	9 (6)	PACS-SE	**ICC $_{(2,2)}$ (total competence) .42 SE** .73 SE+CT
2 Barber et al. (1996)[1]	29 SE 7 CT	1	COMPL	2	PhD/CPs 1 SE expert	No	Yes	9 (6)	PACS-SE	ICC $_{(n/s)}$ (general skills subscale) .77 SE+CT
3 Barber et al. (1997)[2]	32 SE 10 CT 10 IDC	n/s	Probably COMPL	2	PhD/CPs 1 SE expert	No	Yes	5 (4)	ACS-SEC	ICC $_{(2,2)}$ (total score) entire sample .43 SE subsample .33

Table 1 (continued)

Publication	No. of tapes	No. of sessions rated per patient	Length rated tapes	No. of raters	Background rater experience, profession	Rater training	Independence raters	Reporting quality	Competence Scales	Outcome
4 Barber et al. (2003)[2]	20 SE 92 CT 22 IDC	n/s	Probably COMPL	2	CTs with 5 years training	Yes	Yes	12 (6)	CTACS	**ICC $_{(2,2)}$ (total score)** **entire sample .80** CT subsample .73
5 Barber et al. (2004)[2]	302 SE 273 CT 307 IDC	1–2	Probably COMPL	2	SE, CT, IDC experts	n/s	Yes	9 (5)	CTACS ACS-SEC IDCCD-SEC	ICC $_{(2,2)}$ (mean scores) .81 SE .94 CT .74 IDC
6 Dennhag et al. (2012a; 22:6)[2]	Supervisors 436 SE 518 CT 396 IDC	Supervisors 4.6 SE 5.0 CT 4.1 IDC	Probably COMPL	Superv. 3 SE 5 CT 4 IDC	Experts in their fields	n/s	No	11 (6)	ACS-SEC CTACS ACS-IDCCD	ICC $_{(Var, T)}$; MLM; Supervisors .25 SE .24 CT .74 IDC
	Judges 148 SE 192 CT 181 IDC	Judges 1.6 SE 1.9 CT 1.8 IDC		Judges 3 SE 2 CT 2 IDC			Yes			ICC $_{(Var, T)}$; MLM; Judges .20 SE .41 CT .17 IDC
7 Dennhag et al. (2012b, 22:4)[2]	148 SE 192 CT 181 IDC	1.6 SE 1.9 CT 1.8 IDC	Probably COMPL	3 SE 2 CT 2 IDC	Experts in their fields	n/s	Yes	12 (6)	ACS-SEC CTACS ACS-IDCCD	ICC $_{(2,2)}$ .68 SE (cf. Barber 2008) .73 CT (cf. Barber 2003)
8 Blackburn et al. (2001)	102	3	Probably COMPL	2 (ICC) 4 (r)	CT experts	n/s	Yes	9 (6)	CTS-R (13 items) CTS-R (14 items)	**ICC (2-way ANOVA)** 0.63 (13) **0.57 (14)** $r_{Pearson}$ .66 (13) .63 (14)
9 Brueck et al. (2009)	28	2	Last 20 min	3	2 students, 1 expert	Yes	Yes	12 (6)	MITI-d	**ICC $_{(n/s)}$** **.42 (Spirit)** **.56 (Empathy)**
10 Chevron and Rounsaville (1983)[3]	27	3	COMPL	2	Supervisors of other trainees	n/s	Yes	9 (6)	TSRF	Pearson's r .88

Table 1 (continued)

Publication	No. of tapes	No. of sessions rated per patient	Length rated tapes	No. of raters	Background rater experience, profession	Rater training	Independence raters	Reporting quality	Competence Scales	Outcome
11 Dobson et al. (1985)[3]	21	1	COMPL (50-60)	2 of 4	CT experience	Yes	Yes	10 (6)	CTS (11 items)	**Inter-rater correlation (total score) .94**
12 Vallis et al. (1986)[3]	10	1-2	COMPL (50)	1-2	PhD, M.D, CT experts	n/s	n/s	6 (5)	CTS (11 items)	ICC $_{(n/s)}$ (total score) .77
13 Vallis et al. (1988)[3]	20	1	3 * 5-min. sections	2	1 CT expert, 1 research assistant	n/s	Yes	7 (5)	MCTB	Pearson's r .85
14 Dittmann et al. (2017)	30	2–4	COMPL	2	Licensed CBT CPT experience	Yes	Yes	11 (6)	CRS-PTSD CRS-CPT CTS	**ICC $_{(2,2)}$** .97 CRS-PTSD .97 CRS-CPT **.97 CTS (sum score, 11 items)**
15 Hoffart et al. (2005)	12	1	COMPL	2	CT experts	No	Yes	8 (5)	CTS (11 items)	**ICC$_{(3,1)}$ sum score .93**
16 Karterud et al. (2013)	18	2 per therapist (1 high versus. 1 low quality)	Probably COMPL	2 versus 7	MBT clinicians and researchers	Yes	n/s	8 (5)	MBT-ACS	**generalizability coefficient** 2: .68 **7: .88**
17 Kazantzis et al. (2018)	92	2	Probably COMPL	2	Psychology graduates	Yes	Yes	12 (7)	CTS (11 items) CTS-R (12 items)	**Finn's r (total scores) CTS: .93** 3: .89 15: .83 CTS-R: .88 3: .81 15: .85
18 Kuyken and Tsivrikos (2009)	n/s	n/s	n/s	2	1 director 1 expert therapist	No	No	14 (7)	ETBF	kappa .80
19 McGrath (2013)	35	2 random	COMPL (50 min)	2	Doctoral students	Yes	Yes	12 (6)	DUACRS-R	**ICC $_{(2,2)}$ (competence subscale) .86**
20 Reichelt et al. (2003)	48	n/s	COMPL (60 min)	2	Supervisors	Yes	Yes	12 (7)	CTS-R (12 items)	Pearson's r Pre-training .44 Post-training .67

Table 1 (continued)

Publication	No. of tapes	No. of sessions rated per patient	Length rated tapes	No. of raters	Background rater experience, profession	Rater training	Independence raters	Reporting quality	Competence Scales	Outcome
21 Schmidt et al. (2018) [4] Study 1	6	1	Probably COMPL	1-2	CT experts	n/s	Yes	7 (5)	CTS (11 items) + STB	ICC (2 raters, total score) 0.89
Study 2	14	1	Probably COMPL	1-2	1 licensed psychologist2 graduate students	Yes	Yes		CTS (11 items)	**ICC (2 raters, total score) 0.68**
22 Strunk et al. (2010)[4]	240	4	Probably COMPL	2	Psychologists, 1 year practicum	Yes	No	13 (7)	CTS (11 items)	**ICC (2 raters, total score) 0.77**
23 Soygüt et al. (2008)	20	2	Probably COMPL	2	2 PhD/CT experts	Yes	Yes	7 (6)	CTACS	**ICC (2 raters, mean value) 0.60**
24 Svartberg (1989)	31	Mdn 3.5	Probably COMPL	2	Experienced psychiatrists	Yes	Yes	11 (7)	STCRF	**$ICC_{(1,k)}$ .67 (video)** .76 (audio)
25 Tadic et al. (2003)	16	1	Probably COMPL	2	n/s	n/s	No	10 (5)	ICS	**$ICC_{(n/s)}$ mean 0.71**
26 Wittorf et al. (2013)	79	1 random	COMPL (50 min)	2	Trained CPs, CBT-psychosis	Yes	Yes	10 (6)	CTS-Psy	**$ICC_{(n/s)}$ General skills: .86 Technical skills: .96**
27 von Consbruch et al. (2012)[5]	161	1-2 random	COMPL (50-100)	7	6 CPs in training 1 licensed	Yes	n/s	8 (5)	CTCS-SP	**$ICC_{(2,2)}$ (total score) 0.81 (rater A versus all others)**
28 Weck, Bohn et al. (2011)[5]	34	1 random	COMPL (50 min) & 20 min segments 1-3	2 Seg + 2 Entire	3 PhD students/ CT-Trainees, 1 PhD, all with 2-7 years CT experience	Yes	Yes	9 (6)	CTCS- SP	**$ICC_{(2,2)}$ mean .81 (entire)** .84 (seg. 1) .71 (seg. 2) .60 (seg. 3)
29 Weck, Hautzinger et al. (2011)[6]	30	1	COMPL (20 min)	2	1 in training (2 years experience), 1 licensed therapist (7 years experience)	Yes	Yes	10 (6) 11 (7)	CS-P	$ICC_{(2,2)}$ (mean whole scale) .87

Table 1 (continued)

Publication	No. of tapes	No. of sessions rated per patient	Length rated tapes	No. of raters	Background rater experience, profession	Rater training	Independence raters	Reporting quality	Competence Scales	Outcome
30 Weck, Weigel et al. (2011)[6]				4	2 experts (2 and 7 years experience), 2 novices (no experience)					$ICC_{(2,2)}$ (mean whole scale) expert–expert: .87 novice–novice: .89 expert-novice: .86
31 Weck, Hilling et al. (2011)[6]	30	1	Probably COMPL	4	2 experts, 2 psych. undergraduates	Yes	n/s	8 (5)	CTS (14 items)	**$ICC_{(2,2)}$ (mean score)** **expert–expert: .90** novice–novice: .80 expert-novice: .68
32 Weck et al. (2014)	84	1 random	COMPL versus middle segments	4	3 licensed therapists, 1 end of training; 6.3 years experience	Yes	n/s	9 (6)	CTS (14 items)	$ICC_{(1,2)}$ middle segment items 2–6, 8-13: 0.72 entire session items 2–6, 8-13: 0.76 **items 1–14: 0.77**

Superscripts [1–6] publications with one superscript refer to one main study, printed in bold… included in the meta-analyses

Reporting quality from 0 to 14, higher values indicate better reporting (value in brackets as described by Duffy et al. (2013), assuming high quality of reporting if ≥ 5 of 7 reporting criteria were described sufficiently)

n/s not specified, *CSS* cross-sectional study, *RCT* randomized controlled trial, *OBS* observational study, *CCT* controlled clinical trial, *ICC* intraclass correlation coefficient, Independent raters means independent of each other, not blinded to outcome or not knowing the therapists etc., *SE* supportive expressive, *CT* cognitive therapy, *IDC* individual drug counseling, *GDC* group drug counseling, *MI* motivational interviewing, *PT* psychotherapy, *PD* psychodynamic psychotherapy, *IPT* interpersonal psychotherapy, *CPT* cognitive processing therapy, *CBT* cognitive behavior therapy, *ACT* acceptance commitment therapy, *MAPE* manualized active psychoeducation, *CBMT* cognitive behavioral maintenance therapy, *Mdn* median, *MDD* major depressive disorder, *Cocaine Dep* cocaine dependence, *Alcohol Dep* alcohol dependence, *PTSD* post-traumatic stress disorder, *BPS* borderline personality disorder, *Agoraph* agoraphobia, *Panic* panic disorder, *Cluster C* cluster C personality disorder, *Mood* mood disorder, *SP* social phobia, *OCD* obsessive–compulsive disorder, *AUD* audiotape, *VID* videotape, *COMPL* only complete sessions if explicitly stated, *Seg* segment, *PhD* doctoral degree, *CP* clinical psychologist, *CT* cognitive therapist, *MD* Medical Doctor, *MBT* mentalization based therapist, *psych* psychology, *PACS-SE* Penn Adherence/Competence Scale for Supportive-Expressive Psychotherapy, *ACS-SEC* Adherence/Competence Scale for SE Cocaine Dependence, *CTACS* Cognitive Therapy Adherence and Competence Scale, *ACS-IDCCD* Adherence/Competence Scale for IDC for Cocaine Dependence, *CTS-R* Cognitive Therapy Scale Revised, *MITI-d* Motivational Interviewing Treatment Integrity Code, *TSRF* Therapist Strategy Rating Form, *CTS* Cognitive Therapy Scale, *MCTB* Matarazzo Checklist of Therapist Behavior, *CRS-PTSD* Competence Rating Scale for PTSD, *CRS-CPT* Competence Rating Scale for CPT, *MBT-ACS* Mentalization-Based Treatment Adherence and Competence Scale, *ETBF* Evaluation of Therapist's Behavior Form, *DUACRS-R* Drexel University ACT/tCBT Adherence and Competence Rating Scale, *STB* list of specific therapist behavior, *STCRF* Short-term Anxiety-provoking Therapy - Therapist Competence Rating Form, *ICS* Investigation Competence Scale, *CTS-Psy* Cognitive Therapy Scale for Psychosis, *CTCS-SP* Cognitive Therapy Competence Scale for Social Phobia, adapted from CTS, *CS-P* Competence Scale for Psychoeducation, *MLM* (ICC calculated within a multilevel model, *CI* confidence interval

(audio vs. video/both), rater training (yes vs. no/not specified), independence of raters (yes vs. no/not specified), the scale used for the ratings (CTS-based vs. other), therapy (CBT-related vs. other), therapist trainees (yes vs. partly/no/not specified) and patients' diagnosis (depression & anxiety vs. other).

Results

Characteristics of Included Studies

Through our literature search, we identified 1286 records. After we removed duplicates and added records from the reference lists of the included reviews, we screened 908 for their title and abstract. We finally included 20 studies reported in 32 publications in the narrative synthesis. The study flow chart and reasons for exclusion are illustrated in Fig. 1. A detailed description of the reasons for inclusion and exclusion into the statistical analysis are presented in Supplement 1. The 20 IRRs that could be quantitatively combined for quantitatively are highlighted in bold in the evidence map (Table 1), which also illustrates further information. Since one publication (Schmidt et al. 2018) reported two studies and another referred to two samples (Dennhag et al. 2012a), the total numbers and percentages may vary within the following narrative synthesis.

The included studies were conducted between 1983 (Chevron and Rounsaville 1983) and 2018 (Kazantzis et al. 2018; Schmidt et al. 2018), and 17 of the original studies were RCTs. Most studies focused on cognitive therapy (CT), CBT, comparisons with so-called third-wave interventions (Hoffart et al. 2005; McGrath 2013), psychoeducation (Weck, Hautzinger, et al. 2011; Weck, Weigel, et al. 2011), maintenance treatment (Weck, Hilling, et al. 2011) or a CBT-related intervention (motivational interviewing; Brueck et al. 2009). The minority of studies addressed psychodynamic therapy (Svartberg 1989; Tadic et al. 2003) with related interventions such as mentalisation-based treatment (Karterud et al. 2013) or interpersonal therapy (Chevron and Rounsaville 1983). In contrast, the seven publications by Barber et al. (see Table 1, superscripts 1 and 2) compared cognitive and psychodynamic therapy with counseling as well as individual versus group interventions. Most patients included in the studies were diagnosed with depression ($n = 12$, 37.5%), substance dependence ($n = 6$, 18.75%), anxiety and depression ($n = 2$, 6.24%), anxiety alone ($n = 3$, 9.38%), or other diagnoses ($n = 7$, 21.89%), or no diagnosis was specified ($n = 2$, 6.24%). The number of included patients ranged

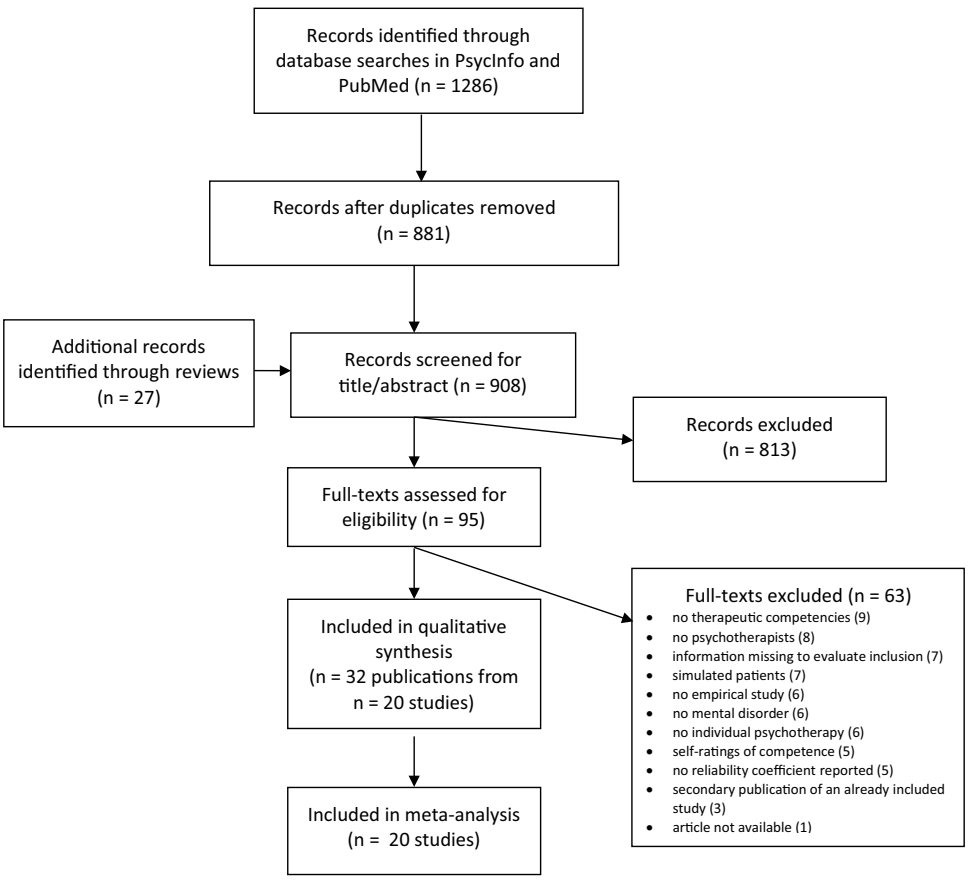

Fig. 1 PRISMA flow chart of study inclusion

from 6 (Schmidt et al. 2018, study 1) to 400 (Barber et al. 2004).

The study therapists were licensed ($n = 11$, 33.33%), in training ($n = 10$, 30.31%), or both ($n = 6$, 18.18%), or their qualification was not described in detail ($n = 6$, 18.18%). The number of therapists ranged from 5 (Svartberg 1989) to 51 (von Consbruch et al. 2012). In 16 publications, the ratings were based on video tapes (50%), in 11 on audio tapes (34.38%), in three on both (9.37%), and in two on other sources (6.25%). The number of tapes that were rated ranged from 10 (Vallis et al. 1986) to several hundred (see Table 1; Dennhag et al. 2012a). One to four sessions were rated per patient, whereas ratings were mostly ($n = 12$, 36.36%) based on one session and were assessed by two raters ($n = 23$, 63.88%). In most cases ($n = 18$, 52.25%), raters were trained; in five cases (15.63%), they received no training; and in nine publications (28.12%), this aspect was not specified. The raters were mostly ($n = 24$, 72.72%) described as independent of each other, whereas sometimes, they were not independent ($n = 4$, 12.12%) or this facet was not specified ($n = 5$, 15.16%).

The quality of reporting of respective studies was above average (i.e., 8–14 points) in 27 publications (84.38%) and below average (≤ 7 points) in five of them (15.62%). In contrast, using the dichotomous scaling (i.e., either 0 or 1) proposed by Duffy et al. (2013), the quality of reporting was rated as "sufficient" (i.e., scores ≥ 5) in $n = 31$ (96.88%) of the studies (see Table 1).

Most often ($n = 16$, 50%), the CTS or CTS-based instruments were used for assessing competence. As an IRR coefficient, most often ($n = 27$, 79.41%), the authors calculated different forms of the ICC. Less often, the generalizability coefficient (Karterud et al. 2013), Pearson's r (Chevron and Rounsaville 1983; Vallis et al. 1988), Finn's r (Kazantzis et al. 2018), the kappa coefficient (Kuyken and Tsivrikos 2009) or so-called inter-rater correlation (Dobson et al. 1985) were used.

Quantitative Synthesis

We conducted a meta-analysis of 20 publications referring to a total sample of $n = 1272$ tapes. The summary correlation was ICC = 0.82 [95% CI (0.74, 0.87), $p < 0.001$], which, at first glance, could be interpreted as appropriate ($\geq .70$; Wirtz 2017) or good IRR ($\geq .75$; Portney and Watkins 2009; Fig. 2). Still, statistical heterogeneity was considerable ($I^2 = 90.39\%$; $Q = 163.06$, $p < .0001$; Higgins and Green 2011). According to the Baujat plot (Supplement 2), the studies by Barber and Crits-Christoph (1996, study 1) and by Dittmann et al. (2017, study 6) were potential outliers. Although these were the studies with the lowest (ICC = 0.42; Barber and Crits-Christoph 1996) and the highest (ICC = 0.97; Dittmann et al. 2017) IRRs, a meta-analysis without the two of them changed the results only marginally.

Visual examination of the funnel plot (Fig. 3), which illustrated symmetry, yielded no indication for publication bias. Accordingly, Egger's test for publication bias was not

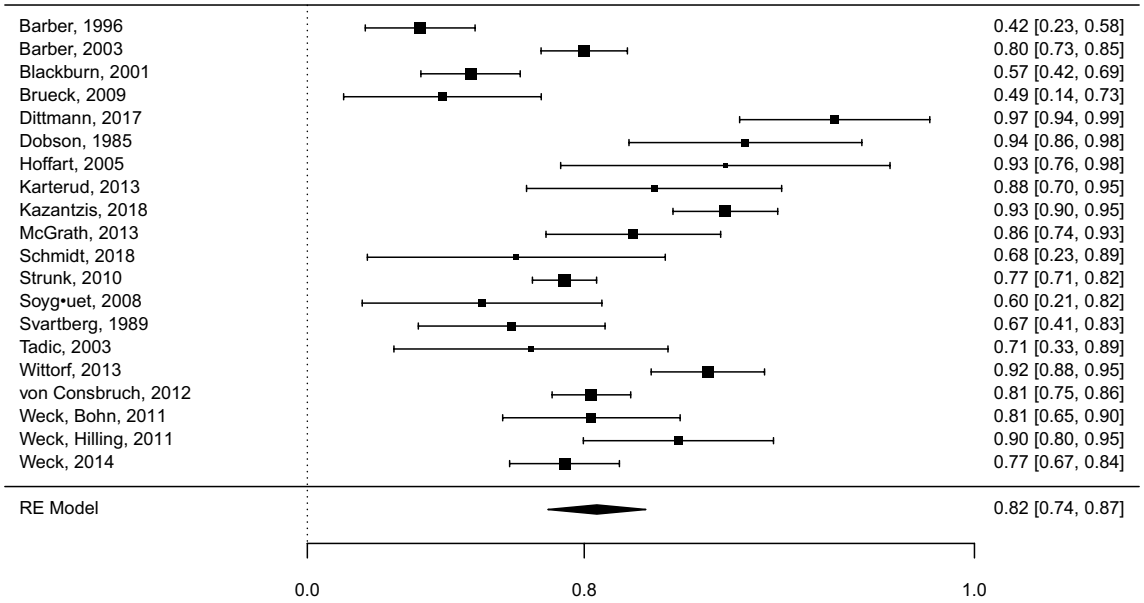

Fig. 2 Forest plot of the average interrater reliability (ICCs with CIs)

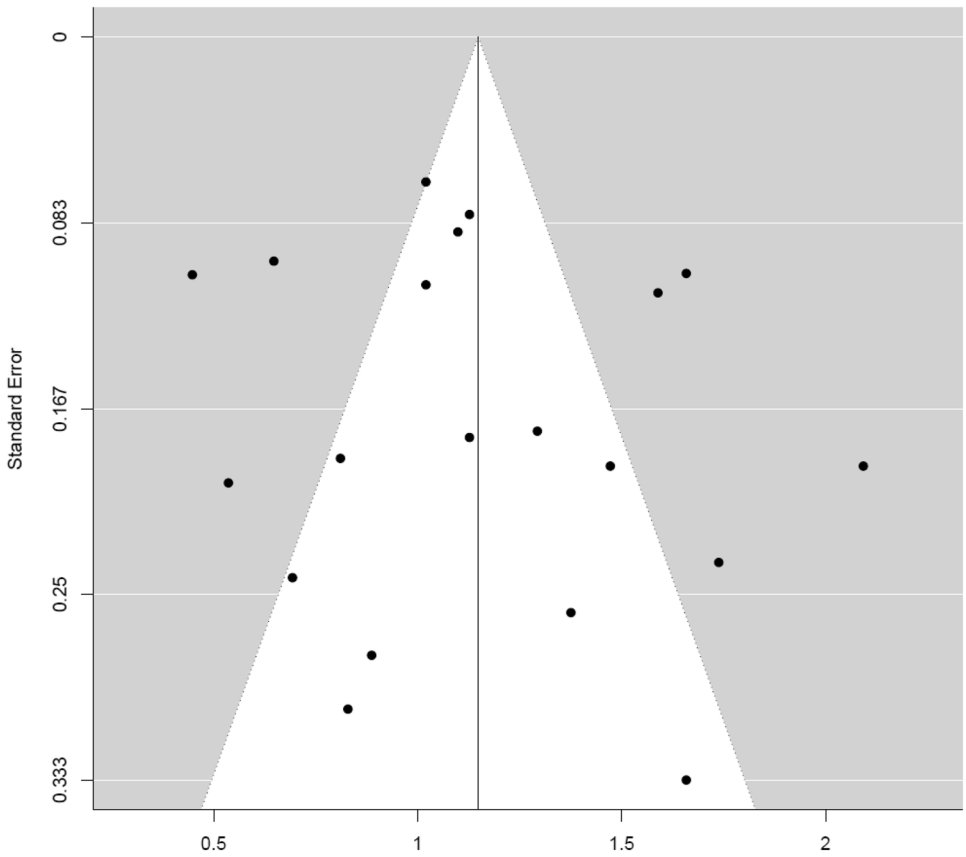

Fig. 3 Funnel plot

significant ($p = 0.56$). However, only 65% (n = 13, instead of the expected 95%) of studies lay within the triangular region of the funnel plot, which clearly indicates heterogeneity again (Higgins and Green 2011).

The Role of Moderators

None of the investigated variables had an individual moderating effect, that is, number of raters [$Q(1) = 0.06$; $p = 0.80$], quality of reporting [$Q(1) = 0.75$; $p = 0.39$], sessions rated per patient [$Q(1) = 0.77$; $p = 0.38$], number of therapists [$Q(1) = 0.04$; $p = 0.84$], number of patients [$Q(1) = 0.05$; $p = 0.82$], study design [$Q(1) = 2.59$; $p = 0.11$], form of therapy [$Q(2) = 0.89$; $p = 0.35$], therapist trainees [$Q(2) = 0.09$; $p = 0.77$], rating material [$Q(1) = 1.11$; $p = 0.29$], independence of raters [$Q(1) = 0.0005$; $p = 0.98$] and patients' diagnosis [$Q(4) = 0.82$; $p = 0.37$]. Two variables, namely, rater training [$Q(1) = 2.96$; $p = 0.09$] and scale used for the ratings [$Q(1) = 3.59$; $p = 0.06$] had a p value of $< .1$.

Discussion

To the best of our knowledge, this is the first evidence synthesis on the reliability of psychotherapeutic competence ratings. The aims of this study were to provide a map of the current evidence, to estimate a pooled IRR, and to investigate moderators of the IRR of psychotherapeutic competence ratings.

In their narrative review, Muse and McManus (2013) reported ICCs for total CTS scores between 0.01 (no agreement) and .94 (nearly perfect agreement), which left uncertainty regarding the ability to rate psychotherapeutic competence. Our meta-analysis revealed a pooled ICC of 0.82 indicative of appropriate reliability, and since both aspects are related to each other, severe heterogeneity (Wirtz and Caspar 2002). Coefficients ranged from ICC = 0.42 (Barber and Crits-Christoph 1996) to ICC = 0.97 (Dittmann et al. 2017). Although these values might be attributable to the file drawer problem (Higgins and Green 2011), that is, the paucity of published studies showing small or no reliability, our results did not support publication bias. Nonetheless, the majority of study authors adhered to basic principles to improve the reliability of ratings, i.e., training raters or using video tapes to maximize the available information (Muse and McManus 2013).

Our qualitative synthesis revealed an evidence map more detailed (Dennhag et al. 2012b) and systematic (Muse and McManus 2013) than the overviews given by previous reviews. Not surprisingly, it showed that most empirical studies referred to CBT and to patients diagnosed with depression. Consequently, the CTS was used most often, as it

was particularly developed for CBT in the context of depression (e.g., Vallis et al. 1988). Although criticized for its specific focus, it is now also used within other diagnoses, such as psychosis, anxiety or personality disorders (Muse and McManus 2013). In addition, other comprehensive measures (e.g., Muse et al. 2017) or treatment-specific instruments (e.g., Machmutow et al. 2018) have been published but are still less commonly used than the CTS. Another perspective may be to successively improve established procedures.

According to our results, the number of tapes that were used ranged from ten to several hundred per study, and ratings were mostly based on a single session. In contrast, Dennhag et al. (2012b) show that, for example, for CT, three patients per therapist and four sessions per patient would be necessary to achieve appropriate reliability, which is far above the actual number. However, since competence ratings by trained raters are rather cost intensive, resource constraints may play a major role (Muse and McManus 2013).

Whereas older studies used Pearson correlation coefficients not controlling for varying variances between raters (Wirtz 2017), the ICC has become the most prevalent reliability measure. In their current publication, Kazantzis et al. (2018) proposed using Finn's *r* as a potentially useful alternative to some ICC, if data are markedly non-normal and there is a restricted number of categories (e.g., if a 7-point scale exists but raters tend to use four options).

Although these results raise confidence in the utility of competence scales, there are still unanswered research questions. Addressing these issues, and thus improving established procedures, may contribute to less clinical and methodological diversity of primary studies, and thereby enhance statistical pooling in the future (Higgins and Green 2011). For example, raters were often described as independent of each other, but authors varied in their explanations of how this independence was achieved, with studies reporting more (Dennhag et al. 2012a) or less detailed information (Kuyken and Tsivrikos 2009). One strategy to enhance rater independence is to view video tapes and give evaluations separately. Another is to view videos and discuss ratings in intervals in order to reduce rater drift, which refers to changing rating criteria over time (Warshaw et al. 2001). Apart from rater drift, other judgment and observational biases (Wirtz 2017) have rarely been investigated in the competence literature thus far—another possible focus of future research.

Furthermore, the amount of rater expertise necessary still remains an empirical question, with some arguing for more experienced raters and others arguing that, presuming the provision of adequate training, novice raters may also provide reliable ratings (Muse and McManus 2016; Weck, Weigel et al., 2011). Furthermore, the study purpose guides the choice of raters, that is, choosing supervisors if broader knowledge about therapists is necessary or independent judges if objectivity is to be maximized (Muse and McManus 2013).

Although no moderators proved significant in our first exploration of moderators, this finding does not indicate their unimportance; moderator analyses require larger samples, especially if studies with varying quality are included (Hempel et al. 2013). The same applies for the fact that nine publications included small samples of ≤ 30 tapes. We only conducted univariate meta-regression analyses due to power considerations, and thus could not simultaneously control for other variables (Meister et al. 2017). Other limitations of our meta-analysis could be the inclusion of rather experienced therapists and a subsample of 20 studies for quantitative synthesis. Combining comparable coefficients for meta-analysis was important to reduce statistical dependency among the coefficients (Quintana 2015).

Despite this strategy, there was considerable between-study heterogeneity, limiting the interpretability of our results. First of all, heterogeneity might be attributable to conceptual differences, as psychotherapeutic competence was defined in different ways in the primary studies. Accordingly, it may be ascribed to differences in the methods used in the original studies, which was evidenced by the fact that only about half of the original studies were RCTs, by the diversity in the quality of reporting, and by the diverse numbers of tapes, patients and therapists included. Adherence and competence ratings are often a by-product of clinical trials. Presumably, researchers do invest in basic strategies to ensure reliable ratings to support the main trials but may not be acquainted with the pitfalls and details accompanying proper competence ratings. Therefore, referring to important standards for rater training, such as clarification of raters' implicit concepts, supervisor feedback, discussion of disagreements, discussion of (a)typical cases or the provision of category definitions (Wirtz 2017), as well as publishing manuals for rater training, and using reporting guidelines (Kottner et al. 2011) will further contribute to advancements in this field of study.

In conclusion, the current meta-analysis indicates first pooled results on the reliability of competence ratings, and highlights considerable heterogeneity within the data. In contrast, meta-analyses are restricted by the results published within primary research (Borenstein et al. 2009), which is why further experimental studies could extend the current results and directly compare relevant competence variables (e.g., contrasting ratings obtained via the CTS, the CTS-R or another instrument). Future studies could further investigate the validity of competence ratings to determine, for example, how to maximize validity (e.g., in relation to a grade received after psychotherapy training or in relation to patient-related outcomes). It remains a vital part of process research to determine the specific bodies of knowledge,

skills and attitudes that constitute an individually competent psychotherapist.

Acknowledgements We would like to thank, Ricarda Löscher B.Sc. psych., for her assistance with study screening and data extraction.

Funding No external funding.

Compliance with Ethical Standards

Conflict of interest Florian Weck is an author of six of the publications included in the review. Franziska Kühne, Ramona Meister, Ulrike Maaß and Tatjana Paunov declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies involving human participants.

Research Involving Animal Rights This article does not contain any studies with animals.

References

References marked with an asterisk indicate studies included in the qualitative summary

American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington VA: American Psychiatric Association.

American Psychiatric Association (APA). (2017). What is psychotherapy? Retrieved from https://www.apa.org/ptsd-guideline/patients-and-families/psychotherapy.

* Barber, J. P., & Crits-Christoph, P. (1996). Development of a therapist adherence/competence rating scale for supportive-expressive dynamic psychotherapy: A preliminary report. *Psychotherapy Research, 6*(2), 81–94. https://doi.org/10.1080/10503309612331331608.

* Barber, J. P., Crits-Christoph, P., & Luborsky, L. (1996). Effects of therapist adherence and competence on patient outcome in brief dynamic therapy. *Journal of Consulting and Clinical Psychology, 64*(3), 619–622. https://doi.org/10.1037/0022-006x.64.3.619.

* Barber, J. P., Foltz, C., Crits-Christoph, P., & Chittams, J. (2004). Therapists' adherence and competence and treatment discrimination in the NIDA Collaborative Cocaine Treatment Study. *Journal of Clinical Psychology, 60*(1), 29–41. https://doi.org/10.1002/jclp.10186

* Barber, J. P., Krakauer, I., Calvo, N., & Badgio, P. C. (1997). Measuring adherence and competence of dynamic therapists in the treatment of cocaine dependence. *Journal of Psychotherapy Practice & Research, 6*(1), 12–24.

* Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the cognitive therapy adherence and competence scale. *Psychotherapy Research*, 13(2), 205–221.

Barber, J. P., Sharpless, B. A., Klostermann, S., & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology: Research and Practice, 38,* 493–500. https://doi.org/10.1037/0735-7028.38.5.493.

Baujat, B., Mahé, C., Pignon, J. P., & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine, 21*(18), 2641–2652.

Beck Institute for Cognitive Behavior Therapy. (2019, October 10). Cognitive Therapy Rating Scale (CTRS). Retrieved from https://beckinstitute.org/wp-content/uploads/2017/03/CTRS-Scale-and-Score-Report-2016.pdf.

* Blackburn, I. M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale (CTS-R): Psychometric properties. *Behavioural and cognitive psychotherapy*, 29(4), 431–446.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken: Wiley.

* Brueck, R. K., Frick, K., Loessl, B., Kriston, L., Schondelmaier, S., Go, C.,…, Berner, M. (2009). Psychometric properties of the German version of the motivational interviewing treatment integrity code. *Journal of Substance Abuse Treatment, 36*(1), 44–48. https://doi.org/10.1016/j.jsat.2008.04.004.

* Chevron, E. S., & Rounsaville, B. J. (1983). Evaluating the clinical skills of psychotherapists: A comparison of techniques. *Archives of General Psychiatry, 40*(10), 1129–1132. https://doi.org/10.1001/archpsyc.1983.01790090091014.

* Dennhag, I., Gibbons, M. B. C., Barber, J. P., Gallop, R., & Crits-Christoph, P. (2012a). Do supervisors and independent judges agree on evaluations of therapist adherence and competence in the treatment of cocaine dependence? *Psychotherapy Research, 22*(6), 720–730. https://doi.org/10.1080/10503307.2012.716528.

* Dennhag, I., Gibbons, M. B. C., Barber, J. P., Gallop, R., & Crits-Christoph, P. (2012b). How many treatment sessions and patients are needed to create a stable score of adherence and competence in the treatment of cocaine dependence? *Psychotherapy Research, 22*(4), 475–488. https://doi.org/10.1080/10503307.2012.674790.

* Dittmann, C., Müller-Engelmann, M., Stangier, U., Priebe, K., Fydrich, T., Görg, N., …, Steil, R. (2017). Disorder- and treatment-specific therapeutic competence scales for posttraumatic stress disorder intervention: Development and psychometric properties. *Journal of Traumatic Stress*. https://doi.org/10.1002/jts.22236.

* Dobson, K. S., Shaw, B. F., & Vallis, T. M. (1985). Reliability of a measure of the quality of cognitive therapy. *British Journal of Clinical Psychology, 24*(4), 295–300. https://doi.org/10.1111/j.2044-8260.1985.tb00662.x.

Duffy, L., Gajree, S., Langhorne, P., Stott, D. J., & Quinn, T. J. (2013). Reliability (inter-rater agreement) of the Barthel Index for assessment of stroke survivors: Systematic review and meta-analysis. *Stroke, 44*(2), 462–468. https://doi.org/10.1161/strokeaha.112.678615.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal Open, 315,* 629–634.

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy, 49,* 373–378. https://doi.org/10.1016/j.brat.2011.03.005.

Fisher, Z., & Tipton, E. (2015). robumeta: An R-package for robust variance estimation in meta-analysis. http://arxiv.org/abs/1503.02220.

Hempel, S., Miles, J. N., Booth, M. J., Wang, Z., Morton, S. C., & Shekelle, P. G. (2013). Risk of bias: A simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic Reviews, 2,* 107. https://doi.org/10.1186/2046-4053-2-107.

Higgins, J., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. 0 [updated March 2011]. The Cochrane Collaboration, 2011. Retrieved August, 29 from www.cochrane-handbook.org.

* Hoffart, A., Sexton, H., Nordahl, H. M., & Stiles, T. C. (2005). Connection between patient and therapist and therapist's competence in schema-focused therapy of personality

problems. *Psychotherapy Research, 15*(4), 409–419. https://doi.org/10.1080/10503300500091702.

* Karterud, S., Pedersen, G., Engen, M., Johansen, M. S., Johansson, P. N., Schlüter, C., …, Bateman, A. W. (2013). The MBT Adherence and Competence Scale (MBT-ACS): Development, structure and reliability. *Psychotherapy Research, 23*(6), 705–717. https://doi.org/10.1080/10503307.2012.708795.

Kazantzis, N. (2003). Therapist competence in cognitive-behavioural therapies: Review of the contemporary empirical evidence. *Behaviour Change, 20*(1), 1–12. https://doi.org/10.1375/bech.20.1.1.24845.

* Kazantzis, N., Clayton, X., Cronin, T. J., Farchione, D., Limburg, K., & Dobson, K. S. (2018). The Cognitive Therapy Scale and Cognitive Therapy Scale-Revised as measures of therapist competence in cognitive behavior therapy for depression: Relations with short and long term outcome. *Cognitive Therapy and Research, 42*(4), 385–397. https://doi.org/10.1007/s10608-018-9919-4.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajeweski, B. J., Hróbjartsson, A., et al. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies, 48,* 661–671. https://doi.org/10.1016/j.ijnurstu.2011.01.016.

Kuyken, W., & Tsivrikos, D. (2009). Therapist competence, comorbidity and cognitive-behavioral therapy for depression. *Psychotherapy and Psychosomatics, 78*(1), 42–48. https://doi.org/10.1159/000172619.

Machmutow, K., Holtforth, M. G., Krieger, T., & Watzke, B. (2018). Identifying relapse prevention elements during psychological treatment of depression: Development of an observer-based rating instrument. *Journal of Affective Disorders, 227,* 358–365. https://doi.org/10.1016/j.jad.2017.11.009.

* McGrath, K. B. (2013). Validation of the Drexel University ACT/tCBT Adherence and Competence Rating Scale: Revised for use in a clinical population. Doctoral dissertation. Retrieved from https://idea.library.drexel.edu/islandora/object/idea%3A3803.

Meister, R., Jansen, A., Härter, M., Nestoriuc, Y., & Kriston, L. (2017). Placebo and nocebo reactions in randomized trials of pharmacological treatments for persistent depressive disorder. A meta-regression analysis. *Journal of Affective Disorders, 215,* 288–298.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive–behavioural therapy. *Clinical Psychology Review, 33*(3), 484–499. https://doi.org/10.1016/j.cpr.2013.01.010.

Muse, K., & McManus, F. (2016). Expert insight into the assessment of competence in cognitive-behavioural therapy: A qualitative exploration of experts' experiences, opinions and recommendations. *Clinical Psychology and Psychotherapy, 23,* 246–259. https://doi.org/10.1002/cpp.1952.

Muse, K., McManus, F., Rakovshik, S., & Thwaites, R. (2017). Development and psychometric evaluation of the Assessment of Core CBT Skills (ACCS): An observation-based tool for assessing cognitive behavioral therapy competence. *Psychological Assessment, 29*(5), 542–555. https://doi.org/10.1037/pas0000372.

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology, 6,* 1549. https://doi.org/10.3389/fpsyg.2015.01549.

R-Core-Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/.

* Reichelt, F. K., James, I. A., & Blackburn, I. M. (2003). Impact of training on rating competence in cognitive therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, 34(2), 87–99.

Roth, A. D., & Pilling, S. (2007). The competences required to deliver effective cognitive and behavioural therapy for people with depression and with anxiety disorders. Retrieved from https://www.ucl.ac.uk/pals/research/clinical-ed.

Santelmann, H., Franklin, J., Bußhoff, J., & Baethge, C. (2016). Inter-rater reliability of schizoaffective disorder compared with schizophrenia, bipolar disorder, and unipolar depression—A systematic review and meta-analysis. *Schizophrenia Research, 176*(2), 357–363. https://doi.org/10.1016/j.schres.2016.07.012.

* Schmidt, I. D., Strunk, D. R., DeRubeis, R. J., Conklin, L. R., & Braun, J. D. (2018). Revisiting how we assess therapist competence in cognitive therapy. *Cognitive Therapy and Research, 42*(4), 369–384. https://doi.org/10.1007/s10608-018-9908-7.

* Strunk, D. R., Brotman, M. A., DeRubeis, R. J., & Hollon, S. D. (2010). Therapist competence in cognitive therapy for depression: Predicting subsequent symptom change. *J Consult Clin Psychol, 78*(3), 429–437. https://doi.org/10.1037/a0019631.

* Soygüt, G., Uluç, S., & Tüzün, Z. (2008). A pilot study of the reliability and validity of the turkish cognitive therapy adherence and competence scale. *Turkish Journal of Psychiatry*, *19*(2).

* Svartberg, M. (1989). Manualization and competence monitoring of short-term anxiety-provoking psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 26*(4), 564–571. https://doi.org/10.1037/h0085477.

* Tadic, M., Drapeau, M., Solai, S., de Roten, Y., & Despland, J. N. (2003). Development of a competence scale for brief psychodynamic investigation: A pilot study. *Schweizer Archiv für Neurologie und Psychiatrie, 154*(1), 28–35. https://doi.org/10.4414/sanp.2003.01333.

Trajković, G., Starčević, V., Latas, M., Leštarević, M., Ille, T., Bukumirić, Z., et al. (2011). Reliability of the Hamilton Rating Scale for depression: A meta-analysis over a period of 49years. *Psychiatry Research, 189*(1), 1–9. https://doi.org/10.1016/j.psychres.2010.12.007.

* Vallis, T. M., Shaw, B. F., & Dobson, K. S. (1986). The Cognitive Therapy Scale: Psychometric properties. *Journal of Consulting and Clinical Psychology, 54*(3), 381–385. https://doi.org/10.1037/0022-006x.54.3.381.

* Vallis, T. M., Shaw, B. F., & McCabe, S. B. (1988). The relationship between therapist competency and cognitive therapy and general therapy skill. *Journal of Cognitive Psychotherapy: An International Quarterly, 2*(4), 237–249.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48.

* von Consbruch, K., Clark, D. M., & Stangier, U. (2012). Assessing therapeutic competence in cognitive therapy for social phobia: Psychometric properties of the Cognitive Therapy Competence Scale for Social Phobia (CTCS-SP). *Behavioural and Cognitive Psychotherapy, 40*(2), 149–161. https://doi.org/10.1017/s1352465811000622.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*(4), 620–630.

Warshaw, M. G., Dyck, I., Allsworth, J., Stout, R. L., & Keller, M. B. (2001). Maintaining reliability in a long-term psychiatric study: An ongoing inter-rater reliability monitoring program using the

longitudinal interval follow-up evaluation. *Journal of Psychiatric Research, 35*(5), 297–305.

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78,* 200–211. https://doi.org/10.1037/a0018912.

* Weck, F., Bohn, C., Ginzburg, D. M., & Stangier, U. (2011). Assessment of adherence and competence in cognitive therapy: Comparing session segments with entire sessions. *Psychotherapy Research, 21*(6), 658–669. https://doi.org/10.1080/10503307.2011.602751.

* 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., Hautzinger, M., Heidenreich, T., & Stangier, U. (2011). Psychoedukation bei depressiven störungen -erfassung von interventionsmerkmalen und behandlungskompetenzen = Psychoeducation for depression - features of interventions and therapeutic competencies. *PPmP: Psychotherapie Psychosomatik Medizinische Psychologie, 61*(3-4), 148–153. https://doi.org/10.1055/s-0030-1269902.

* Weck, F., Hilling, C., Schermelleh-Engel, K., Rudari, V., & Stangier, U. (2011). Reliability of adherence and competence assessment in cognitive behavioral therapy: Influence of clinical experience. *Journal of Nervous and Mental Disease, 199*(4), 276–279. https://doi.org/10.1097/nmd.0b013e318212461.

* Weck, F., Weigel, M., Richtberg, S., & Stangier, U. (2011). Reliability of adherence and competence assessment in psychoeducational treatment: Influence of clinical experience. *Journal of Nervous and Mental Disease, 199*(12), 983–986. https://doi.org/10.1097/nmd.0b013e3182392da1.

WHO. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* (6th ed.). Geneva: World Health Organization.

Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). dplyr: A grammar of data manipulation. R package version 0.4, 3.

Wirtz, M. A. (2017). Interrater Reliability. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–4). New York: Springer.

Wirtz, M., & Caspar, F. (2002). *Interrater agreement and interrater reliability*. Göttingen: Hogrefe.

* Wittorf, A., Jakobi-Malterre, U. E., Beulen, S., Bechdolf, A., Müller, B. W., Sartory, G., …, Klingberg, S. (2013). Associations between therapy skills and patient experiences of change processes in cognitive behavioral therapy for psychosis. *Psychiatry Research, 210*(3), 702–709. https://doi.org/10.1016/j.psychres.2013.08.01.

Young, J. E., & Beck, A. T. (1980). *Cognitive Therapy Scale: Rating manual*. Unpublished Manuscript, University of Pennsylvania, Philadelphia, PA.

Zarafonitis-Müller, S., Kuhr, K., & Bechdorf, A. (2014). Der Zusammenhang der Therapeutenkompetenz und Adhärenz zum Therapieerfolg in der Kognitiven Verhaltenstherapie - metaanalytische Ergebnisse. *Fortschritte der Neurologie-Psychiatrie, 82,* 502–510.