Available online at www.sciencedirect.com

ScienceDirect

Behavior Therapy 51 (2020) 113–122

Behavior Therapy

www.elsevier.com/locate/bt

The Structure of Competence: Evaluating the Factor Structure of the Cognitive Therapy Rating Scale

Simon B. Goldberg
University of Wisconsin-Madison

Scott A. Baldwin
Brigham Young University

Kritzia Merced
Derek D. Caperton
Zac E. Imel
University of Utah

David C. Atkins
University of Washington

Torrey Creed
University of Pennsylvania

The Cognitive Therapy Rating Scale (CTRS) is an observer-rated measure of cognitive behavioral therapy (CBT) treatment fidelity. Although widely used, the factor structure and psychometric properties of the CTRS are not well established. Evaluating the factorial validity of the CTRS may increase its utility for training and fidelity monitoring in clinical practice and research. The current study used multilevel exploratory factor analysis to examine the factor structure of the CTRS in a large sample of therapists (n = 413) and observations (n = 1,264) from community-based CBT training. Examination of model fit and factor loadings suggested that three within-therapist factors and one between-therapist factor provided adequate fit and the most parsimonious and interpretable factor structure. The three within-therapist factors included items related to (a) session structure, (b) CBT-specific skills and techniques, and (c) therapeutic relationship skills, although three items showed some evidence of cross-loading. All items showed moderate to high loadings on the single between-therapist factor. Results support continued use of the CTRS and suggest factors that may be a relevant focus for therapists, trainers, and researchers.

Keywords: Cognitive Therapy Rating Scale; cognitive behavioral therapy; treatment fidelity; adherence and competence; multilevel factor analysis

Drs. Imel and Atkins are co-founders with equity stake in a technology company, Lyssn.io, focused on tools to support training, supervision, and quality assurance of psychotherapy and counseling. The remaining authors report no conflicts of interest.

Funding was provided by the National Institutes of Health / National Institute on Alcohol Abuse and Alcoholism (NIAAA) under award R01/AA018673. Support for this research was also provided by the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Address correspondence to Simon B. Goldberg, Department of Counseling Psychology, University of Wisconsin-Madison, 335 Education Building, 1000 Bascom Mall, Madison, WI, 53703; e-mail: sbgoldberg@wisc.edu

COGNITIVE BEHAVIORAL THERAPY (CBT) is a widely used psychotherapy that treats a range of psychiatric conditions (e.g., anxiety, depression). A

substantial body of empirical literature supports CBT's efficacy when delivered with high quality (Butler, Chapman, Forman, & Beck, 2006; Hofmann et al., 2012). However, considerable variation can occur in the way in which CBT is actually delivered (Webb, DeRubeis, & Barber, 2010), and some have argued that lower quality implementation may be linked to poorer outcomes in routine clinical care (Shafran et al., 2009). Treatment fidelity is conceptualized to have two components: *adherence* refers to whether a therapist provides theory-specified treatment components (Moncher & Prinz, 1991). *Competence* refers to the degree to which a therapist implements these components skillfully, adapting as necessary based on the needs of a given client (McHugh & Barlow, 2010). Thus, competence is predicated on a therapist adhering to treatment principles. Based on the assumption that skillful implementation of treatment-specific ingredients leads to beneficial outcomes, adherence and competence are vital for clinical practice; assurance that treatments are delivered as intended is crucial for research and implementation efforts (Fairburn & Cooper, 2011). However, to date relatively little attention has been paid to the psychometrics of CBT adherence and competence assessment; measurement limitations may in part explain the lack of a consistent link between these factors and treatment outcome (Webb et al., 2010).

Typically, to assess adherence and competence, trained raters provide standardized assessment of a therapist's behavior during a session. Among the more than 60 different measures of CBT fidelity that were identified in a recent review (Muse & McManus, 2013), the most common and widely used observer-rated measure of CBT fidelity was the Cognitive Therapy Rating Scale (CTRS; Young & Beck, 1980). The CTRS has been used as a benchmark for CBT competence in large-scale randomized clinical trials (e.g., Shaw et al., 1999). The measure includes 11 items rated on a 7-point scale ranging from 0 to 6 (Young & Beck, 1980), covering a range of general therapy skills (e.g., interpersonal effectiveness) and CBT-specific skills (e.g., focusing on key cognitions and behaviors).

The psychometric properties of the CTRS were evaluated at the time of its creation and in one recent study. The original validation studies relied on relatively modest amounts of data drawn primarily from the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH TDCRP; Elkin et al., 1989), while the more recent study (Creed et al., 2016) was conducted in a larger community sample. In this more recent study, Creed et al. demonstrated improvements in CTRS scores over the course of training, with most clinicians (79.6%) reaching established competency benchmarks by the final assessment. Although providing support for the construct validity of the CTRS (i.e., increases over the course of training in CBT; Cronbach & Meehl, 1955), Creed et al. did not evaluate the factor structure of the CTRS. Evaluating the CTRS in a community sample may be particularly valuable given the greater variability in therapist performance, relative to clinical trials, as well as greater external validity related to how CBT may be delivered in practice contexts. In addition, evaluations using larger samples of therapists and clients are vital for reliably establishing the psychometric properties of the CTRS.

Existing evaluations of the CTRS have generally been promising. The CTRS has shown excellent internal consistency reliability (α = .95, item-total correlations ranging from .59 to .90; Dobson et al., 1985; Vallis et al., 1986) as well as evidence of inter-rater reliability (ICC = .59 in Vallis et al., 1986, with high reliability [ICC =.84] in a more recent assessment; Creed et al., 2016). Construct validity has been supported with CTRS scores increasing over the course of training in CBT (Creed et al., 2016).

Another important form of validity is structural (or factorial) validity. Structural validity is important for evaluating the theory underlying a given measure (i.e., what constitutes competence in CBT) as well as informing scoring procedures (e.g., use of subscale scores). The CTRS was originally theorized to be composed of two factors: (1) general skills (e.g., collaboration) and (2) cognitive therapy skills (e.g., conceptualization, strategy, and technique; Young & Beck, 1980; Young, Shaw, Beck, & Budenz, 1981). However, early evaluation of the CTRS factor structure did not fully align with this two-factor model. Vallis et al. (1986), which to our knowledge is the only published factor analysis of the CTRS, used principle components analysis on a small sample of n = 90 session recordings from n = 9 therapists. The authors found that items from both the general skills and cognitive therapy skills subscales loaded on the first factor. This first factor was then defined as "overall cognitive therapy quality" (p. 383, Vallis et al., 1986), with the second factor including items related to session structure (agenda, pacing, and homework). One limitation of this early work was the use of repeated measures without adjustment (i.e., multilevel models or clustered standard errors; Baldwin, Murray, & Shadish, 2005). More recently, researchers have suggested that a three-factor structure may more accurately represent the CTRS components: (1) general therapeutic skills, (2) CBT-specific skills,

and (3) case conceptualization (Creed et al., 2016). This proposed structure has not, however, been evaluated empirically.

Typically, therapists are rated multiple times with the CTRS, which means that ratings are nested within therapists. No study to our knowledge has examined the factor structure of the CTRS using multilevel modeling, which can account for the nesting of multiple ratings within a given therapist. Just as multilevel regression can model relationships at the therapist and client levels (e.g., Baldwin & Imel, 2013; Baldwin, Wampold, & Imel, 2007), multilevel factor analysis can model factor structures at the therapist and client levels.

The items in the therapist-level portion are the CTRS item-averages for a therapist (i.e., aggregating across all clients for a given therapist). Thus, the therapist-level model represents how items "hang together" when considering therapists' entire caseload. In contrast, items in the client-level portion are the CTRS within-therapist deviations—how clients differed from their therapists' mean. Thus, the client-level model represents how the items "hang together" when considering specific clients or sessions. It may be, for example, that certain therapist behaviors (e.g., setting an agenda, using specific CBT skills) vary within a therapist's caseload; a therapist might not structure a session with a given client. However, in general, this more specific subscale may not provide unique information after aggregating across clients. Therefore, the client-level competence ratings from the CTRS are most relevant when supervising a specific case whereas the therapist-level ratings are most relevant when assessing therapists' competence over multiple cases.

Based on the limited number of prior psychometric evaluations of the CTRS, only one prior factor analytic study (Vallis et al., 1986), and the need for evaluation in a large sample using multilevel modeling to account for nested observations within therapists, the present study examined the CTRS using multilevel factor analysis. This was conducted in a large sample of therapists (n = 413) and sessions (n = 1,264). Given uncertainty regarding the proposed structure of the CTRS, a lack of prior multilevel factor analyses, and the sample size requirements for reliable within- and between-factor loadings, exploratory (rather than confirmatory) factor analysis was used.

Materials and Methods

PARTICIPANTS

CTRS data were available for n = 413 therapists across n = 1,264 observations. Therapists were drawn from 26 agencies participating in the Beck Community Initiative, a partnership between the University of Pennsylvania and a large publicly funded mental health system that serves more than 120,000 people annually. Therapists participating in this study were involved in a large-scale CBT training and implementation initiative. A detailed description of the training model for therapists is available for review (see Creed et al., 2016). Briefly, training included attending CBT workshops, 6 months of weekly group consultation, and submitting recorded sessions for competency assessment and training purposes. Session recordings were drawn from all points of the training protocol (i.e., preworkshop, postworkshop, 3 months into the 6-month consultation period, end of 6-month consultation period, 2 years postconsultation period). Having sessions drawn from throughout the training procedure was intended to maximize variability in CTRS scores. Therapists received detailed written feedback on their audio submissions. Participants were drawn from a variety of disciplines and varied in their educational backgrounds and level of training (see Creed et al., 2016).

As the focus of this initiative was on training and implementation of an already-established evidence-based practice, no data were collected regarding client-level variables (e.g., client demographics, outcomes). Although a subsample of clients appears on multiple occasions within the data set, client identification variables were not available, making it impossible to separate client- from session-level variability. Thus, inferences from our models provide information about therapist-level competence across clients included in their caseload. The lack of client-level identifiers prohibits drawing conclusions regarding the factor structure of the CTRS for a particular client across time. Based on uncertainty regarding the nested structure of the data, analyses were conducted using maximum likelihood estimation with robust standard errors (see below; White, 1980). The use of robust standard errors addresses a liberal bias in estimates of standard errors (i.e., inaccurately small standard errors) when repeated observations of the same client were included in a therapists' caseload.

PROCEDURES

Data for the current study were drawn from CTRS ratings administered as part of the Beck Community Initiative. Raters were trained using the CTRS manual and a supplemental rater guide developed to improve reliability. Raters were required to demonstrate reliability on ratings of five consecutive audio recordings prior to becoming study raters by scoring within one point of a gold-standard score on each CTRS item, as well as agreement with whether

Table 1
Item and Total Score Descriptive Statistics

Items	Mean	SD	Min	Max	Therapist ICC [95% CI]
1. Agenda	2.53	1.72	0	6	0.04 [0.01, 0.09]
2. Feedback	2.39	1.49	0	6	0.14 [0.08, 0.20]
3. Understanding	3.24	0.91	0	6	0.18 [0.12, 0.24]
4. Interpersonal Effectiveness	3.96	0.94	0	6	0.21 [0.14, 0.27]
5. Collaboration	3.27	1.07	0	6	0.13 [0.08, 0.19]
6. Pacing	2.92	1.14	0	6	0.10 [0.05, 0.16]
7. Guided Discovery	2.73	1.05	0	6	0.13 [0.08, 0.19]
8. Key Cognitions and Behaviors	2.84	1.29	0	6	0.13 [0.07, 0.19]
9. Strategy for Change	2.69	1.47	0	6	0.08 [0.03, 0.14]
10. CBT Technique	2.33	1.39	0	6	0.11 [0.06, 0.17]
11. Homework	2.14	1.50	0	6	0.08 [0.03, 0.14]
Total Score	31.04	11.10	2	62	0.12 [0.06, 0.18]

Note. Based on n = 1,264 ratings. Item numbering based on Young and Beck (1980). ICC = intraclass correlation coefficient representing the between-therapist variation in CTRS scores.

the total score was ≥ 40. A total of 31 doctoral-level cognitive therapy experts served as trained CTRS raters, with a single rater rating each session (i.e., no session was rated by multiple raters in the current data set). Regular reliability meetings were held among all raters to prevent rater drift, wherein raters independently scored the same audio, recorded their scores to track interrater reliability, and then discussed their rationale for all ratings with the group to reach a consensus score for ongoing training purposes. For this study, raters completed a total of 1,264 CTRS ratings. The therapists had an average of 3.06 sessions rated (SD = 1.20, range = 1 to 7). Interrater reliability on CTRS total scores could not be computed directly in the current sample due to a lack of repeated ratings of a given session. However, interrater reliability was high in the larger sample of ratings from which the current subsample is drawn (ICC = .84; Creed et al., 2016).

MEASURES

The CTRS (Young & Beck, 1980) is an observer-rated measure used to evaluate competence in cognitive therapy skills (Beck, 2011). The measure includes 11 items (see Table 1) scored on a 7-point Likert-type scale ranging from 0 (*poor*) to 6 (*excellent*). A score of 40 has been used as a benchmark for CBT competence (Shaw et al., 1999). Items are designed to assess therapeutic relationship skills (e.g., interpersonal effectiveness), CBT-specific skills (e.g., focusing on key cognitions and behaviors), and structure (e.g., agenda setting). Internal consistency across all 11 items was high in the current sample (α = .94).

DATA ANALYSIS

Data were analyzed using R (R Core Team, 2018) and Mplus statistical software (Muthén & Muthén, 1998-2017). Given uncertainty regarding the hypothesized factor structure, a multilevel exploratory factor analysis (EFA) was conducted (see Supplemental Materials Table 1 for Mplus code). Just like single-level EFA, multilevel EFA requires selecting the number of factors, except in multilevel models one selects the number of factors at the therapist and client levels. Fit indices from models with a varying number of factors at the therapist and client levels were compared. Specifically, the number of factors were varied from 0 to 4 at both the therapist and client levels. Models with 0 factors at a specific level only model the covariance among the items at that level. For example, a model with 0 factors at the therapist level would include covariances among all therapist-level items (i.e., an unrestricted covariance matrix). The fit indices used were the Bayesian Information Criterion (BIC; smaller values better), root mean square error of approximation (RMSEA; smaller values better), comparative fit index (CFI; larger values better), Tucker-Lewis index (TLI; larger values better), and standardized root mean square residuals (SRMR; smaller values better). Per Brown (2015), the following cut-off values were used to define acceptable fit: RMSEA < .05, CFI > .95, and TLI > .95. Models were selected on the basis of fit and evaluation of loadings based on clinical utility and rationale.

As noted above, some clients were represented on multiple occasions within the data set, yielding dependencies between observations (i.e., nesting of clients within therapists, nesting of sessions within clients). Modeling this nested structure was not possible due to a lack of client identifiers. To account for this statistically and reduce a liberal bias in standard error estimates (i.e., inaccurately small standard errors), maximum likelihood estimation with robust standard errors was used. This approach does not assume a particular nesting structure within multilevel data (White, 1980).

Results

Descriptive statistics for CTRS items in the current sample are presented in Table 1. Item means ranged from 2.14 (Homework, standard deviation [SD] = 1.50) to 3.96 (Interpersonal Effectiveness, SD = 0.94), with a mean total score of 31.04 (SD = 11.10). Inspection of item-level histograms did not

indicate significant floor or ceiling effects (Figure 1). The overall total score (*Mean* [*M*] = 31.04) was below the clinical competence benchmark score of 40, although there was evidence that scores increased from pretraining (M = 19.88, SD = 6.98, n = 294) to 6-month posttraining follow-up assessment (M = 38.80, SD = 8.88, n = 280). Among the subsample with both pretraining and 6-month posttraining follow-up assessments (n = 171), a large and statistically significant increase was observed (t[170] = 25.76, $p < .001$, d = 2.42).

Between-therapist variation in CTRS scores was measured with intraclass correlation coefficients (ICC; see Table 1, see Supplemental Materials Table 2 for Mplus code). Higher ICCs indicate that a greater proportion of variance in CTRS scores occurred at the between-therapist level (as opposed to within-therapist level). ICCs varied from 0.08 (strategy for change) to 0.21 (interpersonal effectiveness). Due to a lack of client identifiers in the data set, it was not possible to further disaggregate within-therapist variance into client- and session-level components.

Fit indices from multilevel EFA models are presented in Table 2. Models were examined with one to four within-therapist factors and one to four between-therapist factors. Models were also examined with an unrestricted within-therapist covariance structure. BIC values followed a pattern of improved fit as the number of within-therapist factors increased from one to three, with slightly poorer fit with four within-therapist factors. This pattern was evident regardless of the number of between-therapist factors. RMSEA values followed a pattern of improved fit as the number of within-therapist factors increased from one to four, with the exception of models including two within-therapist factors, for which fit was decreased relative to one within-therapist factor. This pattern was consistent regardless of the number of between-therapist factors. RMSEA values reached the recommended level of < 0.05 with three within-therapist factors regardless of the number of between-therapist factors. Similarly, CFI and TLI values reached the recommended level of > 0.95 with three within-therapist factors, regardless of the number of between-therapist factors.

Next, patterns of factor loadings were examined for interpretability and item absence of cross-loading. As it appeared that either three or four

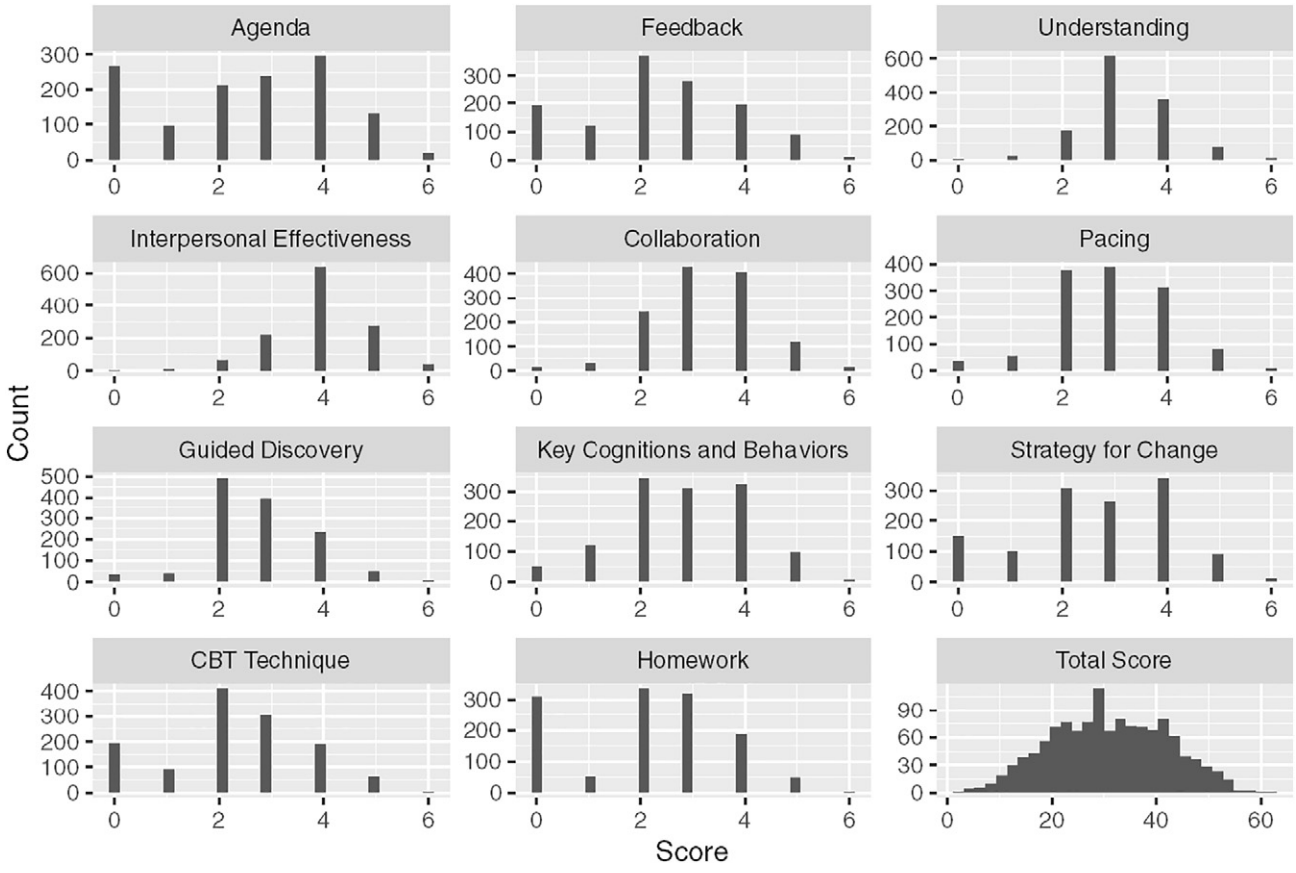

FIGURE 1 CTRS item-level and total score histograms.

Table 2
Exploratory Factor Analysis Model Fit Indices

Within #	Between #	BIC	RMSEA	CFI	TLI	SRMR within	SRMR between
1	1	35995	0.09	0.92	0.90	0.06	0.57
2	1	35703	0.12	0.87	0.81	0.06	0.70
3	1	35460	0.04	0.99	0.98	0.03	0.29
4	1	35463	0.03	0.99	0.99	0.03	0.31
NA	1	35093	0.00	1.00	0.99	0.00	0.34
1	2	35945	0.09	0.93	0.90	0.06	0.72
2	2	35729	0.10	0.92	0.87	0.05	0.18
3	2	35492	0.04	0.99	0.98	0.02	0.13
4	2	35499	0.03	0.99	0.99	0.01	0.10
NA	2	35120	0.00	1.00	0.99	0.00	0.07
1	3	35972	0.09	0.93	0.89	0.05	0.13
2	3	35766	0.09	0.94	0.89	0.04	0.12
3	3	35530	0.04	0.99	0.98	0.01	0.06
4	3	35542	0.03	1.00	0.99	0.01	0.04
NA	3	35160	0.00	1.00	1.00	0.00	0.04
1	4	36011	0.10	0.94	0.88	0.05	0.12
2	4	35810	0.24	0.67	0.29	0.04	0.12
3	4	35576	0.05	0.99	0.97	0.01	0.05
4	4	35589	0.06	0.99	0.96	0.01	0.03
NA	4	35203	0.00	1.00	1.00	0.00	0.02
1	NA	35750	0.07	0.93	0.83	0.05	0.01
2	NA	35500	0.06	0.95	0.85	0.04	0.01
3	NA	35223	0.01	0.99	0.97	0.01	0.00
4	NA	35232	0.00	1.00	0.98	0.01	0.00

Note. NA = unrestricted within covariance; BIC = Bayesian Information Criteria; RMSEA = root mean square error of approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = standardized root mean square residuals.

within-therapist factors fit the data best, factor loadings were examined for these models. The model with three within-therapist factors and one between-therapist factor showed fairly low levels of cross-loaded items and highly interpretable within-therapist factors (Table 3).

At the within-therapist level, Factor 1 was comprised of four items related to session structure (Agenda, Feedback, Homework, Pacing), Factor 2 was comprised of four items related to CBT-specific skills (CBT Technique, Strategies for Change, Key Cognitions and Behaviors, Guided Discovery), and Factor 3 was comprised of therapeutic relationship skills (Collaboration, Interpersonal Effectiveness, Understanding). Collaboration also loaded modestly on Factor 1 (loading = 0.31) and both Pacing and Guided Discovery loaded modestly on Factor 3 (loading = 0.29 for both items). At the between-therapist level, all items showed moderate to high loadings (≥ 0.44) on the single factor. The addition of a fourth within-therapist factor did not appear to improve factor interpretability. One item (Pacing) failed to load strongly on any of the four factors.

Models were then examined with three within-therapist factors and varying numbers of between-therapist factors. Increasing the number of between-therapist factors did not yield interpretable patterns of factor loadings. In a model with three within- and two between-therapist factors, two items (Key Cognitions and Behaviors, Guided Discovery) showed high

Table 3
Within- and Between-Therapist Factor Loadings

	Within			Between
CTRS Item	Factor 1	Factor 2	Factor 3	Factor 1
1. Agenda	**1.09***	-0.26*	0.00	**0.79**
2. Feedback	**0.79***	0.01	0.04	**0.47**
11. Homework	**0.87***	0.01	-0.13*	**0.44**
6. Pacing	**0.40***	0.16*	0.29*	**0.81***
10. CBT Technique	-0.03	**0.96***	0.00	**0.87***
9. Strategies for Change	0.00	**0.96***	-0.04	**0.59**
8. Key Cognitions and Behaviors	0.17*	**0.57***	0.16*	**0.94***
7. Guided Discovery	0.15*	**0.41***	0.29*	**0.79***
5. Collaboration	0.31*	0.00	**0.55***	**0.98***
4. Interpersonal Effectiveness	-0.07	-0.01	**0.77***	**0.82***
3. Understanding	0.00	0.17	**0.65***	**0.95***

Note. Item numbering based on Young and Beck (1980). Loadings bolded to indicate highest factor loading for each item.

cross-loading. Similarly, a model with three within- and three between-therapist factors also failed to yield interpretable factor loadings, with several instance of cross-loaded items (Key Cognitions and Behaviors, Guided Discovery, Homework, Feedback). Thus, it appeared that the model with three within-therapist factors and one between-therapist factor provided the most parsimonious and interpretable factor structure, while simultaneously providing adequate model fit.

Discussion

Evaluation of treatment fidelity is crucial for dissemination and implementation of evidence-based psychotherapies as well as for rigorous psychotherapy research. While the CTRS is a widely used observer-rated measure of CBT treatment fidelity, the measure's factor structure has not been established. The present research is the first large, robust analysis of the CTRS factor structure, using a large sample of community-based therapists (n = 413) being trained in CBT and observed over n = 1,264 observations. Analyses modeled the nesting of observations within therapist, showing that three within-therapist factors and one between-therapist factor yielded a good-fitting model and interpretable factors.

Examination of the pattern of loadings at the within-therapist level may provide insight into the structure of CBT treatment fidelity. The first factor represented structure-related skills, including setting an agenda, assigning homework, eliciting feedback from clients, and pacing the session (CTRS items 1, 2, 6, and 11; Young & Beck, 1980). The second factor was comprised of items specific to CBT, including implementing CBT techniques fluently, engaging in guided discovery, focusing on key cognitions and behaviors, and planning a CBT-oriented strategy for change (CTRS items 7, 8, 9, 10). The third factor was comprised of items reflecting therapeutic relationship skills, including communicating an understanding of clients' thoughts and feelings, interpersonal effectiveness and warmth, and developing a collaborative relationship (CTRS items 3, 4, 5). Thus, it appears that CBT fidelity as assessed via the CTRS in a given session (i.e., within therapist) is composed of a combination of both CBT- and non-CBT-specific skills, along with the ability to structure a session effectively.

In contrast, there appeared to be a single between-therapist factor on which all items loaded, rather than empirically separable domains of competence. Thus, at the therapist level, the CTRS appears to be most useful for making omnibus distinctions of CBT competence. The ability of the CTRS to detect overall, therapist-level skill supports its use in training, supervisory, and quality monitoring contexts. In addition, this omnibus assessment may be further enriched through the three within-therapist factors providing a finer-grained depiction of specific classes of therapeutic behavior that can be targeted for training, supervision, and quality monitoring.

It is worth considering factors that may help contextualize this pattern of multiple within-therapist factors and a single between-therapist factor. One potential contributor is the relatively small between-therapist variability for each item. While generally larger than the proportion of variance in client outcomes attributable to the therapists (i.e., ICC = .05; Baldwin & Imel, 2013), ICCs observed in the current study indicate that the lion's share of variance exists within therapist. This finding puts into question the degree to which competence, as assessed via the CTRS, can be viewed as a therapist-level, rather than client- or session-level, construct. Rather, it may be that some sessions, rather than some therapists, demonstrate competence. There are several theoretically plausible factors that may explain this between-therapist pattern. It may be that therapists' behavior is strongly linked to clients' behavior, such that conceptualizing adherence to specific CTRS domains as a therapist-level trait is less tenable. This could occur for clinically appropriate reasons (e.g., therapists customizing their level of adherence based on a client's needs in a particular session) or could indicate therapists having greater difficulty delivering to a treatment protocol with competence with some clients (e.g., more interpersonally challenging clients; Imel, Baer, Martino, Ball, & Carroll, 2011; Imel et al., 2014). It may also be that most therapists engage in most of the necessary behaviors at some point, such that when scores are aggregated at the between-therapist level, differences between therapists are muted. Further examination of these questions in a data set that includes both therapist and client identifiers is warranted. This would be in keeping with ongoing efforts to establish therapist-level variables that may help explain variation in outcomes across therapists (i.e., therapist effects; Baldwin & Imel, 2013; Goldberg et al., 2018; Johns, Barkham, Kellett, and Saxon, in press; Lingiardi, Muzi, Tanzilli, & Carone, 2017). It may be particularly worthwhile to include CTRS assessments conducted on multiple clients and multiple CTRS assessments conducted on the same therapist-client dyad, in order to increase dependability of therapist-level and dyad-level estimates of adherence, respectively (see Crits-Christoph, Connolly Gibbons, Hamilton, Ring-Kurtz, & Gallop, 2011; Flückiger et al., in press).

This study adds to several decades of work using the CTRS and aids in establishing this measure as a valid and reliable measure of CBT fidelity, marking

the first robust analysis of the measure's structural validity. Although a readily interpretable factor structure was derived using the current data, it will be important for future work to replicate these results, ideally through confirmatory factor analysis and a similarly large sample. Given the high resource demands associated with observer rating systems, it may be valuable to explore the integration of modern technologies such as natural language processing and machine learning to augment and perhaps replace time-intensive human coding (Imel, Steyvers, & Atkins, 2015). The feasibility of this approach has already been demonstrated for assessing motivational interviewing fidelity (Atkins, Steyvers, Imel, & Smyth, 2014) and more recently in the context of CBT (Flemotomos et al., 2018).

While our study lends empirical support to the structural validity of the CTRS, it is worth considering limitations of the CTRS as a measure of CBT fidelity that could be improved through future studies. (We are appreciative to an anonymous reviewer for highlighting these limitations of the CTRS.) For one, the measure was published in 1980. Decades of theoretical and empirical work have continued to clarify both the common and specific mechanisms of action within CBT. It is possible that an updated CTRS could more effectively capture these features than the original version. Relatedly, while the CTRS was presumably developed based on theory and clinical experience, it may be possible to create a more empirically-based fidelity using modern data analytic and measurement methodologies. A second limitation is the measure's emphasis on cognitive techniques. Many modern forms of CBT include behavioral strategies that may not be represented sufficiently on the CTRS (e.g., the word "exposure" does not appear in the CTRS manual; Young & Beck, 1980). Thus, the measure's ability to capture fidelity to some forms of CBT may be less robust.

Several limitations of the current study are worth mentioning. Although our sample size was adequate for conducting EFA, a large enough sample was not available to separate into two portions for conducting both EFA and confirmatory factor analysis, leaving open the question of whether or not the observed factor structure will replicate in other samples. It is therefore crucial that future confirmatory work reevaluate our findings in a separate sample. The potential availability of technologies capable of automating session coding would support this possibility (Imel et al., 2015). Further, although the varied settings in which sessions occurred supports external validity, organizational differences may have also introduced systematic variation (e.g., by workload, productivity demands, staff attitudes towards evidence-based treatment). Unfortunately, the large number of clinics included (n = 26) precluded our ability to test for measurement invariance across clinics. Conducting our study in varied settings limited our ability to include additional measures (e.g., ratings of alliance, treatment outcomes) by which to externally validate our findings. This lack of extra-test correlates greatly limits the degree of validity evidence we can provide in support of the CTRS. Future studies examining the association between CTRS factor scores and key CBT process and outcome measures is therefore a crucial next step.

Another significant limitation was our inability to model nesting of observations within clients. Although statistical techniques designed to account for dependencies within the data (i.e., through use of robust standard errors) were used, lacking client identifiers, it was not possible to disaggregate within-therapist variance into client- and session-level components. While our findings provide insight into therapist competence at the level of their caseload, no information is provided to infer the structure of competence for a particular client over time. It would be valuable to examine this further in a future study. Such a study could assess the degree to which therapist competence appears as a stable therapist-level factor or manifests to varying degrees depending on the particular client (i.e., client-level) or session (i.e., session-level). The relatively modest therapist-level ICCs reported here suggest that a sizable proportion of therapists' competence may depend on the particular client or even session being observed.

Our use of session recordings drawn from a CBT training context is likely both a strength and limitation. Training in CBT may have increased the variability in CTRS scores, which may have increased our ability to reliably estimate factor loadings, a strength provided the validity of ratings is retained. It is also possible that recordings from a CBT training study may not generalize to nontraining contexts (i.e., routine clinical practice in which training in CBT was not being implemented). Being observed within both a training and research context may have influenced therapists' behavior (i.e., Hawthorne effects; Adair, 1984) and therefore the observed structure of the CTRS. It would be valuable for future studies to examine the structure of CTRS scores outside of a training context.

A related limitation was our inability to test for measurement invariance across assessment time points. It is theoretically possible that the structure of the CTRS varies depending on the point in training at which it is assessed. We attempted to conduct a *post hoc* longitudinal measurement

invariance test restricting our sample to the pre- and postworkshop assessments. However, the available sample size (e.g., number of observations per therapist) was limited and model was unidentifiable. Future studies using a larger sample could explore this possibility further.

Conclusion

Despite these limitations, the current study provides the first multilevel factor analytic investigation of the factor structure of the CTRS. The three within-therapist factors and the single between-therapist factors derived from these models provide insight into what characteristics comprise adherent CBT. These results can inform CBT clinical training by identifying component parts of CBT competence that could be targets for training. Results can also inform future investigations studying the CTRS, as well as research on treatment fidelity and therapist differences more generally.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.beth.2019.05.008.

References

Adair, J. (1984). The Hawthorne Effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69* (2), 334–345. https://doi.org/10.1037/0021-9010.69.2.334

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(49). https://doi.org/10.1186/1748-5908-9-49

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change (6th ed.)* (pp. 258–297). Hoboken, NJ: Wiley & Sons.

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or Type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology*, *73*(5), 924–935. https://doi.org/10.1037/0022-006X.73.5.924

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, *75*(6), 842–852. https://doi.org/10.1037/0022-006X.75.6.842

Beck, J. S. (2011). *Cognitive behavior therapy: Basics and beyond*. New York, NY: Guilford Press.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research (2nd ed.)*. New York: Guilford Press.

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: a review of meta-analyses. *Clinical Psychology Review*, *26*(1), 17–31. https://doi.org/10.1016/j.cpr.2005.07.003

Creed, T. A., Frankel, S. A., German, R. E., Green, K. L., Jager-Hyman, S., Taylor, K. P., … Beck, A. T. (2016). Implementation of transdiagnostic cognitive therapy in community behavioral health: The Beck Community Initiative. *Journal of Consulting and Clinical Psychology*, *84*(12), 1116. https://doi.org/10.1037/ccp0000105

Crits-Christoph, P., Connolly Gibbons, M. B., Hamilton, J., Ring-Kurtz, S., & Gallop, R. (2011). The dependability of alliance assessments: The alliance-outcome correlation is larger than you think. *Journal of Consulting and Clinical Psychology*, *79*(3), 267–278. https://doi.org/10.1037/a0023668

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

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

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., … Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Archives of General Psychiatry*, *46* (11), 971–982. https://doi.org/10.1001/archpsyc.1989.01810110013002

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy*, *49*(6-7), 373–378. https://doi.org/10.1016/j.brat.2011.03.005

Flemotomos, N., Martinez, V., Gibson, J., Atkins, D., Creed, T., & Narayanan, S. (2018). Language features for automated evaluation of cognitive behavior psychotherapy sessions. *Proceedings of Interspeech*, *2018*, 1908–1912. https://doi.org/10.21437/Interspeech.2018-1518

Flückiger, C., Hilpert, P., Goldberg, S. B., Caspar, F., Wolfer, C., Held, J., & Vîslă, A. (2019). Investigating the impact of early alliance on predicting subjective change at posttreatment: An evidence-based souvenir of overlooked clinical perspectives. *Journal of Counseling Psychology* Advance online publication

Goldberg, S. B., Hoyt, W. T., Nissen-Lie, H., Nielsen, S. L., & Wampold, B. E. (2018). Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists. *Psychotherapy Research*, *28*(4), 532–544. https://doi.org/10.1080/10503307.2016.1216625

Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440. https://doi.org/10.1007/s10608-012-9476-1

Imel, Z. E., Baer, J. S., Martino, S., Ball, S. A., & Carroll, K. M. (2011). Mutual influence in therapist competence and adherence to motivational enhancement therapy. *Drug and Alcohol Dependence*, *115*, 229–236. https://doi.org/10.1016/j.drugalcdep.2010.11.010

Imel, Z. E., Baldwin, S. A., Baer, J. S., Hartzler, B., Dunn, C., Rosengren, D. B., & Atkins, D. C. (2014). Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients. *Journal of Consulting and Clinical Psychology*, *82*(3), 472–481. https://doi.org/10.1037/a0036158

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computation psychotherapy research: Scaling up the evaluation of patient-provider interactions. *Psychotherapy*, *52*(1), 19–30. https://doi.org/10.1037/a0036841

Johns, R. G., Barkham, M., Kellett, S., & Saxon, D. (2019). A systematic review of therapist effects: A critical narrative update and refinement to review. *Clinical Psychology Review*, *67*, 78–93. https://doi.org/10.1016/j.cpr.2018.08.004

Lingiardi, V., Muzi, L., Tanzilli, A., & Carone, N. (2017). Do therapists' subjective variables impact on psychodynamic psychotherapy outcomes? A systematic literature review. *Clinical Psychology & Psychotherapy*.. https://doi.org/10.1002/cpp.2131

McHugh, R., & Barlow, D. (2010). The dissemination and implementation of evidence-based psychological treatments. *American Psychologist*, *65*(2), 73–84. https://doi.org/10.1037/a0018121

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, *11*(3), 247–266. https://doi.org/10.1016/0272-7358(91)90103-2

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

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (Eight ed.). Los Angeles, CA: Muthén & Muthén.

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

Shafran, R., Clark, D., Fairburn, C., Arntz, A., Barlow, D., Ehlers, A., Freeston, M., et al. (2009). Mind the gap: Improving the dissemination of CBT. *Behaviour Research and Therapy*, *47*, 902–909. https://doi.org/10.1016/j.brat.2009.07.003

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobson, K. S., … Imber, S. D. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology*, *67*(6), 837–846. https://doi.org/10.1037/0022-006X.67.6.837

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

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*(2), 200–211. https://doi.org/10.1037/a0018912

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Young, J., & Beck, A.T. (1980). *Cognitive therapy scale: Rating manual*. Unpublished manuscript, Center for Cognitive Therapy, University of Pennsylvania, Philadelphia, PA.

Young, J,. Shaw, B.F., Beck, A.T., & Budenz, D. (1981). *Assessment of competence in cognitive therapy*. Unpublished manuscript, University of Pennsylvania.

Received: December 28, 2018
Accepted: May 14, 2019
Available online: 24 May 2019