Paper 3

Rasmussen, L-M. P., Patras, J., Handegård, B. H., Neumer, S-P., Martinsen, K. D., Adolfsen, F., Sund, A. M., & Martinussen, M. *(In review).* A validation of the adapted version of the Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS-CBT).

A validation of an adapted group-based version of the Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS CBT)

Abstract

The Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS CBT; Bjaastad et al., 2016) was developed to evaluate the delivery of cognitive therapies for children with clinical anxiety. The present study is an evaluation of the adapted version of the CAS CBT using a sample of group leaders delivering a newly-developed, CBT-based group intervention: EMOTION: *Kids Coping with Anxiety and Depression* (Martinsen, Stark, Rodriguez, & Kendall, 2014). We used a confirmatory factor analysis (CFA) approach in Mplus to test the factor structure of the 11-item instrument. Six raters evaluated a total of N = 239 video-recorded sessions of the EMOTION program. Results showed that we were not able to obtain adequate model fit for the unidimensional 11-item scale χ^2 = 497.076, $p < .05$, df = 44; RMSEA = 0.208, $p < .05$; CFI = .953; and TLI = .941 or the alternate two-subscale solution (i.e., adherence and competence). The final tested model, which removed the items related to session goals, yielded improved but not excellent model fit, χ^2 = 23.26, $p < .05$, df = 11; RMSEA = 0.068, $p = .19$; CFI = .998; and TLI = .997. Further revision of the CAS CBT instrument in order to address group-based interventions may be warranted.

Key words: youths – anxiety – depression – transdiagnostic – observation – validation – confirmatory factor analysis (CFA)

Manual-based interventions consist of prescribed procedures with specified goals and activities designed to produce changes in the target group. Treatment fidelity, or treatment integrity, refers to the therapists' ability to follow the program's core components, which are necessary to produce the desired outcomes (Bond, Evans, Salyers, Williams, & Kim, 2000; Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003). Perepletchikova, Treat, and Kazdin (2007) refer to treatment fidelity using three characteristics. These are 1) adherence, reflecting the therapists' utilization of prescribed intervention procedures, 2) competence, which represents how well the intervention is conducted, and 3) treatment differentiation, which indicates if the treatment differ from others. A high degree of fidelity to an effective program is associated with better treatment outcomes (Carroll et al., 2007; Durlak & DuPre, 2008; Perepletchikova & Kazdin, 2005), but fidelity has received less attention in treatment studies compared to the effectiveness of the intervention (Perepletchikova et al., 2007). It is therefore important to measure both treatment fidelity as well as treatment outcome when evaluating a manual-based intervention. Fidelity is also an important implementation outcome because it measures how well staff have been trained and supported to use the new intervention (Carroll et al., 2007).

Adherence and Competence

Adherence and competence have received a great amount of interest with regard to manualized therapies, mainly to assess and monitor treatment integrity (Perepletchikova & Kazdin, 2005). The present study focuses on a measure comprised of two factors labelled structure and process (Bjaastad et al., 2016), which were originally developed for measuring adherence and competence in CBT for children with anxiety disorders. The structural dimension is analogous to adherence; indicating whether the key components or active ingredients of the program were delivered, and to what extent the intervention was completed (O'Donnell, 2008; Odom, 2008). The process dimension of the measures reflects the quality

of the interaction and relationship between the therapists` and the child(ren) (Justice, Mashburn, Hamre, & Pianta, 2008; O'Donnell, 2008). However, as Bjaastad et al., (2016) found, there are considerable overlap between the adherence questions and competence questions in the two suggested factors, in which the factor structure contains both adherence- and competence-items and vice versa for the process dimension. Researchers recognize the need to address both these dimensions to understand how interventions impact outcome (Odom, Collet-Klingenberg, Rogers, & Hatton, 2010; Webb, DeRubeis, & Barber, 2010). Commonly used methods to assess fidelity are self-reports and observations of the sessions.

Self-report and Observations

In the field of cognitive behavioral therapy (CBT), self-reports such as the Cognitive Behavioral Therapy Checklist (CBTC; Kendall, Gosh, Albano, Ginsburg, & Compton, 2001) have the advantages of being easier to administer and demanding less resources than observations. Filling out self-reports and checklists following delivery can also serve as a reminder to the interventionist about program contents, which in turn can serve to reinforce intervention core components (Bellg et al., 2004). Self-reports, however, rely on individuals' ratings of their own performance, which allows for potential reporter bias (Bellg et al., 2004). Observations, by contrast, are conducted by third parties and are therefore considered a more rigorous and objective measure of treatment adherence (Hogue, Liddle, & Rowe, 1996). Observations, however, require the presence of recording equipment or trained observers in the intervention sessions, which can be time-consuming and expensive. Few of the measures that exist for CBT with children have been evaluated psychometrically, regardless of whether they are self-report or based on observation.

Bjaastad and colleagues (2016) developed the Competence and Adherence scale for CBT (CAS CBT), which is designed for assessing adherence and competence during therapy on youths with anxiety disorders. Anxiety and depression in children are among the most

prevalent psychological problems (Merikangas, Nakamura, & Kessler, 2009), and therefore it is important to develop instruments to measure fidelity when targeting these problems. Also, considering that CBT a commonly used therapy to address these mental health problems (Crowe & McKay, 2017), adequate measures to evaluate adherence and competence in the different interventions being used is therefore evident. Further, to help determine the successfulness of a specific intervention in relation to outcomes, it is important to focus on fidelity (Durlak & DuPre, 2008). It may also help clarify if failures reflect the intervention itself or how it was implemented.

Validation

Construct validity refers to whether a test measures what it is supposed to measure, and is often investigated using a confirmatory factor analysis (CFA) (EFPA, 2013; Floyd & Widaman, 1995). Instrument validation is important to ensure that the instrument used can be applied to similar contexts. For instance, CBT-based interventions for indicated prevention share a lot of common features with clinical therapy, however, conducting interventions in the prevention field involves a number of unverifiable factors (e.g., undefined symptoms in the children, scheduling issues, etc.). Also, resources aligned to support implementation are often limited (Forman, Olin, Hoagwood, Crowe, & Saka, 2009), and typically, assessing adherence and competence is often omitted from prevention studies (Cross & West, 2011; Dane & Schneider, 1998; Dumas, Lynch, Laughlin, Phillips Smith, & Prinz, 2001). Observations of fidelity are particularly rare given the extra resources needed (Hogue et al., 1996; Schoenwald et al., 2011). Furthermore, although highly educated and experienced within their field, many of those working in prevention services do not have formal CBT training. This may impact delivery of a CBT-based program, and is therefore also a reason to evaluate how the interventions are delivered.

The current study

The main goal of the current study was to test the reliability and examine the factor structure of the Competence and Adherence Scale for Cognitive Behavioral Therapy (Bjaastad et al., 2016) with a sample of group leaders delivering a newly developed preventive, CBT-based intervention; EMOTION: *Kids Coping with Anxiety and Depression* (Martinsen et al., 2014). This study was part of a Norwegian multi-site randomized controlled trial (RCT), investigating the effectiveness and the implementation of the EMOTION program (Patras et al., 2016). The CAS CBT has mainly been used with trained therapists working in outpatient clinics treating youth with clinical anxiety. Bjaastad and colleagues (2016), conducted an exploratory factor analysis (EFA) with oblique rotation (direct oblimin principal component analysis), in which two factors were obtained; 1) CBT structure and session goals and 2) Process and relational skills. CAS CBT also showed good internal consistency ($\alpha = .87$), good to excellent interrater reliability (ICC = .83 for Adherence and .64 for Competence) and high rater stability with an ICC = .89 for Adherence and .92 for Competence when the videos were rescored after an average of 17.4 months (Bjaastad et al., 2016). There is a need however, also highlighted by the CAS CBT developers (Bjaastad et al., 2016), to independently validate the instrument using manualized interventions targeting related problem areas, but with different delivery modalities and target groups.

Method

Participants

Participants were trained group leaders ($N = 68$) from different municipal mental health and child welfare services in Norway (e.g., school health services) delivering the EMOTION program. The study sample was 94% women with a mean age of 39.6 ($SD = 9.7$ years). The group leaders were psychologists/specialists (35%), health nurses (14%), educational and psychological counsellors (18%), educators (11%), child-care workers (6%), occupational

therapists (3%) as well as psychology students (5%) and 8% "others" (e.g., counsellor, project leader etc.). Almost 70% of the participants had former experience working with anxiety and depression in youths, and 38% had previously worked with cognitive behavioral therapy (CBT). They received a three-day training, with one-day introduction in general CBT, followed by a two-day workshop in the specific program components of the EMOTION program. The children (N = 266) in the RCT study undergoing the EMOTION program had a mean age of 9.64 years (SD = 0.93), where 56.9 % were girls. The children were recruited based on scores above a predetermined cutoff on anxiety and/or depression instruments. A total of N = 239 sessions (17% of all sessions) were recorded and scored for N = 52 groups.

The EMOTION intervention

The EMOTION program (Martinsen et al., 2014) is a group-based preventive intervention aimed at reducing symptoms of anxiety and depression in children 8-12 years. The intervention builds on regular CBT principles, and during the 20 sessions (one hour sessions, twice per week in a school setting), the main goals were to teach children different sets of skills and strategies to be able to handle their anxiousness or sadness. Additionally, parents received a seven-session course where four of these sessions were together with the children. The parent sessions consisted of themes like positive reinforcements and punishment, positive time with the child, in addition to learning some the same skills as the children learned in their groups. Both the child and parent group sessions were run by two group leaders who were trained in the EMOTION intervention and received regular supervision from an expert in CBT.

Procedure

The research staff distributed video cameras to the intervention group leaders before starting new groups. At the same time a list of the sessions that each leader was to record was distributed. A block of four concurrent children sessions and two concurrent parent sessions

were chosen for each group. The first of each session block was chosen randomly in order to get coverage of a variety of sessions. Sessions were chosen in blocks to simplify the data collection for the group leaders. For example, a group leader may have been randomly assigned to start with session 10, and then follow with sessions 11, 12, and 13. The first and the last session were excluded from the fidelity checks due to the content (introduction and finalization of the groups, respectively). When the groups were finished, the project staff collected and stored the video files at a secure server at one of the participating sites. Ethical approval was obtained from The Regional Committee for Health and Medical Research Ethics (2013/1909/REK Sør-Øst), and the study was registered in clinical trials (NCT02340637).

Measure

The CAS CBT consists of 11-items, divided into three main sections, covering key domains in CBT for children with anxiety (Bjaastad et al., 2016). This includes cognitive therapy structure (e.g., homework, session structure, and progress), process- and relational skills (e.g., reinforcement, collaboration, and flexibility) and treatment goals (specific goals for the session based on the treatment protocol). Adherence is assessed by different items within each of the main sections (e.g., homework, session structure, and progress), while competence is scored globally for each of the main sections. This means that the competence item "cognitive therapy structure" includes an overall competence assessment of both homework and session structure/progress. This potentially causes one of the items being more emphasized than the other(s) within the same section. For instance, the question regarding homework reflects whether the therapists reviewed the participants' homework, and handed out new for next session. The structure/progress question reflects whether the therapists sets an agenda and follows this (including reviewing/presenting new homework), time administration, and general flow of the session. Hence, the structure/progress item generally

receives more emphasis. Further, the item "Flexibility" is rated as a competence score. In addition, there are two questions assessing the overall adherence and competence of the session. These are scored globally, and was added as supplementary items to the scale. The adherence score was rated from *0 = None* to *6 = Thorough*. The competence score ranges from *0 (Poor skills)* to 6 *(Excellent skills)*, with an explanation attached to the ratings, indicating different qualities which needed to be fulfilled. Furthermore, there are three questions about the video quality and challenges with the session (i.e. "What is your evaluation of how challenging this session was?").

In this study, we made a few adaptations of the instrument to fit the EMOTION program in collaboration with the CAS CBT developer. In the original CAS CBT, the parents were included with one item called "parental involvement" (Bjaastad et al., 2016). In EMOTION, the parents received seven individual sessions and therefore this item was removed. The seven parent sessions were rated separately with the same structure as the CAS CBT for children. Also, in the original version, there were two program goals to be rated, but in our version we had up to three goals, so one item was added. The instrument developer(s) approved the modifications.

Rating of items, particularly regarding goals for the sessions were different for each time. During scoring, the items were assessed independently starting at the highest score (*6 = Thorough*) and subtracted accordingly for each element within each goal that was omitted (by the group leaders). If one of the goals for the session was not conducted, the score was *0 (None)*, regardless the reason for this (i.e., prioritizing underway, external factors, time etc.). The group leaders had to present the specific goals in the session, and follow them as described in the program manual to be evaluated by the raters. During the sessions, the two group leaders were scored as one unit, and not assessed individually as they were not assigned a primary and secondary role.

Raters

A scoring team consisting of six people, including both an experienced researcher with previous experience using the instrument, and students with a master's degree or higher in psychology or child care, rated a total number of 239 (17%) videos (170 child sessions and 69 parent sessions). The experienced researcher (scoring 40 individual videos and 66 videos for ICC), with previous clinical practice and video rating experience became the expert rater, which the other raters were tested against. The scoring team received one day of training by the instrument developer in the core elements of the scoring instrument (CAS CBT). In addition, they received a two-day training in the EMOTION program; similar to the group leader-training, focusing on key aspects of the program, session by session. Prior to start up, the raters had to score three of the same videos for training purposes and checking for inter-rater reliability (ICC), and if consensus was met with the expert rater, they could continue. During the project period, ongoing reliability tests were conducted which resulted in a total of 66 randomly selected videos used for interrater reliability (See Table 1 for an overview). Additionally, the team had regular meetings to calibrate, reach consensus and limit drifting. The raters received randomly assigned video recordings for scoring. All raters signed a declaration of confidentiality.

[Insert Table 1 near here]

Statistical analyses

Interrater reliability. The reliability analyses and descriptive analyses were conducted using SPSS statistical packages (24.0). Interrater reliability between raters was calculated using intraclass correlations (ICC, [3, 1]; Shrout & Fleiss, 1979). The ICC`s were calculated by using the model (3, 1) with absolute agreement, which is a Two-Way Mixed Effects Model where people effects are random and measures effects are fixed. The videos were scored by the expert rater and compared against the other observers using the single

measure option. Results were guided by Chicetti`s (1994) principles were ICCs < .40 is considered poor agreement, ICCs between .40 to .59 indicate fair agreement, ICCs between .60 to .74 reflect good agreement and ICCs > .75 show excellent agreement.

[Insert Table 2 near here]

Confirmatory Factor Analyses (CFA). We employed a confirmatory factor analysis (CFA) using Mplus 7.0 statistical software with the weighted least squares estimator (WLSMV; Muthèn & Muthèn, 1998-2010), where the indicators were set as ordered categorical (ordinal). Based on the origin of the instrument and the structure of the items (i.e., competence items depending on the adherence items) the test strategy were as follows (for an overview, see Table 2);

Model 1: Based on the theory behind the instrument, and because the items assessing competence is closely connected to the adherence-items (indicating high correlations), we first tested the fit of a unidimensional model with all items loading on one factor – fidelity.

Model 2: Then, given the structure of the instrument and scoring instructions, we examined a model with the two other factors primarily being evaluated and scored by the instrument (adherence and competence).

***Model* 3a:** Further, we investigated if the originally proposed two-factor structure of CAS CBT (Bjaastad et al., 2016) would be replicated in the current sample (CBT structure and goals and relational skills).

Model 3b: As an extension of the previous model (model 3) and based on methodological issues with reference to how the session goal items were rated during scoring, we tested a two-factor model including 1) structure and goals, and 2) relational skills with correlated item-residuals. The competence items depends highly on the items assessing adherence within the same topic (i.e. competence score on structure and process depends very much on the adherence score structure and process and homework). In addition, the adherence

items are emphasized unevenly (i.e. item 2 “Structure and progress” is emphasized more than item 1 “Homework review and planning new homework”), which indicate a higher correlation between item 2 and item 3, and therefore substantiate the reason to correlate the residuals of these two items.

Model 3c: Furthermore, the item evaluating competence on process and relational skills (item 7) depends highly on the items within the same topic, especially with item 6 (“Flexibility”). Because these items share some common features (i.e., both assesses flexibility and competence, while the two others within this topic is adherence-items), it is expected that the residuals for these items will correlate, and we therefore incorporated this in our model. The correlation between these items were $r = .91$, supporting our indication that these items measure similar constructs.

Model 4: Built on a modified version of the previous model, we investigated a two-factor model; 1) structure and goals, and 2) relational skills, with an alternative structure. Based on the how the items regarding goals for the session were assessed during scoring, not being able to capture why some goals were excluded for instance, could indicate that the latent factor “Structure” is not capable of modelling these particular items adequately. Therefore, we tested whether removing the items evaluating the session goals (items 8-11) would improve model fit. The low correlation between the adherence items within this topic ($r = .04$ to $.42$), reinforced our reasons to examine this model.

To assess how well the model fits the data, multiple fit indices were examined. Chi square is a commonly reported measure of model fit, where a significant result ($p < .05$) indicates misfit (Kline, 2011). Also, Root Mean square Error of Approximation (RMSEA; Steiger and Lind 1980) < .08, and a Bentler’s Comparative fit index (CFI; Bentler, 1990) and Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) > .90 indicate adequate model fit. Preferably, for a good model fit, RMSEA should be < .05, and the CFI and TLI should be

> .95. Also, for RMSEA, a p-value is given, and is interpreted as the probability that the RMSEA < .05).

Correlations. Inter-item correlations between the items were computed using polychoric correlations, which takes into account the ordinal measurement level of the Likert-scale. Correlations between the total scores of structure and CBT session goals, as well the adherence and competence total scores were computed using Persons` *r*.

Internal consistency. Internal consistency for the subscales were estimated with Cronbach`s alpha, where values > 0.70 reflect adequate consistency (EFPA, 2013).

Results

Approximately 20% (*N* = 267) of the total number of sessions were video recorded and scored using the modified version of CAS CBT (Bjaastad et al., 2016). However, some of the videos were not scored (e.g., only parts of the session were recorded due to technical issues, poor video quality or camera placement made scoring impossible, or the group leaders failed to record the whole session). This resulted in 239 (17 %) individually recorded child and parent sessions for 52 groups (*M* = 3.0, *SD* = 1.61 sessions per group).

Interrater reliability

Results showed fair to good interrater reliability (from α = .40 to .74) on all items, and on the mean adherence and mean competence score across all raters compared with the expert rater. See Table 3 for a complete overview of the Mean (*SD*), and ICC scores between the expert rater and the student raters. The items reflecting process and relational skills generally received the lowest scores (.42 to .52), indicating that these were more difficult for the raters to evaluate.

[Insert Table 3 near here]

Confirmatory factor analyses (CFA)

***Model 1*:** Based on the theory behind treatment fidelity and development of the instrument, as well as the structure of the instrument and correlations between the different items, we tested a unidimensional model to check whether the different items loaded on one latent factor (i.e. fidelity). Results showed poor model fit (χ^2 = 497.076, p < .05, df = 44, RMSEA = 0.208, p < .05, CFI = .953, and TLI = .941), which indicates that there are underlying issues within the model which is not taken into account.

Model 2: Further, because the structure of the CAS CBT advocates an adherence- and a competence evaluation of the sessions, we investigated whether the unexplained variance contributed to model by these two constructs which is primarily being assessed by the raters during scoring (i.e. items scored as "adherence" explained by a latent Adherence variable and items scored as "competence" explained by a latent Competence variable). This was also supported by the high inter-item correlations (Table 4). However, results showed that the model covariance matrix was not positive definite, and the technical output in Mplus showed a correlation between the latent factors "adherence" and "competence" > 1 (r = 1.074), which probably implies a model misspecification and that the results are not trustworthy.

Model 3a: Assessing the two component model identified by Bjaastad and colleagues (2016), indicated good fit according to the CFI and TLI, but poor model fit according to the RMSEA and significant misfit according to the chi-square test, χ^2 = 183.69, p < .05, df = 43; RMSEA = 0.117 p < .05; CFI = .985; and TLI = .981.

Model 3: Based on theory, we modified model 3 by allowing the residuals of item 2 (Adherence: Structure and progress) and item 3 (Competence: Cognitive therapy structure) to

correlate, which improved the model results slightly (χ^2 = 162.10, p < .05, df = 42; RMSEA = 0.109, p < .05; CFI = .987; and TLI = .984), but we were not able to get an adequate fit.

Model 3c: Further, as a particularly strong association was expected between item 6 ("Flexibility"), which is also a competence item, and item 7 ("Competence score for Process and relational skills"), we tested a model where we correlated the residuals between item 6 and item 7, but it did not improve model fit (χ^2 =163.37, p < .05, df = 41; RMSEA = 0.112, p < .05; CFI = .987; and TLI = .983).

Model 4: Lastly, we tested a model where all the session goal items were omitted. This model yielded an improved fit over the previous models, χ^2 = 23.26, p < .05, df = 11; RMSEA = 0.068, p = .19; CFI = .998; and TLI = .997. See Table 5 for an overview of the different models tested with pertinent model fit indices.

[Insert Table 4 near here]

[Insert Table 5 near here]

Internal consistency

Based on the models tested in this study, the alpha for the subscale "CBT structure" (excluding session goal items) was .85 which indicated a good reliability, and an excellent alpha for the "Process and relational skills" subscale (α = .93).

Discussion

This study was conducted to examine the factor structure and reliability of the Cognitive and Adherence Scale for Cognitive Behavioral Therapy (CAS CBT; Bjaastad et al., 2016) in a population of children receiving a preventive group intervention for symptoms of anxiety and depression. Results from our study showed that we were not able to estimate a good model fit when conducting a CFA in our sample, particularly when we included the items evaluating

the session goals. We were unable to replicate the results from Bjaastad and colleagues (2016).

Different models were tested to investigate the structure of CAS CBT beyond the hypothesized replication. A unidimensional model, examining if a higher order latent factor could explain the observed variance-covariance matrix was discarded, indicating that it was too simplistic. This implies that an overall dimension, such as fidelity (or treatment integrity), is difficult to model with all 11 items included. Further, the model evaluating the two dimensions being assessed by the raters during scoring (i.e., adherence and competence) was also discarded due to misspecification of the model, as we were not able to estimate all the parameters reliably. Possibly, this was caused by the high correlations and the high dependency between the items in the two suggested factors, particularly the strong associations between the competence-item and the adherence-items within the same topic, which the model is not able to account for.

Issues regarding high dependency between items was associated to different aspects of the instrument. For instance, while we were able obtain an acceptable internal consistency, we did not receive an adequate model fit when we conducted a CFA. The items cohered to such a large extent that it was easy to compare the scores to each other (scoring high on one item ultimately indicated a high score on the next item). Further, the competence questions were consistently being evaluated based on a global assessment of two or three adherence questions, where the previous questions seemed to explain much of the variance within this factor. For instance, within the topic reflecting “Cognitive therapy structure, the competence question (item 3) is strongly associated with the adherence-questions (item 1-2). The competence question (item 7) in the topic “Process and relational skills” is highly based on the adherence-questions (item 4-5), in addition to the other competence question in the same topic (item 6). Within the “Session goal” topic, the competence score for the session goals

(item 11) is also strongly linked to the accompanying adherence questions (item 8-9). During scoring, the raters will therefore base the competence score mostly on the adherence-ratings, but emphasize them differently. Furthermore, in global video observations such as this, the observers are interpreting and evaluating a concept of interest (i.e., adherence and competence). Guided by a scoring manual and a Likert-type scale, the raters make a judgement based on what they observe, and it will therefore have an impact on the scores. Although the raters in the present study met regularly, there should be more focus on training and how the items are rated, along with more frequent meetings to discuss scoring. Another aspect to consider is that the individuals being observed are aware of the situation and might be affected by having the camera nearby (i.e. nervous or more adherent), which may not always give an accurate picture of what is going on (Breitenstein et al., 2010). This could be partly addressed by having the group leaders record all the sessions and then choose sessions randomly for scoring.

In general, the correlations between the different items were high across the two factors (structure and relation) which originally were meant to be tested. High inter-correlations between the items in the instrument supposedly loading on different factors, as well as low correlations between items loading on the same factor, have an impact on the results. Especially, the adherence items rating the goals for the sessions showed low inter-item correlations, and further showed higher correlations with the items reflecting relational skills. This indicated that some of these items did not adequately fit the model. Looking closer at the individual items, the lack of correlation is not a total surprise. The goals for the sessions are independent, indicating that you do not have to complete one before moving to the next one. This may imply some issues regarding the scoring of these items, and possibly elements of the items which were not captured by the scoring, such as the difference between missing (not completed at all) and a total lack of adherence.

The different goals also varies from session to session, which makes them difficult to fit adequately in the instrument structure. This could be reflected by checking the distribution of the response categories, which was highly positively skewed (on a scale from *0 = None* to *6 = Thorough*), but very high on category 1 (almost not present). This indicate an uneven distribution of the adherence to the session goals. One reason for this could be the transdiagnostic origin of the EMOTION manual, which was quite comprehensive, including many elements for each session. For the program developers, choosing two or three main goals per session was challenging, and this could have impacted the completion, and therefore also the scoring of these particular items. This was partly reinforced by removing the items evaluating the session goals (model 4), which improved the model slightly, although good model fit was not achieved. We recommend a focus on the session goals, both in relevance to training and scoring, but also how they should be interpreted and analysed. For example, including information on outer factors that might impact the adherence to the session goal items (i.e., external factors such as time constraints or ongoing prioritizing affecting lack of adherence).

In Bjaastad and colleagues' (2016) original article, they assessed therapists' individual treatment of anxiety, but were not able to conduct a factor analysis on the group condition due to sample size. This could have been a bias in our study, as group condition, with up to 10 children, could potentially contribute with some issues that is not present during individual treatment and which we were not able to assess with the instrument in its current state (e.g. group dynamics, conflicts between the children, noise etc.). This might have affected the completion of the session goals, and subsequently the scoring of the session. Future studies could adapt for this by including additional questions to assess group dynamics. Also, as this was a preventive intervention targeting children with symptoms of anxiety and depression,

many of the children had unspecific symptoms and unestablished issues, which is more difficult to treat. Hence, the session outcome could be more difficult to evaluate.

One possible interpretation of the lack of model fit could also simply be rooted in methodological issues, and that the original factor structure was tested with a different approach (principal component analysis, with oblique rotation) than in our study. Lastly, another element that should be mentioned is that the name of the instrument and the topics being evaluated during scoring (adherence and competence) are not the same as those the structure of instrument reflects according to the factor analyses (i.e., CBT structure and goals *and* Process and relational skills). This introduces some confusion regarding interpretation and use of the instrument, which merit some consideration in future applicability.

Limitations

Group leaders were rated as one unit and were not given specific tasks or roles in advance. Preferably, a unique score for the two individuals would be optimal to be able to detect any variation between the group leaders. Alternately, assigning the group leaders` different roles as primary and secondary, would produce individual scores which is not influenced by the other group leader.

Further, as the instrument was slightly modified to fit our study, we had to remove one of the items assessing parental involvement, and also added one question assessing adherence for a third goal for the session. This implies that the validation of the instrument is not conducted on the exact same items as the original version of the instrument.

Optimally, we would have performed a multilevel confirmatory factor analysis (MCFA) to assess the between (groups) and within-level (sessions) of the data. However, in our study, the group leaders delivered the intervention in pairs, and therefore the groups were treated as the unit of analysis at the between level because the groups comprised different combinations of group leaders. Further, as the individual-level CFA model did not fit the data,

the MCFA was not warranted (Hox, 2002). We did however consider if the group level had an impact, by testing between-group variance on the two factors primarily being tested. The alpha was low ($\alpha = .16$ for Structure and session goals and $\alpha = .20$ for Process and relation skills), and therefore not relevant in this study.

Although within the acceptable range, some of the inter-rater reliability scores were in the lower range < .50, particularly for the items assessing process and relational skills (e.g. Positive reinforcement, Collaboration, Flexibility). This implies that either it was difficult to come to an agreement regarding these items, or there was something with the instrument that makes it difficult to calibrate and reach consensus when scorings these items. Also, the more raters, the more difficult it will be that everyone completely agrees on all items. Focusing on training and conducting accuracy testing frequently should be obliged, as well as keeping the number of raters to a minimum.

Conclusion

In this study, we found similar factors as originally proposed, however, we were not able to replicate the factor structure with adequate model fit in a CFA. There were some issues with the model, particularly when session goal-items were included, mostly reflected by the high correlations and dependence between the items. This indicates that further development of the measure is warranted and that a revision may be in order to adequately assess the use of different manualized CBT group interventions applied on children with symptoms of anxiety and depression.

References:

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., . . . Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH behavior change consortium. *Health Psychology, 23*, 443-451. doi:10.1037/0278-6133.23.5.443

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246. doi:10.1037/0033-2909.107.2.238

Bjaastad, J. F., Haugland, B. S. M., Fjermestad, K. W., Torsheim, T., Havik, O. E., Heiervang, E. R., & Öst, L.-G. (2016). Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS-CBT) for anxiety disorders in youth: Psychometric properties. *Psychological Assessment, 28*, 908-916. doi:10.1037/pas0000230

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H.-W. (2000). Measurement of Fidelity in Psychiatric Rehabilitation. *Mental Health Services Research, 2*, 75-87. doi:10.1023/a:1010153020697

Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing & Health, 33*.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, 40.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. In (Vol. 6, pp. 284-290). US: American Psychological Association.

Cross, W. F., & West, J. C. (2011). Examining implementer fidelity: Conceptualizing and measuring adherence and competence. *Journal of Children's Services, 6*, 18-33. doi:10.5042/jcs.2011.0123

Crowe, K., & McKay, D. (2017). Efficacy of cognitive-behavioral therapy for childhood anxiety and depression. *Journal of Anxiety Disorders, 49*, 76-87. doi:10.1016/j.janxdis.2017.04.001

Dane, A. V., & Schneider, B. H. (1998). Program Integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review, 18*, 23-45. doi:10.1016/S0272-7358(97)00043-3

Dumas, J. E., Lynch, A. M., Laughlin, J. E., Phillips Smith, E., & Prinz, R. J. (2001). Promoting intervention fidelity: Conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial. *American Journal of Preventive Medicine, 20*, 38-47. doi:10.1016/S0749-3797(00)00272-5

Durlak, J., & DuPre, E. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327-350. doi:10.1007/s10464-008-9165-0

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research, 18*, 237-256. doi:10.1093/her/18.2.237

EFPA. (2013). *EFPA Review model for the description and evaluation of psychological tests: Test review form and notes for reviewers version 4.2.6*. In.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286-299. doi:10.1037/1040-3590.7.3.286

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research, Practice, Training, 33*, 332-345. doi:10.1037/0033-3204.33.2.332

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Justice, L. M., Mashburn, A., Hamre, B., & Pianta, R. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly, 23*, 51-68. doi:10.1016/j.ecresq.2007.09.004

Kendall, P. C., Gosh, E. A., Albano, A. M., Ginsburg, G. S., & Compton, S. (2001). *CBT for child anxiety: Therapist treatment integrity checklist.* . Philadelfia: Temple University.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3 ed.). New York: Guilford Publications.

Martinsen, K. D., Stark, K., Rodriguez, K. O., & Kendall, P. C. (2014). *Mestrende barn Manual*. Oslo: Gyldendal Norsk Forlag.

Merikangas, K. R., Nakamura, E. F., & Kessler, R. C. (2009). Epidemiology of mental disorders in children and adolescents. *Dialogues in Clinical Neuroscience, 11*, 7-20.

Muthèn, L., & Muthèn, B. (1998-2010). *Mplus : Statistical analysis with latent variables : User's guide*. Los Angeles, CA: Muthen & Muthen.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*, 33-84. doi:doi:10.3102/0034654307313793

Odom, S. L. (2008). The Tie That Binds: Evidence-Based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education, 29*, 53-61. doi:10.1177/0271121408329171

Odom, S. L., Collet-Klingenberg, L., Rogers, S. J., & Hatton, D. D. (2010). Evidence-Based Practices in Interventions for Children and Youth with Autism Spectrum Disorders. *Preventing School Failure: Alternative Education for Children and Youth, 54*, 275-282. doi:10.1080/10459881003785506

Patras, J., Martinsen, K. D., Holen, S., Sund, A. M., Adolfsen, F., Rasmussen, L.-M. P., & Neumer, S.-P. (2016). Study protocol of an RCT of EMOTION: An indicated intervention for children with symptoms of anxiety and depression. *BMC Psychology, 4*, 48. doi:10.1186/s40359-016-0155-y

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12*, 365-383. doi:10.1093/clipsy.bpi045

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of consulting and clinical psychology, 75*, 829-841. doi:10.1037/0022-006X.75.6.829

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38*, 32-43. doi:10.1007/s10488-010-0321-0

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Buletinl, 86*. doi:10.1037/0033-2909.86.2.420

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10. doi:10.1007/bf02291170

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. doi:10.1037/a0018912

Tables

Table 1

Distribution of Videos per Observer (Single Scored and ICC)

	Observer						
	Expert	R1	R2	R3	R4	R5	Total
Single videos scored	40	37	22	82	27	31	239
Videos used for ICC		19	10	15	12	10	66
ICC Adherence		.67	.54	.83	.69	.86	
ICC Competence		.45	.45	.63	.58	.53	

Note: ICC = Intraclass correlation [3, 1] by Shrout and Fleiss, 1979; Two-Way Mixed Effect model, single measurement (absolute agreement).

Table 2

Overview of the Test Strategy

Model	**Reasoning**	**Factor(s)**
1. Unidimensional	Based on the theory behind the instrument (treatment fidelity/integrity).	1) Fidelity
2. Two factor alternative structure	Based on the structure and scoring instructions of the instrument	1) Adherence 2) Competence
3a. Two factor structure	Based on the original EFA from Bjaastad et al. 2016	1) Structure and goals 2) Relational skills
3b. Two factor; correlating residuals (item 2 and 3)	An extension of the previous model, based on the dependence between adherence and competence-items and unequal emphasis on the items (2-3)	1) Structure and goals 2) Relational skills
3c. Two factor; correlating residuals (item 6 and 7)	An extension of the previous model, based on the dependence and correlation between items and unequal emphasis on the items (6 and 7)	1) Structure and goals 2) Relational skills
4. Two factor; removing session goals	Modified version of Model 1; Based on methodological grounds (scoring of the sessions goals)	1) Structure and goals 2) Relational skills

Table 3

Inter-Rater Reliability Between Expert and Student Raters for the 11-item CAS CBT Scale and Mean Adherence/Competence.

Item/variable	***M* (*SD*)**			**ICC**
	Total	**Expert rater**	**Student raters (*n* = 5)**	
***N* videos**	*n* = 239[a]		*n* = 66[b]	
1. Homework review/planning homework	3.46 (1.97)	4.00 (2.05)	3.21 (2.07)	.60
2. Structure and progress	3.59 (1.52)	3.20 (1.69)	3.12 (1.66)	.60
3. Cognitive therapy structure (items 1-2)	3.36 (1.48)	3.29 (1.74)	3.03 (1.49)	.52
4. Positive reinforcement	3.91 (1.32)	3.83 (1.47)	3.55 (1.54)	.48
5. Collaboration	4.06 (1.38)	4.24 (1.18)	3.83 (1.38)	.40
6. Flexibility	4.00 (1.36)	4.15 (1.26)	3.64 (1.44)	.42
7. Process and relational skills (items 4-6)	3.90 (1.32)	4.23 (1.25)	3.44 (1.42)	.52
8. Session goal 1	3.53 (1.61)	3.15 (2.12)	3.15 (1.85)	.63
9. Session goal 2	2.93 (2.10)	2.58 (2.32)	2.82 (2.15)	.74
10. Session goal 3[c]	2.61 (1.95)	1.65 (1.60)	1.68 (1.67)	.55
11. Session goals (items 8-10)	3.19 (1.47)	3.08 (1.76)	2.75 (1.49)	.56
12. Global adherence	3.60 (1.47)	3.18 (1.87)	3.23 (1.37)	.49
13. Global competence	3.60 (1.40)	3.55 (1.38)	3.21 (1.37)	.51
Mean score adherence (7 items)	3.55 (1.24)	3.43 (1.33)	3.19 (1.23)	.60
Mean score competence (4 items)	3.61 (1.26)	3.69 (1.34)	3.22 (1.30)	.60

Note: Total scale (11 items); Item 1, 2, 4, 5, 8, 9, 10 (and 12) represents adherence scores. Item 3, 6, 7, 11 (and 13) represents competence scores. The adherence score was rated from *0 = None* to *6 = Thorough*. The competence score ranges from *0 (Poor skills)* to 6 *(Excellent skills).*

[a]*N* = 239 individual videos scored only once,

[b]*N* = 66 videos used for interrater reliability calculations;

[c]*N* = 140 videos scored with session goal 3 (not applicable to all sessions).

Table 4

Polychoric Correlations Between Items

Item	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
2.	.58**									
3.	.71**	.89**								
4.	.56**	.57**	.66**							
5.	.55**	.54**	.69**	.72**						
6.	.51**	.56**	.69**	.74**	.82**					
7.	.56**	.56**	.75**	.85**	.88**	.91**				
8.	.46**	.58**	.59**	.43**	.55**	.51**	.51**			
9.	.39**	.60**	.58**	.40**	.33**	.34**	.39**	.24**		
10.	.24**	.50**	.44**	.29**	.24**	.17	.22*	.04**	.42**	
11.	.57**	.78**	.83**	.67**	.67**	.67**	.73**	.67**	.71**	.53**

Note. N = 239.*Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

Table 5

Overview of the Models Tested with Model Fit Indices

Model	df	Chi square	RMSEA	CFI	TLI	WRMR
1.	44	497.076*	0.208*	0.953	0.941	1.779
2.	43	477.588*	0.206*	0.955	0.942	1.728
3a.	43	183.694*	0.117*	0.985	0.981	0.974
3b.	42	162.104*	0.109*	0.987	0.984	0.898
3c.	41	163.370*	0.112*	0.987	0.983	0.890
4.	11	23.263*	0.068[a]	0.998	0.997	0.362

Note: $N = 239$; * $p < .05$; [a] $p = .19$