


MAIN

Development and initial evaluation of a treatment integrity measure for low-intensity group psychoeducational interventions

Jonah Gosling¹, Melanie Simmonds-Buckley^{1,2}, Stephen Kellett^{1,2} , Daniel Duffy³ and Katarzyna Olenkiewicz-Martyniszyn³

¹Department of Psychology, University of Sheffield, Sheffield, UK, ²Rotherham Doncaster and South Humber NHS Foundation Trust, UK and ³Sheffield Health and Social Care NHS Foundation Trust, UK

Corresponding author: Stephen Kellett; Email: Stephen.Kellett@nhs.net

(Received 28 July 2022; revised 27 July 2023; accepted 27 September 2023; first published online 28 November 2023)

Abstract

Background: Despite the importance of assessing the quality with which low-intensity (LI) group psychoeducational interventions are delivered, no measure of treatment integrity (TI) has been developed.

Aims: To develop a psychometrically robust TI measure for LI psychoeducational group interventions.

Method: This study had two phases. Firstly, the group psychoeducation treatment integrity measure-expert rater (GPTIM-ER) and a detailed scoring manual were developed. This was piloted by $n = 5$ expert raters rating the same LI group session; $n = 6$ expert raters then assessed content validity. Secondly, 10 group psychoeducational sessions drawn from routine practice were then rated by $n = 8$ expert raters using the GPTIM-ER; $n = 9$ patients also rated the quality of the group sessions using a sister version (i.e. GPTIM-P) and clinical and service outcome data were drawn from the LI groups assessed.

Results: The GPTIM-ER had excellent internal reliability, good test–retest reliability, but poor inter-rater reliability. The GPTIM-ER had excellent content validity, construct validity, formed a single factor scale and had reasonable predictive validity.

Conclusions: The GPTIM-ER has promising, but not complete, psychometric properties. The low inter-rater reliability scores between expert raters are the main ongoing concern and so further development and testing is required in future well-constructed studies.

Keywords: Adherence; Competence; Group therapy; Improving Access to Psychological Therapies Programme (IAPT); Psychoeducation

Introduction

In response to the high point prevalence of common mental health problems (Bastiaampillai *et al.*, 2019) and the associated need to deliver evidence-based interventions, the Improving Access to Psychological Therapies (IAPT) programme was commissioned to implement the depression and anxiety NICE guidelines (Clark, 2011) in routine services in England. IAPT services (now called *Talking Therapies for Anxiety and Depression*) are based on a hierarchical stepped-care model and provide evidence-based and progressively intensive psychological interventions, according to ongoing severity, risk and responsiveness to previous intervention (Clark *et al.*, 2018). Step 1 entails the screening and ‘watchful waiting’ of patients with minimal symptoms often by the General Practitioner, Step 2 is the delivery of low-intensity (LI) interventions for patients with mild-to-moderate symptoms delivered by accredited Psychological Well-Being Practitioners (PWPs) and

Step 3 contains traditional psychotherapies (predominantly CBT) for patients with moderate-to-severe symptoms delivered by accredited therapists.

LI interventions are delivered one-to-one or in groups across numerous mediums (e.g. via telephone, computerized CBT, in large groups or in small group workshops; Wakefield *et al.*, 2021). One-to-one LI interventions use psychoeducational materials based on the principles of CBT facilitated by a trained practitioner, and have less than 6 hours contact time, with sessions typically lasting for 30 minutes (Shafran *et al.*, 2021). LI group interventions also last for six sessions, but group sessions tend to be longer and are delivered by pairs of facilitators (Delgado *et al.*, 2016). Whilst high-intensity CBT entails some use of psychoeducation, the change methods are more wide ranging, whereas psychoeducation is a central aspect of LI-CBT (Thompson *et al.*, 2021). LI interventions rely on effective therapeutic alliances being rapidly established, maintained, and effectively terminated (Hadjistavropoulos *et al.*, 2017). PWPs therefore need to be able to competently present and use psychoeducational materials and be adherent to the established LI evidence-based treatment protocols in their efforts to reduce unhelpful therapeutic drift (Green *et al.*, 2014).

IAPT services are increasingly delivering LI psychoeducational groups to ensure access and throughput (Burns *et al.*, 2016). As noted, groups can be delivered either didactically via 'teacher-student' lecture-style large groups (e.g. *Stress Control*; White and Keenan, 1990) or via smaller interactive group workshops (Wykes, 2013). Large didactic psychoeducational groups are effective in the treatment of depression (Cuijpers *et al.*, 2009) and anxiety (Dolan *et al.*, 2021). Delgado *et al.* (2016) found significant outcome variability between large didactic psychoeducational groups despite the intervention being manualized (White and Keenan, 1990), and so called for more research evaluating the effectiveness of group LI interventions that concurrently measured the competency and fidelity of practitioners to LI treatment protocols. Ghaemian *et al.* (2020) noted that these large psychoeducational groups can be difficult to manage, with practitioners also needing good teaching skills to deliver the psychoeducational content in an easy-to-understand manner.

Treatment integrity (TI) is an index of when a psychological treatment is delivered as intended (Yeaton and Sechrest, 1981) and is a composite of three elements: adherence, competency and differentiation (Perepletchikova *et al.*, 2007). In the context of LI interventions, adherence refers to when an intervention is delivered according to and in keeping with the LI treatment protocol (Perepletchikova and Kazdin, 2005). Competency would be the skill with which the intervention was implemented by the LI practitioner (Sharpless and Barber, 2009). Differentiation is the degree to which the LI treatment being delivered observably differs from another psychological intervention as defined by the treatment protocol (Southam-Gerow and McLeod, 2013). TI is considered important in (a) ensuring the effective implementation of evidence-based interventions (Landsverk, 2013), (b) strengthening the functional relationship between interventions and their clinical outcomes (Boyle *et al.*, 2020) and (c) enabling better replicability of interventions both across and between services (McCay *et al.*, 2016). Crane *et al.* (2013) championed the use of valid and reliable TI measures supported by detailed scoring manuals in routine services and have completed useful development work on assessment of TI in the delivery of mindfulness-based interventions.

A recent meta-analysis has provided a review of the evidence base between how an intervention is delivered and its associated outcome (Power *et al.*, 2022). The review synthesized 62 hierarchical and non-hierarchical studies with $n = 8210$ patients comprising 45 adherence–outcome correlations, 39 competence–outcome correlations, and seven integrity–outcome correlations. One of the exclusion criteria was internet or telephone-based treatments and this may have excluded some LI studies; no LI studies were included in the meta or sensitivity analyses. The only significant positive associations were found in non-hierarchical studies of the competence/outcome relationship ($r = 0.17$, 95% CI [0.07–0.26], $p < .001$) and between integrity/outcome in both nonhierarchical ($r = 0.15$, 95% CI [0.06–0.23], $p < .001$) and hierarchical study designs ($r = 0.23$, 95% CI [0.01, 0.43], $p < .044$). In this meta-analysis,

hierarchical studies accounted for when therapists treated multiple patients (i.e. when patients were ‘nested’ underneath individual therapists).

In IAPT services, the development and testing of TI measures for LI interventions has not happened. Whilst valid and reliable LI assessment and treatment competency measures have been developed (Kellett *et al.*, 2021), this was limited to one-to-one LI work. There have been previous calls for a psychometrically robust measure of the quality of delivery of group LI interventions (Burns *et al.*, 2016). In response to this, Noble *et al.* (2021) completed a Delphi study that produced 36 LI group facilitator competencies based around four subcategories: group set-up, group content, group process and group closure. The items from the Noble *et al.* (2021) study provided the foundation for the current study and considering the findings of the Power *et al.* (2022) meta-analysis, it was also decided to expand the realms of the measure to include adherence and differentiation items. This enabled an LI-TI measure to be developed and evaluated. The aims of this study were: (1) to pilot the measure and manual, (2) test content validity and (3) to then test validity and reliability.

Method and Results

Phase 1: Development of the measure and manual

As noted, the group psychoeducation expert rater TI measure (GPTIM-ER) and the sister patient version (GPTIM-P) were grounded in items from the Noble *et al.* (2021) study. The Cognitive Behavioural Maintenance-Adherence Scale (CBMT-AS; Weck *et al.*, 2011), group psychoeducation facilitation skills literature (Brown, 2018; Bennett-Levy *et al.*, 2010), UCL competency framework (Roth and Pilling, 2007), the one-to-one LI-CBT treatment competency measure (Kellett *et al.*, 2021) and the (unvalidated) LI-CBT treatment adherence measure (Richards and Whyte, 2009) were all consulted for reference. The GPTIM-ER was informed by the behaviour change model (COM-B model; Michie *et al.*, 2011), which provides the conceptual basis for LI-CBT assessment and treatment (University College London, 2015). Whilst the measure was informed by the COM-B, the items were not intended to directly measure COM-B variables, as there is an extant COM-B measure (Keyworth *et al.*, 2020). Inclusion of the 5-areas information was informed by evidence that 5-areas work is a common feature of LI treatment in IAPT services (Clark *et al.*, 2009). Because the measure was designed to be able to rate TI across the wide variety of psychoeducational groups delivered in IAPT services, then it was decided to adopt a broad LI-CBT perspective, rather than a disorder-specific perspective or specific CBT approach perspective. Therefore, the treatment differentiation item (i.e. ‘the session was underpinned by cognitive-behavioural theory’) was planned to be able to differentiate all LI group interventions based on the CBT approach from other psychological approaches. This avoided the complexity of creating a differentiation item per disorder. A process of truncation was performed by expert agreement in the research team to remove duplicates. This produced nine items: six group psychoeducation delivery competency items and three adherence and differentiation items. GPTIM-ER items are rated using the Dreyfus (1989) 7-point rating system as used in the cognitive therapy scale-revised (CTS-R; Blackburn *et al.*, 2001; 0–1, incompetent; 1–2, novice; 2–3, advanced beginner; 4–5, proficient; and 5–6, expert).

The research team then created the 9-item GPTIM-P which was developed to parallel the items included within the GPTIM-ER but to be able to be completed by patients. Whilst it was assumed that patients would be unable to evaluate differential aspects of an LI intervention (i.e. differentiating LI-CBT from other approaches), most items could be reworded to have high face validity for any patient receiving an LI group intervention (e.g. setting clear aims for the group and homework setting). The GPTIM-P used a more accessible 5-point Likert scale and so patients rated how much they agreed with a statement (i.e. from *strongly disagree* to *strongly agree*). Full scores on both measures were calculated as a total sum of individual item ratings (i.e. scoring range 0–54).

An 18-page detailed manual was created to assist expert raters in using the GPTIM-ER. The GPTIM-P did not have a rating manual. The manual introduced the theory underpinning the GPTIM-ER, providing rating guidelines and providing clear behavioural markers for each item and their associated scores. The GPTIM-ER and the scoring manual are included in the online [Supplementary material](#). The GPTIM-ER therefore covered the start (i.e. agenda setting), middle (i.e. psychoeducational content) and end of sessions (i.e. homework setting). The GPTIM-ER measure and manual were piloted. Five researchers rated a single *Stress Control* session (White and Keenan, 1990) and intra-class correlation coefficients (ICCs) assessed levels of consistency between raters (Koo and Li, 2016). The pilot ICC score was .83 (95% CI .56 to .96) which is a 'good' level of agreement (Koo and Li, 2016). Minor amendments were made to the manual to achieve better clarity. The research team were not included as raters in any subsequent phases.

Phase 1: Participants, design and analysis

Six PWP's drawn from different IAPT services assessed content validity of the GPTIM-ER (i.e. to evaluate how well the measure covered all relevant aspects of TI that the measure intended to rate). These raters needed to be a senior PWP and have delivered at least ten full LI group programmes. Raters had a mean age of 42.8 years ($SD = 8.09$; range = 35–59), 66% were female ($n = 4$), had been qualified for a mean of 9.7 years ($SD = 4.71$; range = 3–18) and had delivered a mean of 60 LI groups ($SD = 63.77$; range = 30–200). Relevance of each GPTIM-ER item was assessed using the content validity index (CVI; Lynn, 1986) and a minimum of six raters is necessary (Yusoff, 2019). Item relevance was rated on a 4-point ordinal scale (1, not relevant; 2, somewhat relevant; 3, quite relevant; 4, highly relevant) to avoid the neutral midpoint (Waltz and Bausell, 1981). The Polit and Beck (2006) guidelines were followed to produce item level (I-CVI) and scale level (S-CVI) scores. The I-CVI was calculated by dividing the number of raters who gave an item a relevance rating of 3 or 4 by the number of overall raters [$I-CVI = (\text{agreed items})/(\text{number of experts})$]. The S-CVI was calculated using two methods. The S-CVI/UA was calculated by adding the proportion of items on the scale that achieved a relevance score of 3 or 4 by all scorers [$S-CVI/UA = (\text{sum of UA scores})/(\text{number of measure items})$]. The S-CVI/Av was calculated by averaging the proportion of items rated relevant across scorers [$S-CVI/Av = (\text{sum of proportion scores})/\text{number of experts}$]. According to Lynn (1986), acquiring convergent scores for items over 0.67 is considered acceptable, whereas a CVI rating higher than 0.90 is deemed excellent (Polit and Beck, 2006).

Phase 1: Results

The I-CVI of the GPTIM-ER ranged from 0.83 to 1.0, with a mean score of 0.96 across all items ('excellent'). The S-CVI/UA score was 0.78 ('acceptable'). The online [Supplementary material](#) provides the full CVI results.

Phase 2: Setting and participants

The second stage was conducted in collaboration with an IAPT service in the North of England. The service has an in-house training programme for delivery of groups, whereby novices sit in on groups to learn from senior facilitators, progress to delivering small aspects of groups under supervision, to eventually then co-facilitating groups. Three types of participants were involved: patients, the PWP's delivering groups and expert PWP raters. Patients ($n = 78$) attending two separate *Managing Stress* psychoeducational courses were recruited. As part of patient participation, demographic, clinical outcomes and attendance data were accessed. Mean patient age was 33 years ($SD = 11.6$ years, range = 18–64 years), over half were female (63%; $n = 49$) and 83% categorized their ethnicity as 'white' ($n = 65$; with 6% selecting 'Asian or Asian British', 5% selecting 'another ethnic group', 4% selecting 'Black or Black British' and 1% who 'did not state'

their ethnicity). Fifty-one percent ($n = 40$) were attending due to 'generalized anxiety disorder', with 42% ($n = 33$) attending due to a 'depressive episode'. Remaining patients attended due to 'panic disorder' ($n = 2$), 'recurrent depressive disorder' ($n = 1$) and 'adjustment disorder' ($n = 2$). Patient demographic data are presented in the online [Supplementary material](#). Eleven patients completed the GPTIM-P.

A group of expert raters ($n = 8$) were recruited: three qualified PWPs and five senior PWPs. Expert raters were excluded if they had been involved in phase 1. Inclusion criteria were being a qualified PWP currently working in an IAPT service and having delivered at least 10 full psychoeducational courses each. Mean age of expert raters was 43 years ($SD = 10.3$, range = 28–60), six were female (75%), they had delivered a mean of 61.4 psychoeducational groups ($SD = 79.1$, range = 6–250) and they had been qualified for 6 years ($SD = 4.1$, range = 2–13). All the raters were providing supervision and had delivered supervision for a mean of 3 years ($SD = 3$, range = 2 months to 7 years).

Phase 2: Procedure

Ten sessions were recorded to generate the rating sample and recordings were drawn from two separate psychoeducational *Managing Stress* courses. Each session was delivered by a pair of qualified and supervised PWPs (i.e. each completed a 1-year postgraduate certificate in LI interventions accredited by the British Psychological Society). Three out of the four PWP facilitator sample were female, and facilitators had a mean age of 32-years ($SD = 6.30$). Mean years post-qualification PWP experience was 2.5 years ($SD = 1.11$). A PWP could not be a rater and a facilitator in the study.

All expert raters received a recorded 2-hour training session on the GPTIM-ER measure and manual. To ensure that the training provided raters with an effective understanding of the GPTIM-ER, a 6-question training evaluation was developed, and this was completed by five raters. Overall, the training was a positive experience and raters valued its usefulness (see online [Supplementary material](#)). The training provided raters with the opportunity to trial the GPTIM-ER alongside the rating manual, by rating a sample *Stress Control* session. Their ratings could then be compared with the benchmark score achieved by the research team in the initial testing phase. Once this benchmark was seen to be equivalent, raters were then signed off as ready-to-rate. Then, four *Managing Stress* sessions were rated by expert raters. Each rater was given two recordings from one pair of facilitators, along with a second set of recordings from the second pair of facilitators. Each expert rater therefore rated four LI group sessions, and this generated a sample of 32 ratings.

Patients attending psychoeducational groups were asked to complete the GPTIM-P. Prior to group sessions, patients were sent a Qualtrics link so they could complete the post-session measure. Patients were requested to do this for all the group sessions attended. Patient outcome data were retrieved and included outcome scores, attendance, ethnicity, presentation/diagnosis, gender and age. The outcome measures included the Patient Health Questionnaire-9 (PHQ-9; Kroenke *et al.*, 2001) and the General Anxiety Disorder-7 (GAD-7; Spitzer *et al.*, 2006) and from these, an overall change score was calculated. This was achieved by subtracting the final score from that taken at the start of the group and then analysing any resultant change as the outcome. The justification for this was to identify change, but without assuming causality in an observational dataset (Tennant *et al.*, 2022). For attendance outcomes, drop-out was defined when a patient attended <2 sessions in keeping with the IAPT manual definition (National Collaborating Centre for Mental Health, 2018). Two study-specific definitions of attendance were also included: (1) patients who dropped out prior to attending all group sessions, and (2) patients who attended less than 50% of sessions.

Phase 2: Data analysis

Construct validity was tested by comparing the GPTIM-ER with the GPTIM-P (i.e. because both measure similar qualities), and this was achieved by calculating Pearson's correlation coefficients between the two measures. Predictive validity (i.e. the ability of the GPTIM-ER to predict patient outcomes) was tested by calculating Pearson's correlation coefficients between (1) patient outcomes and GPTIM-ER scores and (2) patient drop-out and GPTIM-ER scores, when patients attended at least two group sessions. Correlations were interpreted as $p < 0.25$ being small, $p = 0.25-0.50$ being moderate, $p = 0.50-0.75$ as good and $p > 0.75$ as excellent (Portney and Watkins, 2009). Test-retest reliability (i.e. the stability of the GPTIM-ER over time) was investigated by calculating Pearson's correlation coefficients between first and second GPTIM-ER ratings of the same facilitator pair. These results were interpreted as 0.40–0.59 being 'fair', 0.60–0.74 being 'good' and ≥ 0.75 as 'excellent' (Cicchetti, 1994). Inter-rater reliability (i.e. the degree of agreement between independent PWP expert raters that rated group sessions using the GPTIM-ER) was assessed by calculating intra-class coefficients (ICC; Koo and Li, 2016). A power analysis defined the number of GPTIM-ER ratings needed to reliably calculate inter-rater agreement (Arifin, 2018; Walter *et al.*, 1998) and 32 GPTIM-ER ratings were required (i.e. with 10 sessions rated by three raters each). The calculator was set with an alpha significance level of 0.05 and a power rating of 80%. A one-way mixed effects ICC was calculated in the study, where inter-rater reliability was based on average ratings and absolute agreement levels (Hallgren, 2012; Shrout and Fleiss, 1979). ICC scores were interpreted as < 0.5 (poor), 0.50–0.75 (moderate), 0.75–0.90 (good) and > 0.90 (excellent; Koo and Li, 2016). Using the Crane *et al.* (2013) approach, percent agreement was also calculated on absolute agreement between raters and when close agreement had occurred (i.e. when the raters selected the same or adjacent points on the judged level of TI on the scale). Competence agreement between raters was also assessed as a binary classification according whether raters scored facilitators above or below competence on each item (competent = score of 3 or above).

The construct validity of the GPTIM-ER was further tested using exploratory factor analysis (EFA) to assess whether there were any redundant or inappropriate items. Gorusch (1983) recommended a 5:1 participant to item ratio, and as the GPTIM-ER measure has 9-items then a sample size of at least 45 was required. Principal axis factoring with a direct oblimin (oblique) rotation was used. Prior to conducting the EFA, preliminary tests assessed data suitability. Initially, correlations between all items were checked to ensure correlations of ≥ 0.3 were achieved (Cristobal *et al.*, 2007). The Kaiser-Meyer-Olkin (KMO) index and Bartlett's test of sphericity (Bartlett, 1950) were also conducted. The KMO measure rates scores between 0.00 and 0.49 as 'unacceptable', 0.50–0.59 as 'miserable', 0.60–0.69 as 'mediocre', 0.70–0.79 as 'middling', 0.80–0.89 as 'meritorious' and 0.90–1.00 as 'marvellous' (Dodge, 2008). For Bartlett's test of sphericity, the level of significance was set at $p < 0.05$. Cronbach's alpha (i.e. the level of agreement between the individual items of the GPTIM-ER) tested the internal consistency of the measure (Cronbach, 1951). Alpha scores above 0.70 are 'acceptable', 0.80–0.89 deemed 'very good' and scores > 0.90 deemed 'excellent' (Cortina, 1993). Internal consistency of the GPTIM-ER was further tested via item-total correlations and Guttman split-half reliabilities; item-correlation scores $> .30$ are 'acceptable' (Field, 2013). Finally, descriptive statistics were calculated to examine the means, standard deviations, skewness and kurtosis of the GPTIM-ER item data to check the normality of the data (Hopkins and Weeks, 1990).

Phase 2 results: Factor structure and internal consistency

Factor analysis was appropriate as all of the items correlated at least .3, the KMO was .85 and Bartlett's test of sphericity yielded a significant score ($\chi^2[36] = 188.955$, $p = 0.00$). Factor scores are shown in Table 1. The first factor accounted for 61.71% of the variance and the second, third

Table 1. GPTIM factor loadings

| GPTIM item | Loading |
|--|---------|
| The facilitators were clearly using a psychoeducational approach | .85 |
| The psychoeducational information delivered was well matched to the needs of the group | .76 |
| The session was underpinned by cognitive behavioural theory (and not another theory) | .76 |
| The facilitators shared and then abided by an agenda | .78 |
| The facilitators paced the session appropriately | .74 |
| The facilitators presented the materials in an engaging and enthusiastic manner | .72 |
| The facilitators clearly and accurately communicated the psychoeducational information | .86 |
| The facilitators presented change methods with clarity | .84 |
| The facilitators provided guidance on the content of between session work ('homework') | .73 |

Table 2. Item-total and total alpha results

| GPTIM item | Item-total (if deleted) | Cronbach's alpha (if deleted) |
|--|----------------------------|----------------------------------|
| The facilitators were clearly using a psychoeducational approach | .80 | .91 |
| The psychoeducational information delivered was well matched to the needs of the group | .68 | .91 |
| The session was underpinned by cognitive behavioural theory (and not another theory) | .69 | .91 |
| The facilitators shared and then abided by an agenda | .72 | .91 |
| The facilitators paced the session appropriately | .67 | .91 |
| The facilitators presented the materials in an engaging and enthusiastic manner | .65 | .92 |
| The facilitators clearly and accurately communicated the psychoeducational information | .82 | .90 |
| The facilitators presented change methods with clarity | .79 | .90 |
| The facilitators provided guidance on the content of between session work ('homework') | .66 | .91 |

and fourth factors explained 10.89%, 7.62% and 5.70%, respectively. This suggests a uni-dimensional GPTIM-ER factor solution, because of levelling-off in the scree plot eigenvalues after the first factor, plus the insignificance of the second and successive factors. All nine GPTIM-ER items contributed towards the primary factor and met the minimum criteria of having a primary factor loading of $\geq .4$. All items ranged between .72 and .86 and the loadings were equal across the adherence, differentiation and competency items. Table 2 shows that the GPTIM-ER had excellent internal consistency ($\alpha = .92$) and an excellent Guttman split-half coefficient ($r_{SHG} = .90$). The descriptive, skewness and kurtosis statistics (see online [Supplementary material](#)) found that most of the ratings were negatively skewed. One GPTIM-ER item was deemed symmetrical, six items were moderately skewed and two were highly skewed (Bulmer, 1979). With regard to kurtosis, the data were found to be platykurtic (i.e. a lower and broader central peak with longer tails compared with a normal distribution). Visual examination of histograms also confirmed the negative skew.

Phase 2 results: Test-retest reliability

Test-retest results are displayed in Table 3 and show a significant test-retest correlation on the GPTIM-ER ($r = .668$, $p = .005$). Significant correlations were found for all the items and ranged from .640 ($p = .008$; 'psychoeducational approach') to .847 ($p = .000$; pacing). Seven items had 'good' test-retest reliability (0.60–0.74) and two items had excellent test-retest reliability (≥ 0.75).

Table 3. Full GPTIM and item test–retest scores

| GPTIM item | Pearson correlation (<i>r</i>) | Significance (<i>p</i>) |
|--|----------------------------------|---------------------------|
| The facilitators were clearly using a psychoeducational approach | .847*** | .000 |
| The psychoeducational information delivered was well matched to the needs of the group | .694*** | .003 |
| The session was underpinned by cognitive behavioural theory (and not another theory) | .835*** | .000 |
| The facilitators shared and then abided by an agenda | .695*** | .003 |
| The facilitators paced the session appropriately | .640** | .008 |
| The facilitators presented the materials in an engaging and enthusiastic manner | .686** | .003 |
| The facilitators clearly and accurately communicated the psychoeducational information | .748*** | .001 |
| The facilitators presented change methods with clarity | .712*** | .002 |
| The facilitators provided guidance on the content of between session work ('homework') | .738*** | .001 |
| Total score | .668** | .005 |

Significant at $p < .01$ threshold; *significant at $p < .001$ threshold.

Phase 2 results: Inter-rater reliability

Table 4 details the inter-rater reliability for total and item GPTIM-ER scores. There was poor inter-rater agreement on the GPTIM-ER overall (ICC = $-.13$; 95% CI -2.20 to $.70$). One item had moderate agreement; the 'engaging and enthusiastic' item (ICC = $.50$; 95% CI $-.42$ to $.86$). Two items had poor positive agreement; 'clear and accurate communication' (ICC = $.35$; 95% CI $-.85$ to $.82$) and 'change methods' (ICC = $.12$; 95% -1.51 to $.76$). Seven measure items generated negative values (ICC range $-.13$ to $-.97$). Bartko (1976) suggested that when studies are averaging ICC scores and it is suspected that negative values are due to sampling error, then negative scores can be reset to '0.' This correction then only produced a GPTIM-ER ICC score of 0.11 and this remained in the 'poor' range. Poor agreement was evident across the total adherence and differentiation scores (ICC = $-.46$; 95% CI -3.15 to $.60$) and the total competency scores (ICC = $.16$; 95% CI -1.39 to $.77$). There was insufficient variability in expert ratings, as the mean GPTIM-ER score of all rated sessions was fairly high ($M = 44.7/54$, $SD = 6.25$, range = $27-54/54$). Levels of agreement using the exact percent agreement method remained poor (i.e. 17–40% across items). When using close agreement, allowing for exact and adjacent scores, agreement increased to moderate–good levels (i.e. 67–77% across items). Agreement on overall competence classifications was high. There was 100% agreement on four items and 93% agreement on the remaining five items. Only three sessions (9%) had individual items rated as incompetent by expert raters (i.e. a total of 6 items rated incompetent, ranging between 1–3 incompetent items in those three sessions).

Phase 2 results: Construct and predictive validity

The GPTIM-ER significantly positively correlated with GPTIM-P ($r = .68$, $p = .03$). Predictive validity findings are found in the online [Supplementary material](#). In terms of outcome, the GPTIM-ER was significantly associated with mean change scores on both the PHQ-9 ($r = .68$, $p = .03$) and GAD-7 ($r = 0.68$, $p = .03$). In terms of attendance, GPTIM-ER scores were significantly negatively associated with drop-out ($r = -.68$, $p = .03$).

Table 4. Inter-rater reliability for the full GPTIM and individual items

| GPTIM item | ICC | 95% CI | % Agreement (exact) ^a | % Agreement (exact or adjacent score) ^b | Competence classification agreement ^c |
|---|------|--------------|----------------------------------|--|--|
| 1. The facilitators were clearly using a psychoeducational approach | -.97 | -4.58 to .46 | 40% | 73% | 100% |
| 2. The psychoeducational information delivered was well matched to the needs of the group | -.38 | -2.90 to .63 | 23% | 70% | 93% |
| 3. The session was underpinned by cognitive behavioural theory (and not another theory) | -.65 | -3.68 to .55 | 27% | 67% | 100% |
| 4. The facilitators shared and then abided by an agenda | -.44 | -3.10 to .61 | 20% | 77% | 100% |
| 5. The facilitators paced the session appropriately | -.81 | -4.14 to .51 | 23% | 67% | 93% |
| 6. The facilitators presented the materials in an engaging and enthusiastic manner | .50 | -.42 to .86 | 17% | 70% | 93% |
| 7. The facilitators clearly and accurately communicated the psychoeducational information | .35 | -.85 to .82 | 27% | 67% | 93% |
| 8. The facilitators presented change methods with clarity | .12 | -1.51 to .76 | 37% | 70% | 93% |
| 9. The facilitators provided guidance on the content of between session work ('homework') | -.72 | -3.88 to .53 | 30% | 77% | 100% |
| Adherence/Differentiation average (items 1–3) | -.46 | -3.15 to .60 | 30% | 70% | 98% |
| Competency average (items 4–9) | .16 | -1.39 to .77 | 26% | 71% | 95% |
| Total | -.13 | -2.20 to .70 | 0% | 20% | 100% |

ICC, inter-class correlation coefficient; CI, confidence interval.

^aOverall absolute percentage agreement between all raters based on exact scores.

^bOverall close percentage agreement between all raters based on exact scores or adjacent scores (raters rated same or adjacent point on scale).

^cOverall competency agreement between all raters based on a binary classification of competence (rated as competent ≥ 3 or not competent <3).

Discussion

This pragmatic study sought to develop a psychometrically robust TI measure for LI-CBT psychoeducational groups drawn from routine practice, due to the ubiquitous nature of these interventions in modern primary care psychological service settings. A detailed manual was prepared to support the rating process, in keeping with good practice in this area (Powers *et al.*, 2022). After initial piloting and calculation of initial and indicative ICC rates and confirming sufficient content validity, the GPTIM-ER was then found to be an internally consistent and single factor TI measure, that had sufficient construct and predictive validity and good test-retest reliability. Whilst high levels of internal consistency may indicate item redundancy (Hulin *et al.*, 2001), GPTIM-ER items were sufficiently differentiated. The internal reliability results are presumably due to the unifying potential of CBT frameworks (Muse and McManus, 2013).

Whilst rater agreement was satisfactory during the pilot phase, rates of agreement were seen to fall during the second phase. This deterioration was despite checking the training scores against the initial pilot phase scores before expert raters were allowed to progress. Many of the agreement scores were negative values; this occurs when there is very low between-subject variation (i.e. where variability within the group is greater than across groups; Shrout and Fleiss, 1979). Levels of agreement increased when allowing for exact and close percent agreement. It is worth noting that any high percent agreement is probably influenced by the presence of a ceiling effect. Good agreement between expert judges is seen as the most crucial aspect of the reliability of ratings of clinical work (Bobak *et al.*, 2018; Zechariah *et al.*, 2022). The low levels of agreement found do therefore signal the need for future development and testing of the GPTIM-ER.

There are three potential reasons for the low levels of expert rater agreement found here. Firstly, sampling error due to the small number of groups sampled, small number of expert raters and the lack of variability (Lee *et al.*, 2012; Portney and Watkins, 2009). Secondly, the need to rate the facilitators as pairs may have made this task difficult, particularly when there were differences in approach and style between the facilitator pairs. Thirdly, the brief and remote rater training due to the COVID-19 pandemic may have been suboptimal. Whilst the study aimed to provide deliverable and pragmatic rater training (i.e. able to be completed remotely, at a time to suit each individual PWP), similar research has provided such training both in person and for a longer duration (Gordon, 2006; Pierson *et al.*, 2007; McCay *et al.*, 2016; Muse *et al.*, 2017). Whilst less intensive rater training has been found to be previously feasible for LI-CBT competency measures (Kellett *et al.*, 2021), it is acknowledged that longer and more detailed training in the current TI project may have resulted in higher ICCs. Le (2022) also found pre-recorded materials to be ineffective in achieving rating accuracy. Remote training prevents raters being able to debate and verbally agree on a joint consensus rating which enables better eventual consistency through sense-checking (McCay *et al.*, 2016).

Limitations

All rated sessions were gathered from a single service, only on one type of LI group intervention (i.e. the stress control approach) and the sample size (i.e. number of raters, patients and group treatment sessions) was also relatively small. It is acknowledged that there is a wide range of advice on appropriate sample: item ratios for EFA and a greater patient sample size would have been preferable. As the sessions rated were *Stress Control* (White and Keenan, 1990) which is intended for patients with mild-to-moderate anxiety and/or depression, the generalizability of the results to other LI-CBT groups is therefore questionable. The inclusion criteria for being an expert rater was having previously delivered at least 10 full courses, and one of the raters had delivered only six. This was a mistake on the part of the research team in terms of checking the adherence of all the expert raters to the inclusion and exclusion criteria before commencing the training. Selection of the pool of raters was chosen with an arbitrary cut-off regarding group facilitation experience, so experience levels past the minimum standard were not controlled for. Whilst facilitators and raters were separate, because the study was conducted in a single service, then raters may have known the facilitators, and this may have influenced ratings. In terms of the breadth of the GPTIM-ER, it is acknowledged that many aspects of psychoeducation are not about CBT and may be about, for example, biology (e.g. the effect of caffeine intake on sleep quality). This still requires clarity of psychoeducational presentation (Lukens and McFarlane, 2004). Finally, patients were also absent from the patient GPTIM development process.

Methodological and practical implications

Further research to fine-tune and then re-assess the inter-rater reliability of the GPTIM-ER is required. It would be judicious to use more service sites in this effort. The level of burden placed upon patients by asking them to complete an additional measure (i.e. alongside existing outcome measures) needs to be considered in terms of feasibility. Well-conducted LI groups reduces the potential split of patients experiencing the group as either too simple or too complicated (Young-Southward *et al.*, 2020). Future studies should increase training time (as shown in Kühne *et al.*, 2020 and Perepletchikova, 2011) and consider a format change (e.g. in person). Greater and clearer differentiation between the items relating to adherence, competency and differentiation in the GPTIM-ER manual needs to be considered. Groups of raters ranging in experience should be recruited as the reliability of session ratings is influenced by rater experience (Muse and McManus, 2013). If agreement rates can be improved, then the feasibility of uptake by supervisors in routine practice needs to be well considered (Van Der Vleuten, 1996). Clearly, an easy to understand, well-defined and brief measure would enhance the potential feasibility. Testing of the GPTIM-ER with workshop-based psychoeducational groups is needed.

Conclusions

Despite the ubiquity of LI psychoeducational groups within primary mental health care, the clinical governance of these interventions has arguably been piecemeal. Whilst the GPTIM-ER generated a satisfactory range of indices of reliability and validity, the evidence concerning low agreement between expert raters does undermine the measure. Hence, there is insufficient current evidence to deploy this rating tool in LI services, but this project has made a good head start. Further developmental and testing work could use the Crane *et al.* (2013) work as a template and so improve training of raters (i.e. by refreshing the training and standardization phase for new raters), refine the descriptions of the behavioural anchors for each item (i.e. because the tighter these descriptions are, then the higher the likelihood of better ICC scores), further adapt the GPTIM-ER manual and increase the number of rated sessions across a wider variety of LI groups. These efforts would then hopefully improve the psychometric foundations of the GPTIM-ER.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S1352465823000528>

Data availability statement. The data are available on reasonable request from the corresponding author.

Acknowledgements. With thanks to the PWP raters across the two phases of the study.

Author contributions. **Jonah Gosling:** Conceptualization (equal), Data curation (lead), Formal analysis (lead), Investigation (lead), Writing – original draft (lead); **Mel Simmonds-Buckley:** Conceptualization (equal), Methodology (equal), Supervision (equal); **Stephen Kellest:** Conceptualization (lead), Methodology (equal), Supervision (equal), Writing – review & editing (lead); **Daniel Duffy:** Project administration (equal); **Katarzyna Olenkiewicz-Martyniszyn:** Project administration (equal).

Financial support. None.

Competing interests. The authors declare none.

Ethical standards. The study was ethically reviewed and granted approval (IRAS reference number: 293765). The authors abided by the Ethical Principles for Psychologists and Code of Conduct as set out by the British Psychological Society and the British Association for Behavioural and Cognitive Psychotherapies. All participants provided consent for participation and publication of the study. The study had two phases and the methods and results are grouped accordingly. The first phase of the study consisted of measure/manual development, initial piloting, and a content validity check. The second phase entailed psychometric testing of the measure.

References

- Arifin, W. A. (2018). A web-sample size calculator for reliability studies. *Education in Medical Journal*, 10. <https://doi.org/10.21315/eimj2018.10.3.8>
- Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*. <https://doi.org/10.1111/j.2044-8317.1950.tb00285.x>
- Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83, 762–765.
- Bastiampillai, T. J., Allison, S., Harford, P., Perry, S. W., & Wong, M. L. (2019). Has the UK Improving Access to Psychological Therapies programme and rising antidepressant use had a public health impact? *Lancet Psychiatry*, 6, e8–e9. [https://doi.org/10.1016/S2215-0366\(19\)30040-9](https://doi.org/10.1016/S2215-0366(19)30040-9).
- Bennett-Levy, J., Richards, D. A., Farrand, P., Christensen, H., Griffiths, K. M., Kavanagh, D. J., & Proudfoot, J. (2010). Low intensity CBT interventions: a revolution in mental health care. *Oxford Guide to Low Intensity CBT Interventions*, 3, 18.
- Blackburn, I.-M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy*, 29, 431–446. <https://doi.org/10.1017/S1352465801004040>
- Bobak, C., Barr, P. & O'Malley, A. (2018). Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Medical Research Methodology*, 18, 93. <https://doi.org/10.1186/s12874-018-0550-6>
- Boyle, K., Deisenhofer, A. K., Rubel, J. A., Bennemann, B., Weinmann-Lutz, B., & Lutz, W. (2020). Assessing TI in personalized CBT: the inventory of therapeutic interventions and skills. *Cognitive Behaviour Therapy*, 49, 210–227. <https://doi.org/10.1080/16506073.2019.1625945>

- Brown, N. W.** (2018). *Conducting Effective and Productive Psychoeducational and Therapy Groups: A Guide for Beginning Group Leaders*. Taylor & Francis.
- Burns, P., Kellett, S., & Donohoe, G.** (2016). 'Stress Control' as a large group psychoeducational intervention at Step 2 of IAPT services: acceptability of the approach and moderators of effectiveness. *Behavioural and Cognitive Psychotherapy*, 44, 431–443. <https://doi.org/10.1017/S1352465815000491>
- Bulmer, M. G.** (1979). *Principles of Statistics*. Courier Corporation.
- Cicchetti, D. V.** (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284. <https://doi.org/10.1037/1040-3590.6.4.284>
- Cortina, J. M.** (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98. <https://doi.org/10.1037/0021-9010.78.1.98>
- Clark, D., Layard, R., Smithies, R., Richards, D. R., Suckling, R., & Wright, W.** (2009). Improving access to psychological therapy: initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy*, 47, 910–920. <https://doi.org/10.1016/j.brat.2009.07.010>
- Clark, D. M.** (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *International Review of Psychiatry*, 23, 318–327. <https://doi.org/10.3109/09540261.2011.606803>
- Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M.** (2018). Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. *The Lancet*, 391, 679–686. [https://doi.org/10.1016/S0140-6736\(17\)32133-5](https://doi.org/10.1016/S0140-6736(17)32133-5)
- Crane, R. S., Eames, C., Kuyken, W., Hastings, R. P., Williams, J. M., Bartley, T., Evans, A., Silverton, S., Soulsby, J. G., & Surawy, C.** (2013). Development and validation of the Mindfulness-Based Interventions – Teaching Assessment Criteria (MBI:TAC). *Assessment*, 20, 681–688. doi: [10.1177/1073191113490790](https://doi.org/10.1177/1073191113490790)
- Cristobal, E., Flavian, C., & Guinaliu, M.** (2007). Perceived e-service quality: measurement validity and effects on consumer satisfaction and web site loyalty. *Managing Service Quality*, 17, 317–340. <https://doi.org/10.1108/09604520710744326>
- Cronbach, L. J.** (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. <https://doi.org/10.1007/BF02310555>
- Cuijpers, P., Muñoz, R. F., Clarke, G. N., & Lewinsohn, P. M.** (2009). Psychoeducational treatment and prevention of depression: the 'Coping with Depression' course thirty years later. *Clinical Psychology Review*, 29, 449–458. <https://doi.org/10.1016/j.cpr.2009.04.005>
- Delgado, J., Kellett, S., Ali, S., McMillan, D., Barkham, M., Saxon, D., Donohoe, G., Stonebank, H., Mullaney, S., Eschoe, P., Thwaites, R., & Lucock, M.** (2016). A multi-service practice research network study of large group psychoeducational cognitive behavioural therapy. *Behaviour Research and Therapy*, 87, 155–161. doi: [10.1016/j.brat.2016.09.010](https://doi.org/10.1016/j.brat.2016.09.010)
- Dolan, N., Simmonds-Buckley, M., Kellett, S., Siddell, E., & Delgado, J.** (2021). Effectiveness of stress control large group psychoeducation for anxiety and depression: systematic review and meta-analysis. *British Journal of Clinical Psychology*, 60, 375–399. doi: [10.1111/bjc.12288](https://doi.org/10.1111/bjc.12288)
- Dreyfus, H. L.** (1989). The Dreyfus model of skill acquisition. In J. Burke (ed), *Competency Based Education and Training*. Falmer Press
- Dodge, Y.** (2008). *The Concise Encyclopedia of Statistics*. Springer Science & Business Media.
- Field, A.** (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage.
- Ghaemian, A., Ghomi, M., Wrightman, M. and Ellis-Nee, C.** (2020). Therapy discontinuation in a primary care psychological service: why patients drop out. *the Cognitive Behaviour Therapist*, 13.
- Gordon, P. K.** (2006). A comparison of two versions of the Cognitive Therapy Scale. *Behavioural and Cognitive Psychotherapy*, 35, 343–353. <https://doi.org/10.1017/S1352465806003390>
- Green, H., Barkham, M., Kellett, S., & Saxon, D.** (2014). Therapist effects and IAPT Psychological Wellbeing Practitioners (PWPs): a multilevel modelling and mixed methods analysis. *Behavior Research and Therapy*, 63, 43–54. doi: [10.1016/j.brat.2014.08.009](https://doi.org/10.1016/j.brat.2014.08.009)
- Gorsuch, R. L.** (1983). *Factor Analysis* (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hadjistavropoulos, H. D., Pugh, N. E., Hesser, H., & Andersson, G.** (2017). Therapeutic alliance in internet-delivered CBT for depression or generalized anxiety. *Clinical Psychology & Psychotherapy*, 24, 451–461. doi: [10.1002/cpp.201](https://doi.org/10.1002/cpp.201)
- Hallgren, K. A.** (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8, 23. <https://doi.org/10.20982/tqmp.08.1.p023>
- Hopkins, K. D., & Weeks, D. L.** (1990). Tests for normality and measures of skewness and kurtosis: their place in research reporting. *Educational and Psychological Measurement*, 50, 717–729. <https://doi.org/10.1177/0013164490504001>
- Hulin, C., Netemeyer, R., & Cudeck, R.** (2001). Can a reliability coefficient be too high? *Journal of Consumer Psychology*, 10, 55–58. https://doi.org/10.1207/S15327663JCP1001&2_05
- Kellett, S., Simmonds-Buckley, M., Limon, E., Hague, J., Hughes, L., Stride, C., & Millings, A.** (2021). Defining the assessment and treatment competencies to deliver low-intensity cognitive behavior therapy: a multi-center validation study. *Behavior Therapy*, 52, 15–27. <https://doi.org/10.1016/j.beth.2020.01.006>

- Keyworth, C., Epton, T., Goldthorpe, J., Calam, R., Armitage, C. J. (2020). Acceptability, reliability, and validity of a brief measure of capabilities, opportunities, and motivations ('COM-B'). *British Journal of Health Psychology*, 25, 474–501. doi: [10.1111/bjhp.12417](https://doi.org/10.1111/bjhp.12417)
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16, 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- Kühne, F., Heinze, P., & Weck, F. (2020). Standardized patients in psychotherapy training and clinical supervision: study protocol for a randomized controlled trial. *Trials*, 21, 276. <https://doi.org/10.1186/s13063-020-4172-z>
- Landsverk, J. (2013). Reflections on TI in a dissemination and implementation framework. *Clinical Psychology: Science and Practice*, 20, 114. <https://doi.org/10.1111/cpsp.12027>
- Le, K. (2022). Pre-recorded lectures, live online lectures, and student academic achievement. *Sustainability*, 14, 2910. <https://doi.org/10.3390/su14052910>
- Lee, K. M., Lee, J., Chung, C. Y., Ahn, S., Sung, K. H., Kim, T. W., . . . & Park, M. S. (2012). Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. *Clinics in Orthopedic Surgery*, 4, 149–155. <https://doi.org/10.4055/cios.2012.4.2.149>
- Lukens, E. P., & McFarlane, W. R. (2004). Psychoeducation as evidence-based practice: considerations for practice, research, and policy. *Brief Treatment and Crisis Intervention*, 4, 205–225. <https://doi.org/10.1093/brief-treatment/mhh019>
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35, 382–385. <https://doi.org/10.1097/00006199-198611000-00017>
- McCay, E., Carter, C., Aiello, A., Quesnel, S., Howes, C., & Johansson, B. (2016). Toward TI: developing an approach to measure the TI of a dialectical behavior therapy intervention with homeless youth in the community. *Archives of Psychiatric Nursing*, 30, 568–574. <https://doi.org/10.1016/j.apnu.2016.04.001>
- Michie, S., Ashford, S., Sniehotta, F., Dombrowski, S., Bishop, A., & French, D. (2011). A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. *Psychology & Health*, 26, 1479–1498. <https://doi.org/10.1080/08870446.2010.540664>
- Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review*, 33, 484–499. <https://doi.org/10.1016/j.cpr.2013.01.010>
- Muse, K., McManus, F., Rakovshik, S., & Thwaites, R. (2017). Development and psychometric evaluation of the Assessment of Core CBT Skills (ACCS): an observation-based tool for assessing cognitive behavioral therapy competence. *Psychological Assessment*, 29, 542. <https://doi.org/10.1037/pas0000372>
- National Collaborating Centre for Mental Health (2018). *The Improving Access to Psychological Therapies Manual: Appendices and Helpful Resources*. Retrieved from: [england.nhs.uk/wp-content/uploads/2018/06/the-iapt-manual-v5.pdf](https://www.england.nhs.uk/wp-content/uploads/2018/06/the-iapt-manual-v5.pdf)
- Noble, L. A., Firth, N., Delgado, J., & Kellett S. (2021). An investigation of the competencies involved in the facilitation of CBT-based group psychoeducational interventions. *Behavioural and Cognitive Psychotherapy*, 16, 1–13. doi: [10.1017/S1352465821000084](https://doi.org/10.1017/S1352465821000084)
- Pierson, H. M., Hayes, S. C., Gifford, E. V., Roget, N., Padilla, M., Bissett, R., & Fisher, G. (2007). An examination of the motivational interviewing treatment integrity code. *Journal of Substance Abuse Treatment*, 32, 11–17. <https://doi.org/10.1016/j.jsat.2006.07.001>
- Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice*, 18, 148–153. <https://doi.org/10.1111/j.1468-2850.2011.01246.x>
- 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. <https://doi.org/10.1037/0022-006X.75.6.829>
- Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: issues and research recommendations. *Clinical Psychology: Science and Practice*, 12, 365. <https://doi.org/10.1093/clipsy/bpi045>
- Polit, D. E. & Beck, C. T. (2006). *Essentials of Nursing Research* (6th edn). Lippincott Williams & Wilkins, Philadelphia.
- Portney, L. G., & Watkins, M. P. (2009). *Foundations of Clinical Research: Applications to Practice* (vol. 892). Upper Saddle River, NJ: Pearson/Prentice Hall.
- Power, N., Noble, L.A., Simmonds-Buckley, M., Kellett, S., Stockton, C., Firth, N., & Delgado J. (2022). Associations between treatment adherence-competence-integrity (ACI) and adult psychotherapy outcomes: a systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, 90, 427–445. doi: [10.1037/ccp0000736](https://doi.org/10.1037/ccp0000736)
- Richards, D., & Whyte, M. (2009). *Reach Out: National Programme Student Materials to Support the Delivery of Training for Psychological Wellbeing Practitioners Delivering Low Intensity Interventions*. Rethink Mental Illness.
- Roth, A., & Pilling, S. (2007). *The Competences Required to Deliver Effective Cognitive and Behavioural Therapy for People with Depression and with Anxiety Disorder*. Department of Health.
- Shafraan, R., Myles-Hooton, P., Bennett, S., & Öst, L. G. (2021). The concept and definition of low intensity cognitive behaviour therapy. *Behaviour Research and Therapy*, 138, 103803. doi: [10.1016/j.brat.2021.103803](https://doi.org/10.1016/j.brat.2021.103803).

- Sharpless, B. A., & Barber, J. P. (2009). A conceptual and empirical review of the meaning, measurement, development, and teaching of intervention competence in clinical psychology. *Clinical Psychology Review*, 29, 47–56. <https://doi.org/10.1016/j.cpr.2008.09.008>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86, 420. <https://doi.org/10.1037/0033-2909.86.2.420>
- Southam-Gerow, M. A., & McLeod, B. D. (2013). Advances in applying TI research for dissemination and implementation science: introduction to special issue. *Clinical Psychology: Science and Practice*, 20, 1. <https://doi.org/10.1111/cpsp.12019>
- Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, 166, 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>
- Tennant, P. W. G., Arnold, K. F., Ellison, G. T. H., & Gilthorpe, M. S. (2022). Analyses of ‘change scores’ do not estimate causal effects in observational data. *International Journal of Epidemiology*, 51, 1604–1615. <https://doi.org/10.1093/ije/dyab050>
- Thompson, M. Parker, H. & Cave, C. (2021). Exploring which aspects of a low-intensity CBT intervention were found to contribute to a successful outcome from the service user point of view: a mixed methods study. *Counselling and Psychotherapy Research*, 22, 279–291.
- University College London (2015). *National Curriculum for the Education of Psychological Wellbeing Practitioners*. Department of Health.
- Van Der Vleuten, C. P. M. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Science Education*, 1, 41–67. <https://doi.org/10.1007/BF00596229>
- Wakefield, S., Kellett, S., Simmonds-Buckley, M., Stockton, D., Bradbury, A., & Delgadoillo, J. (2021). Improving Access to Psychological Therapies (IAPT) in the United Kingdom: a systematic review and meta-analysis of 10-years of practice-based evidence. *British Journal of Clinical Psychology*, 60, 1–37. <https://doi.org/10.1111/bjc.12259>
- Walter, S.D., Eliasziw, M. & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistical Methods*, 17, 101–110. <http://www.ncbi.nlm.nih.gov/pubmed/9463853>
- Waltz, C. F., & Bausell, R. B. (1981). *Nursing Research: Design, Statistics, and Computer Analysis*. FA Davis Company.
- Weck, F., Hilling, C., Schermelleh-Engel, K., Rudari, V., & Stangier, U. (2011). Reliability of adherence and competence assessment in cognitive behavioral therapy: influence of clinical experience. *Journal of Nervous and Mental Disease*, 199, 276–279. <https://doi.org/10.1097/NMD.0b013e3182124617>
- White, J., & Keenan, M. (1990). Stress control: a pilot study of large group therapy for generalized anxiety disorder. *Behavioural and Cognitive Psychotherapy*, 18, 143–146. <https://doi.org/10.1017/S0141347300018267>
- Wykes, C. (2013). Are gains made in IAPT psychoeducational groups maintained over time? A qualitative study. D Clin Psy thesis, University College London.
- Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, 49, 156–167. <https://doi.org/10.1037//0022-006X.49.2.156>
- Yusoff, M. S. B. (2019). ABC of response process validation and face validity index calculation. *Education in Medicine Journal*, 11, 55–61. <https://doi.org/10.21315/eimj2019.11.3.6>
- Young-Southward, G., Jackson, A., & Dunan, J. (2020). Group CBT for mild to moderate depression and anxiety: an evaluation of patient satisfaction within a primary care mental health team. *the Cognitive Behaviour Therapist*, 13.
- Zechariah, S., Waller, J. L., Stallings, J., Gess, A. J., & Lehman, L. (2022). Inter-rater and intra-rater reliability of the INSPECT (Interactive Nutrition Specific Physical Exam Competency Tool) measured in multi-site acute care settings. *Healthcare*, 10, 212. doi: [10.3390/healthcare10020212](https://doi.org/10.3390/healthcare10020212)

Cite this article: Gosling J, Simmonds-Buckley M, Kellett S, Duffy D, and Olenkiewicz-Martyniszyn K (2024). Development and initial evaluation of a treatment integrity measure for low-intensity group psychoeducational interventions. *Behavioural and Cognitive Psychotherapy* 52, 317–330. <https://doi.org/10.1017/S1352465823000528>