The Development of the Acceptance and Commitment Therapy Fidelity Measure (ACT-FM): A Delphi Study and Field Test.

Lucy O'Neill[1,2]*, Gary Latchford[1,2], Lance M McCracken[3,4], Christopher D. Graham[1,2,5]

[1]Leeds Institute of Health Sciences, University of Leeds, UK

[2]Department of Clinical and Health Psychology, Leeds Teaching Hospitals NHS Trust, UK

[3]King's College London, Institute of Psychiatry, Psychology & Neuroscience, London, UK

[4]Department of Psychology, Uppsala University, Sweden

[5]Department of Psychology, Queen's University Belfast, David Keir Building, Belfast, Northern Ireland

*corresponding author: Christopher Graham, christopher.graham@qub.ac.uk
Department of Psychology, Queen's University Belfast, David Keir Building, Belfast, BT7 1NN

1

Being able to assess whether psychological therapies are delivered according to their own principles is helpful for assuring treatment quality in research and training. We aimed to develop and preliminarily test a measure of therapist fidelity to Acceptance and Commitment Therapy (ACT) that is concise in measuring key therapist behaviours, reliable, practicable and potentially applicable across therapy contexts. This measure was developed via expert consensus in a Delphi study (Study 1). Here, thirteen expert ACT practitioners (average of 11 years' experience with ACT, half ACBS Peer reviewed ACT trainers) participated in three iterative rounds of online questionnaires. A preliminary draft of the measure was used to initiate discussion. In the first two rounds, participants rated and commented on existing items, the manual, and structure of the measure, and generated new items for consideration. In a third round, participants commented on the emergent draft of The ACT Fidelity Measure (ACT-FM). The Delphi study resulted in a 24-item measure with items structured around the three-part model of psychological flexibility ("Tri-flex") alongside Therapist Stance. Eighty-three percent of the chosen items met the specified criteria for consensus. In Study 2, to investigate usability and preliminary psychometric properties of the ACT-FM, a separate group of nine clinicians used the ACT-FM to rate a video of an ACT therapy session. Inter-rater reliability was moderate to excellent, and based on clinician feedback, the measure was expanded to 25 items. To reach the stated aims, further work is required – particularly evaluating the utility of the ACT-FM across therapy contexts.

Acceptance and Commitment Therapy (ACT; Hayes, Strosahl, & Wilson, 2012) is form of cognitive behavioural therapy, where a range of therapeutic tools and methods (e.g. clinical conversation, metaphor, mindfulness and perspective taking exercises) may be used to engender psychological flexibility. Psychological flexibility can be defined as: "…the capacity to persist or to change behaviour in a way that 1) includes conscious and open contact with thoughts and feelings (openness), 2) appreciates what the situation affords (awareness), and 3) serves one's goals and values (engagement)"(McCracken & Morley, 2014, p. 225). The ACT approach, including the aim of improving psychological flexibility, is informed by a radical behavioural account of language and cognition called Relational Frame Theory and a worldview called Functional Contextualism (Hayes, Barnes-Holmes & Wilson, 2012). Functional Contextualism assumes that the functions of, and influences on, behaviours are incompletely understood outside of the contexts in which they occur. Additionally, it comprises a pragmatic a-ontological stance (Long, 2013), focusing on success in reaching stated goals rather than uncovering what is essentially 'true' or 'real'.

Both Functional Contextualism and Relational Frame Theory concern normal psychological processes and behaviour, and do not assume qualitatively different psychological processes occur in, for example, those given psychiatric diagnoses (Hayes, Strosahl, & Wilson, 2012). Ergo, psychological flexibility is considered a clinically relevant process of change that is common across human populations and situations. Commensurate with this, psychological flexibility is consistently associated with measures of physical, social, and emotional functioning across many different contexts (Kashden & Rottenberg, 2010; Ruiz, 2010).

Systematic reviews have found that ACT improves important outcomes, such as daily functioning and quality of life, in many different clinical presentations, including chronic pain (Hann & McCracken, 2014; Veehof, Trompetter, Bohlmeijer, & Schreurs, 2016), anxiety and depression (A-Tjak et al., 2015), and chronic illness (Graham, Gouick, Krahé, & Gillanders, 2016). There is debate regarding the methodological quality of trials of ACT to date, however (Adkins et al., 2017; Gaudiano, 2009; Öst, 2008, 2014). One clear limitation is that comprehensive treatment fidelity checks are often omitted in trials of ACT (Adkins et al., 2017; A-Tjak et al., 2015; Graham et al., 2016; Öst, 2008, 2014).

Fidelity measures in psychotherapy and in ACT

Fidelity measures are tools that "measure the extent to which an intervention or practice is implemented as intended" (Bond, Becker, & Drake, 2011, p. 127). They are

3

important in several ways for the development and delivery of psychotherapy. In research, when assessing the efficacy of a therapy, fidelity checks help establish that the treatment under study is delivered as intended. In routine clinical practice, fidelity measures can be used to highlight training needs.

Several ACT adherence and competency scales exist (e.g. Luoma, Hayes, & Walser, 2007; 2017; McGrath & Forman, 2012; Plumb & Vilardaga, 2010; Pollard, 2010; Twohig et al., 2010). Each has been developed for specific purposes, and therefore each has strengths and limitations for wider use. The ACT Core Competency Self-Rating Form (Luoma, Hayes, & Walser, 2007; 2017) appears useful for encouraging self-reflection and to scaffold learning, and has the advantage of being designed for use across multiple therapy contexts. However, it is lengthy (50 items), and it is unclear how the tool was developed, or how it performs in practice (e.g. whether it enables reliable measurement of ACT fidelity or leads to improvements in practice.) Plumb and Vilardaga (2010) created a widely used treatment fidelity manual with fewer items. However, it was designed specifically for a trial of ACT for people diagnosed with OCD (Twohig et al., 2010), and includes a reasonably lengthy manual. Pollard (2010) developed a measure called the ACT for Psychosis Adherence and Competence Scale (APACS), which has shown acceptable reliability, and is brief in terms of number of items, but includes a long instruction manual. McGrath and Forman (2012) have published a dissertation on developing an ACT and traditional CBT (tCBT) adherence and competence rating scale, designed for trials comparing ACT to tCBT. It was found to have acceptable interrater reliability and could distinguish ACT from tCBT.

In summary, existing measures have features that may limit their wider uptake reflecting the purposes for which they were originally developed. Some have a large number of items and/or extensive instruction manuals. This is important because the time and cost involved is often cited as a reason for omitting fidelity measurement in clinical trials (Waltz et al., 1993.) Some appear condition or context-specific, while others have not yet been assessed for utility or reliability. This perhaps represents a missed opportunity: A brief measure with acceptable psychometric properties applicable across the many contexts in which ACT is delivered may enhance uptake, facilitate compilation of data from trials of ACT, and could be used further to evaluate the ACT model. It may be helpful for assessing whether theoretically consistent therapist behaviours elicit the expected changes in participant behaviour.

Adopting a communitarian approach, we set out to develop an ACT fidelity measure with the potential to meet several key criteria. First, we intended for it to be concise and

practicable, including a limited number of items that capture key observable and indicative ACT consistent and inconsistent behaviors. Next, we aimed for a wide scope of use, applicable across the contexts in which ACT is used. Finally, we wanted it to provide a framework for the reliable measurement of ACT-therapist behaviours. Developing such a measure necessitates a process of continual research and refinement; much of which necessarily is conducted once the initial measure is developed, via clinical use and assessment in multiple contexts, leading to incremental refinement. However, this does not mean that such aims must be neglected at the initial measure-development stage. Therefore, we report two studies undertaken to initiate progress towards these aims in developing a new measure of ACT fidelity – the ACT Fidelity Measure (ACT-FM).

i)   Expert Delphi.

An expert Delphi study was conducted to develop the items and format of the ACT-FM. While Delphi Studies can be used to enhance the emergent face or construct validity of a measure (DeVon et al., 2007; Hardesty & Bearden, 2004; e.g., Salavati et al., 2017), they may also be used to advance a pragmatic conceptualisation of validity (Long, 2013; Wilson, 2001), that is more consistent with a functional contextualist worldview. We aimed to solicit the suggestions and repeated evaluations of a sufficiently diverse and expert group (in ACT and therapy) to reach consensus on a small item pool that might capture indicative and observable ACT consistent and inconsistent therapist behaviours. We hoped that the use of Delphi methodology with a diverse sample would be helpful in several ways. First, a Delphi study enables experts to suggest new items for inclusion, or alterations to existing items as opposed to just reducing a pre-defined list of items. We assumed that the anonymity, sharing of opinions and iterative approach would reduce some of the social pressures, like dominance of particular individuals, or responding based on social desirability, that could skew the content and format in several detrimental ways. For example, if final item choice reflected the specific therapy experience or context of one or two dominant individuals, then this may limit the scope of the measure. If the language, items and format are not clear to the majority, then this will decrease the likelihood of consistent scoring. Indeed, if the measure appears irrelevant, uses inaccessible language or formats, or the items/format necessitates excessive explication via a lengthy manual, then this will affect uptake.

ii)   Field study

A field study, involving clinicians, was undertaken involving a group of experienced but less expert ACT clinicians. This was study had several purposes. First, it was used to assess the reliability of the ACT-FM for measuring ACT consistent and inconsistent behaviours in practise. Second, via application, it was used to assess the usability of the ACT-FM, and to gain feedback on the clarity of items and to suggest alterations.

Study 1: Delphi Study to develop the Acceptance and Commitment Therapy Fidelity Measure (ACT-FM)

Material and Methods

Design

Delphi methodology (Dalkey & Helmer, 1963) was used to integrate the opinions of a group of ACT experts to develop the ACT-FM. Delphi methods have four necessary features, which we considered when designing the study. First, panellist anonymity, which reduces the influence of social pressures (e.g. influential members dominating discussion, responding based on social desirability) on participant response. Second, controlled feedback of the panellists' judgements, to present participants with other panellist's viewpoints. Third, iteration, to enable participants to refine their views in light of presented information. Finally, statistical aggregation of the panellists' responses - to enable an interpretation of the data (Linstone & Turoff, 1975, p. 3). The study was approved by the University of Leeds School of Medicine Research Ethics Committee (Approval date: 19/6/2017, Approval ref: MREC16-120).

Participants

Delphi Panel

Choosing suitably qualified participants is crucial to the Delphi process (Hsu & Sandford, 2007). Consequently, the inclusion criteria were: Professionals who have worked in the field of ACT either clinically and/or research based for a minimum of 5 years and/or professionals who are recognised as a peer reviewed ACT trainer by the Association of Contextual Behavioural Science (ACBS).

ACT experts initially were recruited via existing contacts and reputation. Further participants were recruited via snowballing. As recommended (Rowe & Wright, 2001), we aimed to recruit between five and 20 experts.

6

Research Team

The Research Team consisted of Lucy O'Neill (Psychologist in Clinical Training), Dr Chris Graham, Dr Gary Latchford and Professor Lance McCracken (clinical psychologists who have many years clinical experience and/or significant clinical and research experience with ACT). The Research Team considered comments and adjudicated suggestions made by the panel in the context of clinical experience and literature on ACT and on fidelity measures – and then made changes to the measure.

Measures

Initial draft of the ACT-FM

It was decided to structure the ACT-FM around the Tri-flex (Openness, Awareness, Engagement; Hayes, Strosahl, & Wilson, 2012) and the general ACT therapeutic stance, as opposed to using a hex-a-flex (six-process) structure. The aim was to keep the measure concise and practicable – limiting the number of items to a workable level yet still enabling a reasonably granular classification of therapist behaviours to components of psychological flexibility. Informed by knowledge of ACT, and by reviewing some existing ACT fidelity measures, a preliminary pool of 42 items was generated by the Research Team. Commensurate with guidelines recommending prescribed and proscribed behaviours, ACT consistent and inconsistent domains were included within each of the four sections. This resulted in eight domains in total. A manual was included that comprised instructions for raters and definitions of the ACT processes under study. An initial scoring system was adapted from the measure developed by Plumb and Vilardaga (2010), where therapist behaviours are rated for frequency and extensiveness on a one-to-five scale. Two example items are: "Therapist links behaviour change to client's personal values (i.e. emphasises that behaviour change serves the purpose of greater contact with values)." (Engaged Consistent); "Therapist uses coercion or attempts to persuade the client." (ACT Stance Inconsistent).

Delphi survey

The questionnaires were hosted on the Bristol Online Surveys website (BOS; University of Bristol, 2009). The first round included a draft of the ACT-FM manual with space for comments, the list of 42 initial items for scoring and comments, and a nomination form to suggest further participants. When assessing each individual item, panelists were asked to respond to the following questions on a 7-point rating scale, where, 1= 'not at all' and 7= 'definitely': 1) 'How well does this item capture the above ACT concept?' 2) 'How

7

observable is this therapist behaviour?' 3) 'Do you think this item should be included in the final measure?'.

The second round included the revised draft of the manual for comments and the revised items. Items were each assessed via one question ('Do you think this item should be included in the final measure?'), which was rated on a 7-point scale (where 1= definitely do not include, 7= definitely do include). Respondents were also invited to give general comments on the emergent measure and items.

The third round questionnaire was designed to collect feedback on the emergent version of the ACT-FM, which was emailed as a pdf to the panel. The survey invited comments on: 1) the manual and scoring, 2) the items, 3) the layout, 4) any other comments and suggestions.

Procedure

Potential Delphi Panellists were sent successive emails that contained links to the appropriate round of the Delphi survey. Only participants who took part in Round 1 were invited to take part in Rounds 2 and 3. Each survey took approximately 30 to 60 minutes to complete.

After Panellist's had completed the Round 1 survey, ACT-FM items were ranked based on aggregated responses to question 3 for each domain. The Research Team considered these aggregated ratings and comments for each item, and made decisions about the inclusion, exclusion and modification of items. There is no accepted gold standard for measuring consensus in a Delphi study (Von der Gracht, 2012; Hsu & Sandford, 2007). In this study the consensus criterion chosen was that 80 percent of participants' ratings for an item fell within the 6-to-7 range on the 7 point Likert scale (Ulschak, 1983; cited in Hsu & Sandford, 2007). Any items scoring below consensus cut-off were removed or modified to incorporate panel suggestions. New items were added – for inclusion in the subsequent round - if directly suggested. New items could also be developed by the Research Team and included in the subsequent round, if the panel suggested that an aspect of psychological flexibility was not covered. Suggestions for the structure and content of the scoring manual were also considered. This process resulted in a revised version of the ACT-FM, including the manual, which was input into the Round 2 survey.

Several weeks after completion of Round 1, an e-mail was sent containing a link to the Round 2 Delphi survey. Attached to this e-mail was a document containing the feedback from Round 1 of the Delphi process (panellists' aggregated scores and comments for each

item and comments on the overall manual, alongside Research Team rationale for their decisions). Following completion of the Round 2 Delphi survey, items were again ranked and panellists' comments considered. The Research Team then selected 3 items for each of the 8 sections of the ACT-FM.

Several weeks later the Round 3 Delphi Questionnaire was sent to participants via e-mail, and two documents were appended to this e-mail: a pdf version of the ACT-FM and, again, a document containing feedback on Round 2 of the Delphi process. Final comments and suggestions from the panel were also discussed and actioned by the Research Team.

Delphi results
Participants

Forty-seven individuals were invited to take part in Round 1. The response rates were as follows: 13/47 (28%) for Round 1, 10/13 (77%) for Round 2 and 9/13 (69%) for Round 3. Participants' demographics are presented in Table 1. Participants worked across many clinical specialities and therapy contexts: adult mental health (9), physical health (4), chronic pain (3), neuropsychology (3), supervision and training (3), psychosis (2), paediatrics (1), grief (1) and work and sport (1). The mean number of years of experience working with ACT was 11 (range 5 years – 23 years), and approximately half of the participants in each round were recognised by the ACBS as a Peer Reviewed ACT Trainer.

Round 1 results
Item results

It was notable that several items showed strong consensus among the panel (i.e. over 80% of panellists felt strongly that they should be included.) Supplementary Table 1 outlines panellist's ratings for each item, and the decision made by the research team regarding whether to keep, edit or delete. To illustrate the process of scoring and altering items, Table 2 shows the progress of one item through the rounds of the Delphi study.

Actions from Round 1

Some items and aspects of the manual were modified based on feedback. Key changes are described below.

1. Rating scale

The panel suggested that where no behaviours were observed a score of zero should be recorded, and that the rating scale descriptions could be made clearer. In response, the

rating scale was changed from 1-5 to 0-3, and the focus of this rating was altered from measuring the depth and extensiveness of the behaviour to rating how consistently the therapist enacted the behaviour.

2. Necessity for ACT inconsistent items

Several panellists questioned whether the inclusion of ACT inconsistent items was necessary. The Research Team referred to fidelity measurement literature, which highlights the importance of including inconsistent or 'proscribed' therapist behaviours to determine deviations from fidelity (e.g. Waltz et al., 1993; Plumb & Vilardaga, 2010). Thus, proscribed items were retained. Yet, in response, text was added to the manual explaining the utility of rating ACT inconsistent behaviours.

3. Scoring participant's response to therapist behaviour

The initial manual stated that the therapist's behaviour should be scored irrespective of how the client responds. However, two participants suggested that the client's response should also be considered. After discussion, the Research Team considered it particularly challenging for a listener/viewer to discern and rate client's response reliably and therefore these suggestions were not actioned.

4. Capturing functions of behaviour

A panellist suggested that, given that ACT is informed by Functional Contextualism, the measure should avoid making absolute statements about the fidelity of an observed behaviour without considering the function of the behaviour. Commenting on one item - 'Therapist does not lecture' - the panellist mentioned that there may be times when it is effective for the therapist to lecture. The Research Team noted that it is impractical and difficult to rate the function of several concurrent therapist behaviours. However, recognising the importance of the criticism, several changes were made to the measure. First, new items were proposed for the Therapist Stance section for inclusion in the Round 2 survey to try to capture therapist behaviours that enhance awareness of function, e.g. 'Therapist gives the client opportunities to notice the effectiveness of their behaviours (i.e. whether behaviours help/helped them to achieve results consistent with their values).' Second, the Research Team adapted the manual to clarify that the rater should bear the context of the therapy session in mind and consider the function of the therapist behaviour where possible.

Round 2 results
Item results

10

Panellists' ratings (Supplementary Table 2) and comments on the iteration of the fidelity measure emerging from Round 2 were again discussed by the Research Team. Three items for each area of the Tri-flex and Therapeutic Stance were then selected for inclusion in the ACT fidelity measure. For four out of the eight sections, the included items were those with the highest ratings. For the other four sections, some of the highest rated items had overlapping focus or content, which meant there was a lack of breadth of flexibility processes within the ACT-FM. Therefore, the next highest scoring item that assessed a different dimension was chosen (see Supplementary Table 2). For three of these areas, the fourth highest scoring item enabled coverage of sufficient breadth, and for one section a lower scoring item was selected.

Developments to the measure

One theme was apparent in panellist feedback, comprising concerns that a therapist could score highly on the resultant measure because responses are technically correct but not skilful, such as when they are not sensitive to what is happening in the room with their client. The Research Team noted that this comment overlapped with themes in the previous round (Scoring participant's response to therapist behaviours; Capturing functions of behaviour). As changes had been made previously, no changes were made on this occasion and panellists' attention was drawn to the previously adapted instructions to raters (i.e. that they should bear the context of the therapy session in mind, and consider the function of the therapist behaviour).

Round 3 results

Several themes were apparent in the feedback provided by the panel on the final version of the measure.

1. Impact of therapist behaviour on participant behaviour

Again, overlapping with previous comments, one participant suggested that the measure should focus on a participant's response to therapist methods and define therapist ACT consistency in terms of whether the therapist's behaviour engendered psychological flexibility in the client. In response to these comments, the wording in the manual was changed from "The clinician's behaviour should be scored irrespective of how the client responded to the clinician's attempt" to "The focus of this measure is on the therapists behaviour". The research team felt that this wording kept the focus on the therapist's behaviour but did not discount the rater considering participant response.

11

2. Workability and functional analysis

One participant commented on the need to include an item assessing workability and/or functional analysis. The Research Team noted that this was another consistent theme in the comments. A workability item was introduced in Round 2 in response to similar comments, but had not scored high enough to be included in the final measure. Data from Round 2 was reviewed and an alternative workability item was adapted from the ACT consistent Engaged section that had achieved consensus, with 90% of the panel rating it as a 6 or 7 (item 5 in Supplementary Table 2): "Therapist gives the client opportunities to notice workable and unworkable responses (i.e. whether their actions move them towards or away from their values)." After discussion, the Research Team included this item in place of an existing one that had greater overlap with items already included in the ACT-FM.

The developed ACT Fidelity Measure (ACT – FM).

The final measure consisted of four areas: therapist stance, open response style, aware response style and engaged response style, each split into ACT consistent and ACT inconsistent items, making 8 sections in all. Each item was rated on a 4 point rating scale, 0 (behaviour never occurred) to 3 (therapist consistently enacts this behaviour). The ACT-FM also included a brief manual comprising guidance on completion, and definitions of the psychological flexibility processes under observation.

Study 2: Field test method
Design

The Delphi methodology used in Study 1 sought to produce items and formats that would enable the measurement of key indicative and observable ACT consistent and inconsistent behaviours. To make further progress on this aim and to assess the preliminary reliability of the measure a field test was conducted. Here clinicians used the ACT-FM to rate the fidelity of a 20-minute ACT therapy video.

Participants

Opportunity sampling was used to recruit participants from two local ACT Special Interest Groups. We aimed to recruit a sample representative of one section of potential users: clinicians experienced enough in ACT to identify and use ACT principles but not necessarily experts. The inclusion criteria were: Clinicians who currently use ACT in their practice and have a minimum of three years' experience with ACT in a clinical and/or research capacity.

12

A minimum of 5 participant raters was required to achieve the statistical significance for an alpha-value set at 0.05 for an ICC with the minimum power of at least 80% for 24 observations per participant (Bujang & Baharum, 2017).

Materials

ACT video

A 20-minute video of an ACT session was included to be rated by clinicians using the ACT-FM. The therapy video was edited from several videos exemplifying the application of various ACT principles and techniques created by Dr David Gillanders (Clinical Psychologist and Peer Reviewed ACT Trainer). He models ACT consistent therapy sessions with another clinician who is role playing a client living with irritable bowel syndrome. We edited the clips in order to create one film to include examples of therapist behaviours from each part of the Tri-flex. The videos can be found here: https://vimeo.com/davidgillandersactvideos and we received permission to use the videos for the purpose of testing the fidelity measure. The edited video used for this research can be obtained by contacting the corresponding author.

Usability Questionnaire

A 5-item usability questionnaire was developed that required participants to answer five questions on a scale of 1 to 7 (where 1= not at all and 7= extremely): 1) How easy to understand was the fidelity measure? 2) How easy to use was the fidelity measure? 3) How easy to use was the response format? 4) Were any items difficult to understand? 5) Were any items difficult to rate?

Additionally, demographic information was gathered and space was provided for participants to write qualitative comments and suggestions. This usability questionnaire took approximately 5 minutes to complete.

Procedure

After consenting to take part and sharing their demographic information, clinicians were given time independently to become familiar with the ACT-FM manual and items before rating the same 20-minute ACT video. After rating, participants were invited to share their scores with a researcher and each other, and discrepancies were discussed, enabling the identification of ambiguous items and instructions that may be misunderstood. Participants then filled in the usability questionnaire.

Field test results

Participants

Nine clinicians were recruited from a range of service contexts: adult physical health (5), neurological conditions (4), adult mental health (2), paediatrics (2), adult pain (1) and older adult psychology (1). All participants were currently working clinically and had between 3 and 10 years' experience with ACT (mean= 4.7, SD= 2.19).

Inter-rater reliability

The obtained level of inter-rater agreement in rating the ACT video was moderate to excellent (Intra-class Correlations Coefficients ICC 2, 1 = 0.73, 95% CI, 0.60 - 0.93; Koo & Li, 2016)

Usability feedback

Ratings made by participants

Table 3 shows participants' ratings for usability. The scores suggested that clinicians found the ACT-FM reasonably acceptable and usable. However, lower scores on questions four and five suggested that some items were experienced as difficult to understand or rate. Participants comments on these problematic items were again considered by the Research Team, and the ACT-FM was modified as detailed below.

1. Clarify the instruction manual

Participants made several suggestions to improve usability of the instruction manual. These included: emphasising the rationale for taking notes during the session and for scoring at the end; clarifying that a therapist behaviour can be scored across multiple items rather than rating the one most suitable item, and clarifying that the ACT consistent and inconsistent items are not opposites of each other. One participant voiced a concern that a score may be affected by opportunity to perform behaviours if a session was short or at a particular stage in therapy (perhaps the initial or final session). In response, the ACT-FM was modified to include session length and number. While not formally used to alter scoring, this information can provide context for understanding the scores achieved in any given session.

2. Changes to specific items

Seven items were considered difficult to understand, and these items were modified to increase comprehension. For example, ambiguous language was changed: "Therapist helps the client to experience that they are bigger than or contain their psychological experiences" became "…bigger than and/or separate from their psychological experiences" to avoid

misinterpretation of the phrase 'contain' as referring to Bion's theory of 'Containment' (Bion, 1985).

Almost all participants found the following item difficult to rate: "Therapist uses experiential methods (e.g. exercises and metaphors) that are sensitive to the situation". This wording led to confusion about whether the focus was intended to be on techniques that are sensitive to the situation (i.e. in contrast to manualized or 'one size fits all' interventions), or on the therapist using experiential techniques (i.e. in contrast to didactic methods). After discussion, it was decided that both interpretations are valid and represent important aspects of ACT fidelity. Therefore, this item was separated into two items, which were both included within the ACT consistent therapist stance section of the ACT-FM.

Discussion

To improve the quality of clinical trials and enhance clinical practice, as a long-tern goal, we aimed to develop a new ACT fidelity measure that is concise in capturing key and observable therapist behaviours, practicable, reliable, and can function across multiple therapy contexts. The two-stage research process described in this study are the first steps towards these aims. The Delphi study resulted in a 24-item measure that was expanded to 25 items following the field study. The resultant measure is called the ACT Fidelity Measure (ACT-FM), attached.

To our knowledge, this is the first such tool for ACT to be developed through Delphi methodology, incorporating detailed feedback and suggestions from the potential users of the designated measure. Although, further research is required regarding the psychometric properties of the measure, we feel that this process was helpful in several ways. First, a sufficiently expert sample was recruited, and the method enabled us to derive a small number of items achieving group consensus for capturing observable actions indicative of ACT consistent and inconsistent therapist behaviours. Second, this sample was diverse, including experts from many different therapy contexts. Although this can only be assessed comprehensively once the measure has been trialled in several therapy contexts, we hope this will increase the likelihood that the behaviours assessed by the ACT-FM are measurable across contexts in which ACT might be used. Third, the field study suggested that the ACT-FM may be reasonably clear for use by a less expert group (also a target group of potential end-users of the scale), and the initial exploratory assessment of the tool's reliability was encouraging.

15

Delphi study

As a result of the Delphi process, all of the 24 items included in the final ACT-FM differed from those in the initial draft that was used to instigate Round 1, and several aspects of the manual also were changed. Nonetheless, the process of developing the ACT-FM was challenging, as it required the research team to make difficult decisions. We were required to balance Delphi panellist suggestions, with guidelines for developing fidelity measures, and the *a priori* purpose and scope of the ACT-FM (concise and observable, tri-flex structure, etc.) This meant that some suggested changes were only partially actioned. For example, the suggestion that ACT inconsistent items are not necessary is inconsistent with the aim of creating a robust fidelity measure (e.g. Waltz et al., 1993; Plumb & Vilardaga, 2010).

Perhaps the most challenging comments were those that questioned the consistency of some items or concepts with the philosophy of functional contextual that underlies ACT. Several changes were made to the measure to accommodate these comments. However, the Delphi study draws attention to a tension between the necessity to class behaviours as consistent/inconsistent within fidelity measures - an exercise that relies heavily on the topography of behavior - and Functional Contextualism, which emphasises the function and influences exerted on behaviours as opposed to their form. In trying to address this discrepancy, we considered the function of the ACT-FM. The *a priori* stated goals were to develop a parsimonious measure that would allow assessment of ACT fidelity via observation. Because consistent and accurate inferring of function appears beyond the reach of passive observation alone, the ACT-FM is what it is, a measure that relies largely on the form of therapist responses. So, while ideally ACT therapy methods ought to be defined functionally - as therapist responses that enhance psychological flexibility - the ACT-FM is not a pure measure of function. It could be argued that this issue is not uncommon among the frequently used process and fidelity measures currently used in ACT research. For example, the most widely used measure of psychological flexibility, the Acceptance and Action Questionnaire II (Bond et al., 2011) includes items such as "I'm afraid of my feelings" and "Emotions cause problems in my life" that appear to lack a granular focus on the context or function of behaviours. While this is not ideal, these forms are derived from principles, experience and data: In this example, the items of the AAQ-II are meant to reflect behaviours serving experiential avoidance functions, and they appear to improve via manipulation of PF within trials of ACT (albeit as part of an overall change in the AAQ II; e.g. Hann & McCracken, 2014; A-Tjak et al., 2015.)

We openly acknowledge these issues and limitations and, given the goals for the measure, we hope that the ACT-FM can continue to improve over time based on experience in practice and evidence. Modifications, further testing, and improvements are invited and welcomed; particularly additions that enable greater sensitivity to function of therapist behavior.

Field test

The inter-rater reliability of the ACT-FM was moderate to excellent. Nonetheless, subsequent feedback regarding ambiguous wording led to alterations of several items. Also, a decision was made to split one of the items into two, meaning that one section has four items, whereas the other sections only have three. Ultimately, here we chose to prioritise the practicality and utility of the measure over a too-rigid adherence to a predefined structure.

Limitations and future research

We acknowledge that the two studies included here represent limited progress toward the goal of developing of a concise ACT fidelity measure that can reliably measure ACT fidelity across various contexts. Much, therefore, remains to be done. For example, it would be useful to repeat inter-rater reliability investigations with a larger set of data with more variance, such as from multiple therapy sessions, therapy contexts, therapists and clients. As we only used a 20 minute video of a therapy session it is possible that not all Tri-flex processes were adequately covered. Further, given the content of the video, it is less likely to have enabled assessment of ACT inconsistent behaviours. It is important to state that, while we have integrated feedback from clinicians working in multiple contexts on what behaviours are ACT (in)consistent and observable, the ACT-FM has not yet been used in multiple contexts. This means that utility has not been demonstrated: we do not yet know if it is reliable beyond the field test used here (video), or whether it leads to improvements in the uptake or quality of ACT delivery or training.

Some of the solutions for increasing the function and context sensitivity of the ACT-FM may be sub-optimal. For example, instructions in the manual for raters to "use your clinical judgement when scoring, bearing in mind the context of the therapy session and considering the function of the therapist behaviour" may introduce error in the form of inconsistency between raters: those more experienced with ACT may take into account different factors than those less experienced.

In contrast to clinical trials where raters are often trained in using the fidelity measure prior to application, in the field test raters received no training. This was a conscious decision, so as to assess how the measure performs with minimal instruction. While encouraging reliability statistics were apparent without training, this may have served to reduce the inter-rater reliability. We expect that greater inter-rater reliability would be achieved if coders are trained - including a re-calibration process - prior to implementation.

Indeed, the field study was conducted using a therapy video. Although the measure is designed to be used for both audio and video recordings, it has not yet been tested for its usability with audio only. It is unknown how much the clinicians took non-verbal factors into account when scoring and it is unknown if the same inter-rater reliability result would be achieved with audio-only.

For an RCT it is helpful to know what constitutes high fidelity in order to be able to state that the therapy being used in the trial met this condition. Currently it is unknown what score would indicate high or low therapist fidelity to ACT. Of course ultimately evidence ought to show the relationship between fidelity, treatment process changes, and treatment outcome. Further research could focus on gaining normative data from a large pool of therapists, raters and settings to begin to grade the quality of ACT delivery using the ACT-FM and then proceed to address these other important questions.

Finally, it is important to acknowledge that the small research team, including members variously experienced with ACT, were not a passive part of developing the measure, and made decisions regarding what suggestions were and were not actioned and how these might be actioned. So, while the Delphi consultation process appeared helpful in producing a context for the democratic expression of expert opinion, this likely only partially limited the disproportionate influence of a few individuals on the emergent initial version of the ACT-FM that is presented here.

Conclusions

A 25-item fidelity measure was co-developed via consultation with an expert panel in a Delphi study and then preliminary tested in a field study. Initial feedback on the reliability and practicability of the measure was promising. However, work is required to assess comprehensively the utility and psychometric properties of the ACT-FM across contexts. Further careful tailoring of the ACT-FM is likely to be necessary, and is welcomed.

References

A-Tjak, J. G. L., Davis, M. L., Morina, N., Powers, M. B., Smits, J. A. J., & Emmelkamp, P. M. G. (2015). A meta-analysis of the efficacy of acceptance and commitment therapy for clinically relevant mental and physical health problems. *Psychotherapy and Psychosomatics*, 84(1), 30-36.

Association for Contextual Behavioural Science (2018, March) *State of the ACT evidence*. Retrieved from https://contextualscience.org/ACT_Randomized_Controlled_Trials

Atkins, P. W. B., Ciarrochi, J., Gaudiano, B. A., Bricker, J. B., Donald, J., Rovner, G., Smout, M., Livheim, F., Lundgren, T., & Hayes, S. C. (2017). Departing from the essential features of a high quality systematic review of psychotherapy: A response to Öst (2014) and recommendations for improvement. *Behaviour Research and Therapy*, 97, 259-272.

Bion, W. R. (1985). Container and contained. *Group relations reader*, 2(8), 127-133.

Bond, G. R., Becker, D. R., & Drake, R. E. (2011). Measurement of fidelity of implementation of evidence-based practices: Case example of the IPS Fidelity Scale. *Clinical Psychology: Science and Practice,* 18, 125–140.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., ... & Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire–II: A revised measure of psychological inflexibility and experiential avoidance. *Behavior therapy*, 42(4), 676-688.

Bujang, M. A., & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. *Archives of Orofacial Science*, 12(1).

Dalkey, N. & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management science, 9(*3), 458–467.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship*, 39(2), 155-164.

Graham, C. D., Gouick, J., Krahé, C., & Gillanders, D. (2016). A systematic review of the use of Acceptance and Commitment Therapy (ACT) in chronic disease and long-term conditions. *Clinical psychology review*, 46, 46-58.

Gaudiano, B. A. (2009). Öst's (2008) methodological comparison of clinical trials of acceptance and commitment therapy versus cognitive behavior therapy: Matching Apples with Oranges?. *Behaviour Research and Therapy*, 47(12), 1066-1070.

Hann, K. E. J., & McCracken, L. M. (2014). A systematic review of randomized controlled trials of Acceptance and Commitment Therapy for adults with chronic pain: Outcome domains, design quality, and efficacy. *Journal of Contextual Behavioral Science*, 3(4), 217-227.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, 57(2), 98-107.

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual behavioral science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science*, 1(1-2), 1-16.

Hayes, S. C., Hayes, L. J., Reese, H. W., & Sarbin, T. R. (Eds.). (1993). Varieties of scientific contextualism. Reno, NV: Context Press.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour research and therapy*, 44(1), 1-25.

Hayes, S. C., Strosahl, K., & Wilson, K. G. (2012). *Acceptance and commitment therapy: The process and practice of mindful change* (2nd ed.). New York, NY: Guilford Press.

Hsu, C. & Sandford, B. (2007). The delphi technique: making sense of consensus. *Practical Assessment, Research and Evaluation,* 12(10), 1-8.

IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review*, *30*(7), 865–878. https://doi.org/10.1016/j.cpr.2010.03.001

Linstone, H. A., & Turoff, M. (2011). Delphi: A brief look backward and forward. *Technological Forecasting and Social Change*, *78*(9), 1712-1719.

Long, D. M. (2013). Pragmatism, realism, and psychology: Understanding theory selection criteria. *Journal of Contextual Behavioral Science*, *2*(3-4), 61-67

Luoma, J. B., Hayes, S. C., & Walser, R. D. (2017). *Learning ACT: An acceptance and commitment therapy skills-training manual for therapists*. Second Edition. New Harbinger Publications.

McCracken, L. M., & Morley, S. (2014). The psychological flexibility model: a basis for integration and progress in psychological approaches to chronic pain management. *J Pain*, *15*(3), 221–234. https://doi.org/10.1016/j.jpain.2013.10.014

McGrath, K. B., & Forman, E. M. (2012). *Validation of the Drexel University ACT/tCBT Adherence and Competence Rating Scale: Revised for Use in a Clinical Population*. Drexel University.

Öst, L. G. (2008). Efficacy of the third wave of behavioral therapies: A systematic review and meta-analysis. *Behaviour research and therapy*, *46*(3), 296-321.

Öst, L. G. (2014). The efficacy of Acceptance and Commitment Therapy: An updated systematic review and meta-analysis. *Behaviour Research and Therapy*, 61, 105–121.

Plumb, J. C., & Vilardaga, R. (2010). Assessing treatment integrity in acceptance and commitment therapy: Strategies and suggestions. *International Journal of Behavioral Consultation and Therapy*, *6*(3), 263.

Pollard, S. (2010) Development and psychometric testing of the Acceptance and commitment therapy for psychosis adherence and competence scale. *Unpublished thesis*, La Trobe University, Melbourne, Australia

Powell, C. (2003). The Delphi technique: myths and realities. *Journal of advanced nursing*, *41*(4), 376-382

Roth, A., Fonagy, P. (2006). *What works for whom?: a critical review of psychotherapy research*. Guilford Press.

Rowe, G. Wright, G. (2001) Expert opinions in forecasting: the role of the Delphi technique, in: J.S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners,* Kluwer Academic Publishers, Boston et al., p. 125–144.

Ruiz, F. J. (2010). A review of Acceptance and Commitment Therapy (ACT) empirical evidence: Correlational, experimental psychopathology, component and outcome studies. *International Journal of Psychology and Psychological Therapy*, *10*(1).

Salavati, M., Waninge, A., Rameckers, E. A. A., van der Steen, J., Krijnen, W. P., van der Schans, C. P., & Steenbergen, B. (2017). Development and face validity of a cerebral visual impairment motor questionnaire for children with cerebral palsy. *Child: care, health and development*, *43*(1), 37-47.

Twohig, M. P., Hayes, S. C., Plumb, J. C., Pruitt, L. D., Collins, A. B., Hazlett-Stevens, H., & Woidneck, M. R. (2010). A randomized clinical trial of acceptance and commitment therapy vs progressive relaxation training for obsessive compulsive disorder. *Journal of Consulting and Clinical Psychology*, *78,* 705−716.

Ulschak, F. L. (1983). *Human resource development: The theory and practice of need assessment*. Reston, VA: Reston Publishing Company, Inc.

University of Bristol (2009). Bristol online surveys from http://www.survey.bris.ac.uk

Veehof, M. M., Trompetter, H. R., Bohlmeijer, E. T., & Schreurs, K. M. G. (2016). Acceptance- and mindfulness-based interventions for the treatment of chronic pain: a meta-analytic review. *Cognitive Behaviour Therapy*, 45(1), 5-31.

Von der Gracht, H. (2012). Consensus measurement in Delphi studies: review and implications for future quality assurance. *Technological forecasting and social change*, *79*(8), 1525-1536.

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

Wilson, K. G. (2001). Some notes on theoretical constructs: Types and validation from a contextual-behavioral perspective. *International Journal of Psychology and Psychological Therapy*, *1*, 205–215.

Table 1. Panelists' demographics across the three rounds of the Delphi Study

		Round 1	Round 2	Round 3
Number of participants		13	10	9
Male, female		10, 3	7, 3	7, 2
Continent of Residence	UK	8 (61.5%)	7 (70%)	6 (66.7%)
	The rest of Europe	2 (15.4%)	1 (10%)	1 (11.1%)
	North and South America	3 (23.1%)	2 (20%)	2 (22.2%)
Years of experience with ACT	Range	5 - 23	5 - 13	5 - 14
	Mean	11.3	10.8	11.2
	SD	4.2	2.6	2.8
Number Peer Reviewed ACT Trainer by the ACBS	Recognised	6 (46.2%)	5 (50%)	5 (56%)
	Not recognised	7 (53.8%)	5 (50%)	4 (44%)
Type of Work	Mainly Clinical	4 (30.8%)	4 (40%)	3 (37.5%)
	Mainly Research	6 (46.2%)	5 (50%)	3 (37.5%)
	Clinical and Research Equally	3 (23.1%)	1 (10%)	2 (25%)

Table 2. An example journey of an item through all stages of this study

Stage	Item	Feedback
Initial item proposed in Round 1	"Therapist lectures the client.*"	54% rated as 6 or 7, three participants commented on needing to consider function, as giving psycho-education could be useful and might sometimes be perceived as lecturing. Two suggested to overcome this by defining 'lectures' with a suggestion of adding that the therapist gives prolonged explanations or suggests the client is incorrect.
Updated item proposed in Round 2	"Therapist lectures the client (e.g. gives prolonged advice and/or explanations)."	78% of the panel rated as a 6 or 7, it was selected for the measure as the third highest scoring item. No feedback from the panel to improve this item.
Same item presented in Round 3 in ACT-FM format	As above	One participant commented that they agreed that the therapist shouldn't lecture the client, but that prolonged explanations are not the same as some things do require a lot of verbal unpacking. We therefore took out the explanations part and replaced it with the therapist tries to convince the client.
Item presented in field study	"Therapist lectures the client (e.g. gives prolonged advice or tries to convince the client)."	Two participants commented that the word "prolonged" made the item confusing as it could be misinterpreted to be about talking too much rather than in the way in which the therapist was talking. This was therefore taken out.
Final item	"Therapist lectures the client (e.g. gives advice, tries to convince the client, etc)."	

*Item from "ACT Inconsistent Therapist Stance" section

Table 3. Participants' scores on the usability questionnaire.

Usability question*	Mean	SD
a) How easy to understand was this fidelity measure?	5.22	1.20
b) How easy to use was this fidelity measure?	4.78	1.39
c) How easy to use was the response format?	5.11	1.05
d) Were any items particularly difficult to understand?	3.33	1.66
e) Were any items particularly difficult to rate?	3.33	1.12

*1-7 Likert scale, where 1=not at all and 7=extremely