Measuring the longitudinal course of voice hearing under psychological interventions: a systematic review

Article  (Published Version)


This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/108572/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher’s version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk
Clinical Psychology Review 97 (2022) 102191

Contents lists available at ScienceDirect
Clinical Psychology Review
journal homepage: www.elsevier.com/locate/clinpsychrev

Measuring the longitudinal course of voice hearing under psychological interventions: A systematic review

Sofia Loizou a,*, David Fowler a, Mark Hayward a,b

a School of Psychology, University of Sussex, Falmer BN1 9QH, UK
b Research & Development Department, Sussex Partnership NHS Foundation Trust, Nevill Avenue, Hove BN7 3HZ, UK

ARTICLE INFO
Keywords:
Voice hearing
Auditory verbal hallucinations
Psychological interventions
Longitudinal course

ABSTRACT
Trials of psychological interventions targeting distressing voices have used a range of variables to measure outcomes. This has complicated attempts to compare outcomes across trials and to evaluate the effectiveness of these interventions. Therefore, this review aimed to identify the variables that have been used to measure the longitudinal course and impact of voice hearing under these interventions and to evaluate how these variables change over time. Inclusion and exclusion criteria were applied, resulting in a total of 66 articles. Of these, 60 studies (28 RCTs, 23 uncontrolled, 9 non-randomised) were published in peer-reviewed journals, whilst 6 were recently completed or currently ongoing. The findings of this review suggest that a range of variables that are not directly relevant to psychological interventions have been used (e.g., depression, characteristics of voice hearing experience), whilst those directly impacted by psychological interventions (e.g., voice-related distress), broader concepts of outcome (e.g., functioning) and specific associated processes (e.g., self-schema) have received less attention. Findings also showed that the majority of variables demonstrated improvements, but effect sizes varied considerably across trials. This may be attributed to methodological differences such as statistical power, binding, control groups and different methods of measurement. Our review highlights the importance of determining a set of outcomes that are directly targeted and should change under psychological interventions. Recommendations include the use of voice-related distress as a primary outcome. This can ultimately facilitate comparisons across studies and inform the development of psychological interventions.

1. Introduction
Voice hearing can be described as the perceptual experience of hearing voices in the absence of external stimuli (American Psychiatric Association, 1994; Beck & Rector, 2003). This is comparatively common in the general population with no negative impact (Beevan, Read, & Cartwright, 2011; Johns, 2005; Johns et al., 2014; Johns & van Os, 2001; Moritz & Larøi, 2008). However, some voice hearers can experience distress and poor functioning, consequently needing care (Johns et al., 2014; Larøi, 2012; Larøi et al., 2012). A central factor that appears to distinguish voice hearers with and without the need for care is negative voice content (Larøi, 2012). Some voice hearers can experience critical voices (e.g., ‘you can’t do anything right’), which has been linked to increased risk of suicide (Kjelby et al., 2015; Larøi, 2012; Nayani & David, 1996). On some occasions, voices can give instructions which individuals may comply with to minimise distress (Byrne, Birchwood, Trower, & P.E., 2007). The severity of these instructions can range from benign orders (e.g., ‘get the milk’) to commands to harm (e.g., physically injuring one-self or others) (Nayani & David, 1996).

The cognitive model of distressing voices has been developed to illustrate why some individuals who hear voices do not experience distress whilst others do (Birchwood et al., 2004; Birchwood, Meaden, Trower, Gilbert, & Plaistow, 2000; Chadwick & Birchwood, 1994). The model proposes that appraisals of voice power, intent and identity can lead to emotional, physiological and behavioural responses to voices and particularly distress. Schemata (e.g., ‘I am weak’) influenced by early life experiences can also play an important role in shaping the beliefs people hold about themselves (e.g., low self-esteem, of low social rank) and their voices (e.g., voices are powerful). This has been corroborated by a plethora of research, showing that voice hearers who perceive themselves in a negative light (e.g., inferior), also tend to view voices as powerful, giving rise to distress (Birchwood et al., 2000; Birchwood et al., 2004; Gilbert et al., 2001; Paulik, 2012). It has been further suggested that safety behaviours including unhelpful coping

* Corresponding author at: School of Psychology, Pevensey 1 Building, University of Sussex, Falmer BN1 9QH, UK.
E-mail address: s.loizou@sussex.ac.uk (S. Loizou).

https://doi.org/10.1016/j.cpr.2022.102191
Received 23 November 2021; Received in revised form 10 June 2022; Accepted 4 August 2022
Available online 10 August 2022
0272-7358/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
styles, compliance and appeasement, are often used to mitigate perceived threats, which prevents the disconfirmation of negative voice-related appraisals, hence maintaining distress (Byrne et al., 2007; Chadwick & Birchwood, 1994; Griffiths, Michail, & Birchwood, 2012; Morrison, 1998). The cognitive model has been particularly influential in the development of psychological interventions for distressing voices, including Cognitive Behavioural Therapy for psychosis (CBTp), which involves assisting people to understand their psychotic experiences and to modify unhelpful beliefs associated with distress and disability (Johns, Isham, & Manser, 2020; Morrison & Barratt, 2010).

Cognitive Behavioural Therapy for psychosis (CBTp) has been recommended as a first-line intervention and as an adjunct to pharmacotherapy for the treatment of psychotic experiences including distressing voices (National Institute for Health and Care Excellence, 2014). While its effectiveness has been supported by a body of research, CBTp has only been found to have small to medium effects (Hazell, Hayward, Cavanagh, & Strauss, 2016; Jauhar et al., 2014; Turner, Burger, Smit, Valmaggia, & van der Gaag, 2020; van der Gaag, Valmaggia, & Smit, 2014; Wykes, Steel, Everitt, & Tarrier, 2008; Zimmermann, Favrod, Trieu, & Pomini, 2005). This has been partly attributed to the use of broad psychotic symptom severity measures that fail to sufficiently assess outcomes directly impacted by CBTp (e.g., distress), subsequently influencing the magnitude of effect sizes reported in trials (Birchwood, Shiers, & Smith, 2014; Peters, 2014; Thomas, 2015b; Thomas et al., 2014; van der Gaag et al., 2014). As a result, symptom-specific psychological approaches have been developed and evaluated to improve outcomes (Lincoln & Peters, 2019; Thomas et al., 2014). This has also led to trials moving away from broad psychotic symptomatology outcomes and focussing instead on outcomes relating to the characteristics of the voices, the characteristics of the voice hearer and associated processes.

The characteristics of the voices are the variables that have been used to assess and describe the voices, such as their frequency (McLeod, Morris, Birchwood and Dovey, 2007b). These are not targeted by psychological interventions and therefore unlikely to change. The characteristics of the voice hearer, on the other hand, refer to variables that have been used to assess and describe how voices are experienced by the voice hearer. These include the emotional impact or problem behaviour in relation to voices (e.g., distress; Hayward, Jones, Bogen-Johnston, Thomas, & Strauss, 2017, compliance; Trower et al., 2004), which are thought to be more relevant and have been recommended as primary outcomes in trials of psychological interventions (Birchwood & Trower, 2006), as well as broader concepts of outcome, focussing on the impact of voices on broad life domains (functioning; Wykes et al., 2005, distress; Chadwick et al., 2016). Associated processes are also experienced by the voice hearer and have been proposed to play a role in the development and maintenance of distressing voices, such as depressed mood, anxiety, low self-esteem (Close & Garety, 1998; Freeman & Garety, 2003; Garety, Kuipers, Fowler, Freeman, & Bebbington, 2001), dissociation (Longden, Madil, & Waterman, 2012), beliefs about the voices (Chadwick & Birchwood, 1994) and negative relating (Hayward, Berry, McCarthy-Jones, Strauss, & Thomas, 2014). Although their importance has been supported by previous research (e.g., Birchwood et al., 2004; Fannon et al., 2009; Hayward, Denney, Vaughan, & Fowler, 2008; Smith et al., 2006), these processes have not yet been the main focus in previous trials and therefore our understanding of the process of change needs to evolve, ultimately contributing to the development of more targeted interventions (Thomas et al., 2014).

The heterogeneity of the aforementioned variables suggests that there are differing views as to what the outcomes of these interventions should be. Alongside this, a variety of measures have been used to assess these outcomes. For example, the impact of voices has been assessed using measures of voice-related distress (e.g., Psychotic Symptoms Rating Scale for Auditory Hallucinations [PSYRATS-AH; Haddock, McCarron, Tarrier, & Faragher, 1999]), anxiety and depression (e.g., Depression Anxiety Stress Scale-21 [DASS-21; Henry & Crawford, 2005]). The range of variables and measures used in these trials has hindered efforts to compare outcomes and to evaluate these psychological interventions (Thomas, 2015a; Thomas et al., 2014). Hence, there is a need to clarify outcomes and their measurement, as this will enable trials to be more consistent, which can inform psychological interventions for distressing voices.

It is therefore important to identify the variables that have been used to measure the longitudinal course and impact of voice hearing and examine whether these variables remain constant or vary over time under psychological interventions. This can provide an insight as to how, when and what is changing and comparing this to what we are expecting to change, which can inform the development and targeting of psychological interventions. To our knowledge, no review thus far has focused on the selection and measurement of variables that have been used to examine the longitudinal course and impact of voice hearing under psychological interventions. The aims of this review are as follows:

1. To identify the psychological variables that have been used in studies of psychological interventions to measure the longitudinal course and impact of voice hearing.
2. To determine how these variables relate to the characteristics of the voices.
3. To determine how these variables relate to the characteristics of the voice hearer.
4. To determine how these variables relate to proposed processes that can maintain problematic voice hearing experiences.
5. To determine how these variables change over time under psychological interventions.

2. Methods

This review was pre-registered on Prospero (ref. number: CRD42020182578).

2.1. Searches

PsycINFO and PubMed were individually searched on the 28th of April 2022 based on titles, abstracts and keywords using the following search string: (“trial” OR “treat” OR “intervention” OR “therapy” OR “training”) AND (“auditory hallucination” OR “voice hear”)

The International Standard Randomised Controlled Trials Number (ISRCTN) registry was also searched to identify ongoing or recently completed trials that have not yet been published. The terms “hallucinations” or “voice hearing” were employed when searching the registry. Following database searching, reference lists of book chapters, reviews, meta-analyses and relevant papers were thoroughly examined with the aim of identifying articles not found in the original searches.

2.2. Inclusion and exclusion criteria

2.2.1. Sample

We included studies reporting on individuals experiencing voices. There were no restrictions on diagnosis and mental health settings. Studies involving participants with at-risk mental state, non-clinical or subclinical levels of voice hearing were excluded.

2.2.2. Studies

Quantitative studies investigating voice hearing over time under psychological intervention either ongoing, recently completed or published in peer-reviewed journals were included. These consisted of: 1) Randomised Controlled Trials (RCT); 2) non-randomised studies; and 3) descriptive studies, provided they incorporated at least two time points of assessment (i.e., pre- and post-intervention, follow-ups). We excluded reviews, meta-analyses, therapy manuals, guides, narratives,
2.3. Interventions

Psychological interventions for distressing voices were deemed appropriate if they attempted to either: 1) make links between life experiences, thoughts, feelings or behaviours and voice hearing; 2) evaluate appraisals and/or change behaviour linked to voices; 3) promote the monitoring of thoughts, emotions or behaviours related to voices; 4) implement effective coping strategies for the management of voices; 5) decrease emotional and/or behavioural responses to voice hearing; 6) enhance well-being and/or functioning; 7) increase compassion, positive memories, resilience, awareness and/or acceptance related to voices; or 8) change social relating/relating to voices. There was no restriction on the format of these interventions in terms of: 1) group/individual delivery; 2) online-face-to-face delivery; 3) type of setting; 4) level of care; 5) level of therapist contact; 6) duration of intervention. Additionally, we included any type of comparison groups ranging from treatment as usual (TAU), waiting list, to any active control. We also excluded non-psychological interventions for voices.

2.4. Study selection

The first stage consisted of removing all duplicates using Endnote. Following that, the remaining articles were imported into Rayyan QCR to manage the review process. A reviewer initially screened all articles based on titles and abstracts and then inspected the full text of the remainder. We applied inclusion and exclusion criteria at all points and documented reasons for exclusion. Inter-rater reliability (99.4% agreement) was established as a second reviewer independently screened a random sample of papers (10%) resulting from the searches. A third reviewer was introduced in instances of disagreement. Finally, the reference lists of the eligible papers along with those of relevant book chapters and reviews were searched to locate articles we may have missed. Fig. 1 shows the selection process.

2.5. Data extraction

Following study selection, the author extracted the study design, sample characteristics, outcomes and measures of the identified articles. The findings of RCTs were also extracted. In the event that RCTs did not report effect sizes, these were calculated by the first author (Cohen’s d).

2.6. Risk of bias (quality) assessment

The quality of the studies was assessed using the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018). The MMAT covers five domains: qualitative, RCTs, non-randomised, descriptive and mixed methods. Each domain has five different criteria that are rated as either ‘Yes’, ‘No’ or ‘Can’t tell’ (See Table 1). A total number is calculated based on the number of criteria met, ranging from zero (low quality) to five (high quality).

2.7. Synthesis of evidence

We followed a qualitative approach to data synthesis. To address questions one to four, we first identified all the variables that had been used in published studies of psychological interventions. We then grouped these variables based on themes (related to shared commonalities). As for recently completed or ongoing studies, we reported the measures that were used. Finally, to understand how these variables changed over time, we reported the standardised effect sizes (Cohen’s d) of RCTs. The following equation was used:

$$d = \frac{\bar{x}_{\text{intervention}} - \bar{x}_{\text{control}}}{S_p}$$

Cohen’s d (Cohen, 1988) was interpreted as small (d = 0.2), medium (d = 0.5) and large (d = 0.8). In instances where means and standard deviations were not reported, but the t value and degrees of freedom, or odds ratio were reported, we calculated Cohen’s d effect sizes using this website: http://www.psychometrica.de/effect_size.html. We also converted Eta squared effect sizes to Cohen’s d.

3. Results

Sixty-four articles that were identified through database searches met the eligibility criteria. A further two studies meeting the inclusion criteria were identified through other sources, resulting in a total of 66 articles (60 published in peer review journals, three recently completed and three ongoing). Details of these studies can be seen in Table S1 and Table S2 (see Supplementary Material).

Of the 60 published articles, nine were non-randomised controlled, 23 were uncontrolled and 28 were RCTs. These focussed on the following interventions: behavioural/symptom management; Coping Strategy Enhancement (CSE), Cognitive Behavioural Therapy (CBT), Coping with Voices, Avatar Therapy, Virtual Reality Therapy (VRT), Cognitive Therapy for Command Hallucinations (CTCH), Acceptance and Commitment Therapy (ACT), Person-Based Cognitive Therapy (PBCT), Mindfulness-Based Cognitive Therapy (MBCT), mindfulness, Hallucination focused Integrative Treatment (HIT), Relating Therapy (RT), Competitive Memory Training (COMET), Imagery Rescripting, Imaginal exposure, Psychoeducation, Cognitive Nursing Intervention (CNI) and Experience Focussed Counselling (EFC). Three of the 28 RCT articles appeared to be from the same HIT trial (Jenner, Nienhuis, van de Willige, & Wiersma, 2006; Jenner, Nienhuis, Wiersma, & van de Willige, 2004; Wiersma, Jenner, Nienhuis, & van de Willige, G., 2004) and two from the same CNI trial (England, 2007, 2008). Four uncontrolled studies used data from the Sussex Voices Clinic, but the timeframe of the data collection differed and therefore samples were not identical (Clarke, Jones, & Hayward, 2021; Hayward, Edgecumbe, Jones, Berry, & Strauss, 2018; Hayward, Frost, Naito, & Jones, 2022; Morrice, Jones, Burgio, Strauss, & Hayward, 2021). A further 2 uncontrolled studies used data from the same HIT programme (Jenner, van de Willige, & Wiersma, 1998; Wiersma, Jenner, Van de Willige, Spakman, & Nienhuis, 2001) and two from the same Imagery Rescripting case series (Paulik, Newman-Taylor, Steel, & Arntz, 2020; Paulik, Steel, & Arntz, 2019). The majority of the RCTs included participants of White identity, with a diagnosis of schizophrenia-spectrum or psychotic-spectrum disorders, had at least one follow-up point (three months to two years) and had used a non-active control group. Sample sizes of RCTs ranged from 19 to 197. In the following sections we provide a summary of the variables that have been used to assess the longitudinal course and impact of voice-hearing under psychological interventions and how these change over time. Table S2 (see Supplementary Material) shows the methodological quality of the studies.

3.1. Characteristics of voices

Although the characteristics of voices are not targeted by psychological interventions, a number of these characteristics have been extensively measured in the included studies. One of the most prevalent variables is the ‘voice hearing experience’, which has been measured in 35 studies. This has been mostly assessed using the total score of the PSYRATS-AH (Haddock et al., 1999), capturing several aspects of the voice hearing experience including the characteristics (e.g., location, loudness), appraisals (e.g., controllability, beliefs regarding origin) and impact of voices (e.g., amount of distress, disruption). In fact, the total score of the PSYRATS-AH is currently the most frequently used primary...
Fig. 1. Flowchart of the selection process.
outcome in studies of psychological interventions that have pre-defined outcomes including RCTs (Bell et al., 2020; Craig et al., 2018; Gottlieb et al., 2017; Leff, Williams, Huckvale, Arbuthnot, & Leff, 2013; Penn et al., 2009; Schnackenberg, Fleming, & Martin, 2017; Wykes et al., 2005), uncontrolled (Brand, Bendall, Hardy, Rossell, & Thomas, 2020; Dellazizzo, Potvin, Phraxayavong, & Dumais, 2020; Dodgson et al., 2021; Gottlieb, Romeo, Penn, Mueser, & Chiko, 2013; Varese et al., 2020) and non-randomised controlled studies (Kay et al., 2021; Newton et al., 2005; Wykes, Parr, & Landau, 1999). The sum of items, however, may conceal therapeutic effects, as we would expect some variables (impact of voices e.g., distress) to change more than others (characteristics of voices, e.g., location) under psychological interventions. This was followed by the frequency of voices, which has been measured in 24 studies. Variables that have been examined less consistently include hallucination severity, negative content, loudness, location, duration, clarity, tone, distractibility, reality, number of voices, the severity of voice-related distress either as a secondary outcome (Brand et al., 2020; Chadwick et al., 2016; Craig et al., 2018; Dellazizzo et al., 2020; Dodgson et al., 2021; Gottlieb et al., 2013; Gottlieb et al., 2017; Hayward et al., 2017; Hayward et al., 2018; Hayward et al., 2021; Hayward et al., 2022; Hazell et al., 2018; Jones et al., 2021; Louise et al., 2019; Newton et al., 2005; Paulik et al., 2019; Paulik, Jones, & Hay ward, 2018; Penn et al., 2009; Trower et al., 2004; Wykes et al., 1999; Yang, Lee, Lo, & Beckstead, 2015), whereas the remaining studies did not pre-define outcomes. Competitive memory training (COMET) was the only intervention that explicitly aimed to reduce depression through strengthening memories linked to positive self-esteem (van der Gaag et al., 2012). Moreover, anxiety has been examined in 24 studies of psychological interventions, the majority of which have used this as a secondary outcome (Brand et al., 2020; Chadwick et al., 2016; Craig et al., 2018; Dodgson et al., 2021; Gottlieb et al., 2013; Gottlieb et al., 2017; Hayward et al., 2017; Hayward et al., 2018; Hayward et al., 2021; Hayward et al., 2022; Jones, Straus, & Hayward, 2021; Morrice et al., 2021; Paulik et al., 2019; Paulik et al., 2020) have used this as a primary outcome. The remaining 22 studies have used voice-related distress either as a secondary outcome (e.g., Craig et al., 2018) or have not pre-classified outcomes (e.g., Haddock, Slade, Bentall, Reid, & Faragher, 1998). While the majority of studies have used the voice-related distress subscale of the PSYRATS-AH (Haddock et al., 1999; Woodward et al., 2014), some have used other instruments (e.g., Hamilton Program for Schizophrenia Voices Questionnaire [HPSVQ; Van Lieshout & Goldberg, 2007], Delusion and Voices Self-Assessment [DV-SA; Pinto, Gigantesco, Morosini, & La Pia, 2007]). Different methods of scoring have also been introduced, such as individual items that may be less sensitive to the detection of change (e.g., Personal Questionnaire Rapid Scaling [PQRST ‘distress’ item; Mulhall, 1978], PSYRATS-AH ‘amount of voice-related distress’ item; Haddock et al., 1999). Less emphasis has been placed on engagement, resistance, pre-occupation, compliance to command hallucinations (CHs) and the subjective negative impact of psychotic experiences including that of voices. It should be noted that compliance has been the primary outcome in three trials that explicitly aimed to reduce harmful compliance behaviour through altering beliefs about the power of the voices (Birchwood, Michail, et al., 2014; Trower et al., 2012). However, confidence to resist harmful CHs and confidence in coping with commands were used instead in Trower et al. (2012), since only 64% of the sample had ever complied and about 42% complied to harmful commands in the last six months. Additionally, the subjective negative impact of voices has been the primary outcome in one study of MBCT (Louise, Rossell, & Thomas, 2019). This has been assessed using an adapted version of the Subjective Experiences of Psychosis Scale (SEPS; Haddock et al., 2011), which is one of the few instruments in the area that has been developed in collaboration with individuals with lived experience of psychosis.

3.2.2. Broader Symptomatology

Despite the apparent move away from broad psychotic symptomatology outcomes to voice-related outcomes, psychotic symptoms have been measured relatively often in the identified studies. Specifically, positive and negative symptoms have been assessed in 14 and 15 studies of psychological interventions, respectively. Psychiatric symptoms have also been assessed in 15 studies. These have been predominately measured using the PANSS (Kay, Fiszbein, & Opler, 1987). Variables relating to delusions including the experience, phenomenology, negative impact and severity of delusions, have been examined in 13 studies. Of these, one uncontrolled study of brief Psychoeducation aiming to change the attribution of auditory hallucinations and reduce secondary delusions, has used delusional ideation as a primary outcome (Shiraishi et al., 2014).

In terms of general psychopathology symptoms, depression has been measured in 52 studies of psychological interventions, of which two RCTs have defined this as a primary outcome (Leff et al., 2013; van der Gaag, van Oosterhout, Daalman, Sommer, & Korrelboom, 2012), 21 studies as a secondary outcome (Birchwood, Michail, et al., 2014; Brand et al., 2020; Chadwick et al., 2016; Craig et al., 2018; Dellazizzo et al., 2020; Dodgson et al., 2021; Gottlieb et al., 2013; Gottlieb et al., 2017; Hayward et al., 2017; Hayward et al., 2018; Hayward et al., 2021; Hayward et al., 2022; Hazell et al., 2018; Jones et al., 2021; Louise et al., 2019; Newton et al., 2005; Paulik et al., 2019; Paulik, Jones, & Hayward, 2018; Penn et al., 2009; Trower et al., 2004; Wykes et al., 1999; Yang, Lee, Lo, & Beckstead, 2015), whereas the remaining studies did not pre-define outcomes. Competitive memory training (COMET) was the only intervention that explicitly aimed to reduce depression through strengthening memories linked to positive self-esteem (van der Gaag et al., 2012). Moreover, anxiety has been examined in 24 studies of psychological interventions, the majority of which have used this as a secondary outcome (Brand et al., 2020; Chadwick et al., 2016; Craig et al., 2018; Dodgson et al., 2021; Gottlieb et al., 2013; Gottlieb et al., 2017; Hayward et al., 2017; Hayward et al., 2018; Hayward et al., 2021; Hayward et al., 2022; Hazell et al., 2018; Newton et al., 2005; Paulik et al., 2018; Paulik et al., 2019; Wykes et al., 1999; Yang et al., 2015), while some studies did not pre-specify outcomes. Both depression and anxiety have been assessed using a variety of instruments (e.g., Beck Depression Inventory-II [BDI-II; Beck, Steer, & Brown, 1996], Hospital Anxiety Depression Scale [HADS; Zigmond & Snaith, 1983], Generalised Anxiety Disorder-7 [GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006]).

Table 1

<table>
<thead>
<tr>
<th>Study design</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative randomised</td>
<td>1. Is randomisation appropriately performed?</td>
</tr>
<tr>
<td>controlled trials</td>
<td>2. Are the groups comparable at baseline?</td>
</tr>
<tr>
<td></td>
<td>3. Are there complete outcome data?</td>
</tr>
<tr>
<td></td>
<td>4. Are outcome assessors blinded to the intervention provided?</td>
</tr>
<tr>
<td></td>
<td>5. Did the participants adhere to the assigned intervention?</td>
</tr>
<tr>
<td>Quantitative non-randomised</td>
<td>1. Are the participants representative of the target population?</td>
</tr>
<tr>
<td>studies</td>
<td>2. Are measurements appropriate regarding both the outcome and intervention?</td>
</tr>
<tr>
<td></td>
<td>3. Are there complete outcome data?</td>
</tr>
<tr>
<td></td>
<td>4. Are the confounders accounted for in the design and analysis?</td>
</tr>
<tr>
<td></td>
<td>5. During the study period, is the intervention administered as intended?</td>
</tr>
<tr>
<td>Quantitative descriptive</td>
<td>1. Is the sampling strategy relevant to address the research question?</td>
</tr>
<tr>
<td>studies</td>
<td>2. Is the sample representative of the target population?</td>
</tr>
<tr>
<td></td>
<td>3. Are the measurements appropriate?</td>
</tr>
<tr>
<td></td>
<td>4. Is the risk of nonresponse bias low?</td>
</tr>
<tr>
<td></td>
<td>5. Is the statistical analysis appropriate to answer the research question?</td>
</tr>
</tbody>
</table>

5
All other general symptoms have been measured less consistently, for example, dissociation, post-traumatic stress symptoms and trauma-related outcomes have been primarily assessed in studies that have targeted these variables (Brand et al., 2020; Paulik et al., 2019; Paulik et al., 2020; Steel et al., 2019; Varese et al., 2020). Three of these studies have addressed trauma memories associated with their voices (Brand et al., 2020; Paulik et al., 2019; Paulik et al., 2020) and one study targeted dissociative experiences linked to trauma in voice hearers (Varese et al., 2020).

3.2.3. Broader concepts of outcome

Less attention has been placed on broader concepts of outcome compared to other variables (e.g., depression, anxiety). Specifically, functioning has been measured in 12 studies of psychological interventions, with one trial of group CBT using social functioning/disability as a main outcome (Wykes et al., 2005). Similarly, subjective recovery has been examined in 11 studies of psychological interventions and has been mostly assessed using the Choice of outcome in CBT for psychoses scale (CHOICE; Greenwood et al., 2010; Webb et al., 2021), which has been developed in consultation with service users.

Recovery has been primarily used as a secondary outcome (Chadwick et al., 2016; Clarke et al., 2021; Hayward et al., 2017; Hayward et al., 2018; Hayward et al., 2021; Hayward et al., 2022; Hazell et al., 2018; Jones et al., 2021; Varese et al., 2020), although one trial of EFC has used this as a primary outcome (Schnackenberg et al., 2017) and another trial did not pre-specify outcomes (Knott et al., 2020). Moreover, a variety of health-related outcomes (e.g., well-being, psychological distress, quality of life, satisfaction with life) have been measured in five or fewer studies. Psychological distress was the only variable relating to health-related outcomes that was used as a primary outcome in two studies of PBCT (Chadwick et al., 2016; Dannahy et al., 2011).

3.2.4. Voice-related appraisals

Voice-related appraisals have been widely measured and targeted in studies of psychological interventions; for example, omnipotence has been measured in 21 studies and perceived controllability in 18 studies. Furthermore, malevolence has been measured in 15 studies and benevolence in 13 studies. These studies mainly involved interventions that aimed to identify and challenge beliefs associated with the perceived power (omnipotence) and intention of the voices (malevolence, benevolence). For instance, Avatar Therapy and VRT focused on changing perceived voice power and control (Craig et al., 2018; du Sert et al., 2018) and CTCH explicitly targeted voice power appraisals (Birchwood, Michail, et al., 2014; Trower et al., 2004). Perceived controllability has been mostly assessed using the ‘controllability’ item of the PSYRATS-AH (Haddock et al., 1999), whilst appraisals of malevolence, benevolence and omnipotence have generally been assessed using the Beliefs About Voices Questionnaire (BAVQ; Chadwick & Birchwood, 1995, BAVQ-Revised; Chadwick, Lees, & Birchwood, 2000). The Voice Power Differential scale (VPD; Birchwood et al., 2000) has been used less consistently to measure omnipotence/perceived power (Birchwood, Michail, et al., 2014; Craig et al., 2018; Stefanjak, Sorokosz, Janicki, & Wciórka, 2019; Trower et al., 2004).

Beliefs regarding origin of voices have been assessed in six studies of psychological interventions, of which one study used cognitive restructuring to challenge distressing thoughts involving beliefs about the content or origin of the voices (Gottlieb et al., 2013). Finally, omnipotence has been measured in only two trials of CTCH (Birchwood, Michail, et al., 2014; Trower et al., 2004) using the Omnipotence Scale (Birchwood et al., 2004).

3.2.5. Coping

Several variables relating to coping with voices have been assessed in a small number of studies of psychological interventions, all of which have dealt with teaching effective coping techniques. In particular, the number of coping strategies has been measured in five studies of psychological interventions (Bell et al., 2020; Jenner et al., 2004; Newton et al., 2005; Wykes et al., 1999; Wykes et al., 2005). The types and effectiveness of coping strategies have been assessed in one study of group CBT (Wykes et al., 1999) and the frequency of coping strategies, confidence in coping with voices and understanding of voices have been measured in one study of CSE (Bell et al., 2020).

3.2.6. Cognitions

Self-esteem appeared to be the most frequently used variable relating to cognitions, as it has been measured in 11 studies of psychological interventions, several of which have addressed self-esteem (Craig et al., 2018; Hazell et al., 2018; Newron et al., 2005; Paulik et al., 2018; Penn et al., 2009; Wykes et al., 1999; Wykes et al., 2005). Additionally, beliefs about self have been measured in only two studies of CBT that have targeted both negative and positive self-schemata (Hayward et al., 2021; Hazell et al., 2018), despite being targeted in several studies of psychological interventions (e.g., Chadwick et al., 2016; Dannahy et al., 2011; Jones et al., 2021). Insight has been measured in five studies of psychological interventions, whilst cognitions associated with trauma have been measured in one study of Imaginal Exposure (Brand et al., 2020) and were used as mechanisms of change.

3.2.7. Attitudes and skills

Variables relating to attitudes and skills have also been measured in a limited number of studies. Mindful awareness has been measured in two studies of mindfulness-based therapies (Louise et al., 2019; Lüdtke et al., 2020) and acceptance and action-based beliefs have been primarily measured in studies that have focussed on increasing acceptance in relation to voices or CHs (El Ashry, Abd El Dayem, & Ramadan, 2021; Knott et al., 2020; Langlois et al., 2026; Shawyer et al., 2012), with the exception of one study (Craig et al., 2018). Voice relating has been measured in five studies (Dannahy et al., 2011; Hayward et al., 2017; Hayward et al., 2021; Hazell et al., 2018; Steel et al., 2019) and social relating in three studies of psychological interventions (Hayward et al., 2017; Hayward et al., 2021; Hazell et al., 2018), all of which involved reconstructing the way in which individuals relate to their voices and/or to other people.

3.3. Change over time

The majority of variables demonstrated improvements across most RCTs. However, effect sizes varied considerably from minimal to large. Due to the range of variables used in these RCTs, only the most frequently measured variables are discussed in detail. These include the voice hearing experience, voice-related distress, depression and appraisals of voice omnipotence (see Tables S5–8 in Supplementary Material).

3.3.1. Voice hearing experience

The voice hearing experience has been measured in trials of psychological interventions using the total score of the PSYRATS-AH. Improvements were seen in most trials, with effects in the medium-to-large range. Compared to TAU, Wykes et al. (2005) found that group CBT had no effect on the voice hearing experience at post-treatment, whilst Knott et al. (2020) found a large effect. When group CBT was compared to Supportive Therapy, a minimal improvement was observed at post-treatment, but this was not sustained at 3-month and 1-year follow-ups (Penn et al., 2009). Similarly, PBCT had a small effect on the voice hearing experience at post-treatment compared to TAU, but this was not maintained at 10-month follow-up (Chadwick et al., 2016).

Online CBT had no effect on the voice hearing experience both at post-treatment and at 3-month follow-up in comparison to Usual Care (Gottlieb et al., 2017). On the other hand, a pilot RCT found a large effect size in favour of guided self-help CBT at post-treatment compared to TAU (Hazell et al., 2018). The feasibility RCT of guided self-help CBT demonstrated a medium effect size in favour of the intervention at post-
treatment and a small effect size at 28-week follow-up, compared to TAU (Hayward et al., 2021). When the therapy was compared to Supportive Counselling, a large improvement was seen in favour of guided self-help CBT at post-treatment, but this decreased substantially to a minimal effect at follow-up (Hayward et al., 2021). Both COMET and EFC exhibited small effects on the voice hearing experience following treatment, whilst CSE demonstrated a medium effect and RT, ACT and computer/VR-assisted therapies found large effects, all compared to inactive control groups (Bell et al., 2020; du Sert et al., 2018; El Ashry et al., 2021; Hayward et al., 2017; Leff et al., 2013; Schnackenberg et al., 2017; Stefanik et al., 2019; van der Gaag et al., 2012). The effects of Avatar Therapy were medium at post-treatment and small at 3-month follow-up, when compared to Supportive Counselling (Craig et al., 2018). CTCH had no effect both at post-treatment and at 18-month follow-up compared to TAU (Birchwood, Michail, et al., 2014). Finally, the voice hearing experience demonstrated medium improvements with HIT at post-treatment compared to TAU (Jenner et al., 2004), while small and medium effects were seen when complete cases were used at post-treatment and at 18-month follow-up, respectively (Jenner et al., 2006).

3.3.2. Voice-related distress

Voice-related distress also demonstrated medium-to-large improvements in the majority of trials. Group CBT had a small effect on the amount of voice-related distress item of the PSYRATS-AH at post-treatment compared to TAU (McLeod, Morris, Birchwood and Dovey, 2007b). Additionally, PBCT had a medium effect on the intensity of voice-related distress item of the PSYRATS-AH at post-treatment in comparison to TAU, but this diminished to a minimal effect at 10-month follow-up (Chadwick et al., 2016). Individual CBT had large effects on the amount and intensity of voice-related distress items of the PSYRATS-AH at post-treatment relatively to TAU, and small effects at 12-week follow-up (Shukla, Padhi, Sengar, Singh, & Chaudhury, 2021). Compared to TAU, Hazell et al. (2018) found that guided self-help CBT had a large effect on the voice impact subscale of the HPSVQ at post-treatment, whilst Hayward et al. (2021) found that guided self-help CBT had a medium effect on the voice impact subscale of the HPSVQ and a small-to-medium effect on the voice-related distress subscale of the PSYRATS-AH at post-treatment. At 28-week follow-up, the effect of the intervention on the voice impact subscale of the HPSVQ was only minimal, but the voice-related distress subscale of the PSYRATS-AH showed a marginal improvement (medium effect) (Hayward et al., 2021). When the therapy was compared to Supportive Counselling, medium effect sizes were observed in favour of guided self-help CBT for both the voice impact subscale of the HPSVQ and the voice-related subscale of the PSYRATS-AH at post-treatment. However, the effects of the therapy on both scales were not sustained at follow-up (Hayward et al., 2021).

An online self-guided mindfulness module only had a minimal effect on the DV-SA at post-treatment compared to Waiting list (Lüdtke et al., 2020), whereas RT produced large gains on the voice-related distress subscale of the PSYRATS-AH at post-treatment and at 36-week follow-up compared to TAU (Hayward et al., 2017). Avatar (Stefanik et al., 2019) and VRT (du Sert et al., 2018) also had large effects on the intensity and amount of voice-related distress items and the voice-related distress subscale of the PSYRATS-AH at post-treatment, respectively, compared to inactive control groups. A medium effect on the voice-related distress subscale of the PSYRATS-AH was observed in favour of Avatar Therapy at post-treatment and a small effect at 3-month follow-up, when compared to Supportive Counselling (Craig et al., 2018).

Trials comparing CTCH to TAU showed different findings. In particular, a small RCT (Trower et al., 2004) found that the therapy demonstrated medium improvements on the intensity of voice-related distress item of the PSYRATS-AH at post-treatment and small improvements at 12-month follow-up, whilst the larger scale RCT (Birchwood, Michail, et al., 2014) found no effect on the amount and intensity of voice-related distress items of the PSYRATS-AH at post-treatment and at 18-month follow-up. Similarly, the voice-related distress subscale of the PSYRATS-AH showed no improvements with TORCH at post-treatment and at 6-month follow-up compared to both Befriending and Waiting list (Shawyer et al., 2012). In contrast, large improvements in favour of ACT were seen for the amount and intensity of voice-related distress items and the negative emotional content subscale of the PSYRATS-AH at post-treatment and at 3-month follow-up compared to TAU (El Ashry et al., 2021). With respect to HIT, a medium effect size for the distress index of the PSYRATS-AH was found at post-treatment, compared to TAU (Jenner et al., 2004), but after using complete data, a small effect size was observed at post-treatment and a medium effect size at 18-month follow-up (Jenner et al., 2006).

3.3.3. Depression

Most trials reported small-to-medium effects against active and inactive control groups on depression. Group CBT had a minimal effect on the BDI-II at post-treatment compared to Supportive Therapy, but this increased to a small effect at 3-month follow-up and to a medium effect at 1-year follow-up (Penn et al., 2009). PBCT and online CBT also demonstrated small-to-medium improvements compared to TAU. More specifically, PBCT had a small effect on the HADS at both post-treatment and at 10-month follow-up (Chadwick et al., 2016) and online CBT had a small effect on the Brief Psychiatric Rating Scale (BPRS) and a medium effect on the BDI at post-treatment (Gottlieb et al., 2017). Improvements on the BPRS were small at 3-month follow-up, whilst improvements on the BDI remained stable (Gottlieb et al., 2017).

In terms of guided self-help CBT, small improvements were seen on the HADS at post-treatment compared to TAU in the pilot RCT (Hazell et al., 2018), whereas the feasibility trial demonstrated relatively greater improvements (Hayward et al., 2021); medium effects were seen in favour of the intervention at post-treatment and at 28-week follow-up when compared to TAU. A medium effect size in favour of the intervention was also seen at post-treatment when compared to Supportive Counselling, which diminished to a small effect at follow-up (Hayward et al., 2021). RT also demonstrated a medium effect on the HADS at post-treatment compared to TAU, which increased to a large effect at 36-week follow-up (Hayward et al., 2017). Computer/VR-assisted therapies showed contradicting findings. Compared to delayed therapy, Leff et al. (2013) found that Avatar Therapy had no effect on the Calgary Depression Scale (CDS), whereas du Sert et al. (2018), demonstrated a large effect size on the BDI, favouring VRT. When compared to Supportive Counselling, Avatar Therapy had a small effect on both the CDS and the DASS-21 at post-treatment (Craig et al., 2018). At 3-month follow-up, the effect on the CDS was minimal, whilst the effect on the DASS-21 disappeared. A comparable pattern was seen in trials comparing CTCH to TAU; Trower et al. (2004) found that CTCH had a small effect on the CDS at post-treatment and a medium effect at 12-month follow-up, whilst Birchwood, Michail, et al., 2014 found no effect at both post-treatment and at 18-month follow-up. A relatively small trial that specifically aimed to target depression by comparing COMET to TAU, showed that the therapy had a medium effect on the BDI-II at post-treatment (van der Gaag et al., 2012).

3.3.4. Omnipotence

Several trials of psychological interventions demonstrated medium-to-large effects on omnipotence. Group CBT demonstrated large improvements on the omnipotence subscale of the BAVQ-R at post-treatment compared to TAU (McLeod, Morris, Birchwood and Dovey, 2007b), but no effects were seen when compared to Supportive Counselling (Penn et al., 2009). However, minimal effects were observed at 3-month and 1-year follow-ups (Penn et al., 2009).

Compared to TAU, online CBT had a medium effect on the omnipotence subscale of the BAVQ-R at post-treatment and a minimal effect at 3-month follow-up (Gottlieb et al., 2017), whilst guided self-help CBT demonstrated a large effect size at post-treatment (Hazell et al., 2018).
Two trials of computer/VR-assisted therapies also exhibited large improvements on the omnipotence subscale of the BAVQ-R (du Sert et al., 2018) and the perceived power item of the VPD (Stefaniak et al., 2019), in contrast to delayed therapy. Nonetheless, Craig et al. (2018) found that Avatar Therapy had a medium effect on the omnipotence subscale of the BAVQ-R and a small effect on the perceived power item of the VPD at post-treatment compared to Supportive Counselling. At 3-month follow-up, the effect on the BAVQ-R was small, whilst the effect on the VPD had diminished. In a similar manner, Birchwood, Michail, et al., 2014 showed that CTCH had small effects on the omnipotence subscale of the BAVQ-R and the perceived power item of the VPD at post-treatment compared to TAU. The effect on the BAVQ-R was sustained at 18-month follow-up, whereas the effect on the VPD somewhat decreased but was still within the small range. Large effect sizes were observed for the perceived power item of the VPD, favouring CTCH at post-treatment and at 12-month follow-up compared to TAU (Trower et al., 2004). TORCH had no effect on the omnipotence subscale of the BAVQ-R both at post-treatment and at 6-month follow-up compared to Befriending (Shawyer et al., 2012).

3.3.5. Additional variables

Trials of psychological interventions either had no effect on benevolence or demonstrated minimal-to-small effects when compared to both active and inactive control groups (Birchwood, Michail, et al., 2014; Craig et al., 2018; Hayward et al., 2021; Hazell et al., 2018; Penn et al., 2009). Service-user led recovery as measured by the CHOICE appeared to have medium-to-large improvements both at post-treatment and at follow-up in three trials of cognitive-behavioural approaches compared to TAU (Hayward et al., 2017; Hayward et al., 2021; Hazell et al., 2018). However, a small effect was observed in favour of guided self-help CBT at post-treatment, when compared to Supportive Counselling, which was not sustained at 28-week follow-up (Hayward et al., 2021). Minimal-to-small improvements were also observed in one trial of PBCT at post-treatment (Chadwick et al., 2016). Reductions in compliance were seen in one trials of CTCH that aimed to decrease compliance through reducing the power imbalance between the voices and the voice hearer. The pilot trial demonstrated large improvements (Trower et al., 2004), whereas the full-scale trial showed minimal-to-small improvements (Birchwood, Michail, et al., 2014), at post-treatment and at follow-ups. Controllability demonstrated small-to-large improvements across trials of psychological interventions (Chadwick et al., 2016; El Ashry et al., 2021; Jenner et al., 2004; Jenner et al., 2006; Shukla et al., 2021; Stefaniak et al., 2019; Trower et al., 2004). Mixed findings were seen for other variables, as effect sizes noticeably ranged across RCTs for malevolence, hallucination severity, positive and negative symptoms, self-esteem, anxiety, functioning, delusion experience, voice frequency and many more. All effect sizes can be seen in Table S1 (see Supplementary Material).

4. Discussion

This review aimed to identify the variables that have been used to measure the longitudinal course and impact of voice hearing under psychological interventions, and to examine how these variables changed over time.

4.1. Main findings

We found that a range of variables relating to the characteristics of both the voices and the voice hearer have been used in studies of psychological interventions for distressing voices. Surprisingly, depression appeared to be one of the most prominent variables. However, only two trials that pre-defined outcomes have used this as a primary outcome (Leff et al., 2013; van der Gaag et al., 2012), of which one intervention explicitly aimed to reduce levels of depression (COMET; van der Gaag et al., 2012), indicating a discrepancy between outcome measurement and targeting of specific processes.

It is also important to highlight that the majority of studies have defined the voice hearing experience as a primary outcome, but this has been largely measured using the total score of the PSYRATS-AH, which contains items that are not directly impacted by psychological interventions and subsequently less likely to change. Similarly, a range of variables that are not directly impacted by psychological interventions have been examined in the identified studies including the frequency of voices and the severity of positive and negative symptoms. Moreover, voice-related distress has been consistently measured in studies of psychological interventions, but only three trials have used this as a primary outcome, despite being more directly impacted by psychological interventions as opposed to other outcomes (e.g., voice hearing experience).

It has been increasingly recognised that psychological trials should also consider broader outcomes (Cuijpers, 2019; Thomas, 2015a). Indeed, psychological interventions for schizophrenia and psychotic experiences have several aims including the reduction of associated distress, the improvement of functioning and the promotion of recovery (National Institute for Health and Care Excellence, 2014; Scottish Intercollegiate Guidelines Network, 2013). A variety of broader concepts of outcome such as functioning, satisfaction with life, well-being and subjective recovery have been measured in recent studies, suggesting that there is a growing acceptance of these outcomes. However, the heterogeneity of these outcomes also indicates that there is a lack of consensus among researchers about what should be measured.

Studies of psychological interventions have used a range of variables relating to proposed processes involved in the maintenance of voice hearing experiences, with some of these processes having been measured and targeted more frequently than others. Voice-related appraisals (e.g., malevolence, benevolence) have been measured in a large number of studies, the majority of which have targeted these variables. On the other hand, many studies of cognitive-behavioural approaches have not measured self-schema or self-related appraisals, despite cognitive models of voices highlighting their role in voice hearing (Birchwood et al., 2000; Birchwood et al., 2004) and some interventions targeting these appraisals (e.g., Chadwick et al., 2016; Dannahy et al., 2011). Other processes (e.g., dissociation, trauma memory, voice relating, social relating) have been measured in a small number of studies of psychological interventions that explicitly targeted these variables. Furthermore, while the importance of social mechanisms (e.g., stigma, isolation) has been consistently emphasised (Ruddle, Mason, & Wykes, 2011; Vilhauer, 2017), these were not measured by any of the identified studies.

Our findings also indicate greater improvements (i.e., medium-to-large effect sizes) across several variables for recent trials of symptom-specific interventions compared to previous trials of CBtp (i.e., small-to-medium, effect sizes - see meta-analyses; Hazell et al., 2016; Jauhar et al., 2014; Turner et al., 2020; van der Gaag et al., 2014; Wykes et al., 2008; Zimmermann et al., 2005), thereby offering some promising evidence about the effectiveness of symptom-specific psychological approaches. Specifically, several trials established medium-to-large effects on the voice hearing experience (e.g., Bell et al., 2020; El Ashry et al., 2021; Hazell et al., 2018; Knott et al., 2020), voice-related distress (e.g., du Sert et al., 2018; Hayward et al., 2017; Hazell et al., 2018; Stefaniak et al., 2019), omnipotence (e.g., du Sert et al., 2018; Gottlieb et al., 2017; Hazell et al., 2018; Trower et al., 2004) and subjective recovery (e.g., Hayward et al., 2017; Hayward et al., 2021; Hazell et al., 2018). Levels of depression also tended to decrease under psychological interventions, with small-to-medium effect sizes seen across most trials (e.g., Birchwood, Michail, et al., 2014; Chadwick et al., 2016; Gottlieb et al., 2017; van der Gaag et al., 2013), in line with cognitive models of psychosis suggesting that depression can play a key role in the maintenance of voice hearing (Garety et al., 2001; Morrison, 2001). Previous research has demonstrated that negative schematic beliefs about self may lead to appraisals of voice power, depression and voice-related
distress (Birchwood et al., 2004). The improvements in depression may therefore be attributed to changes in cognitive processes (e.g., beliefs about self, voice-related appraisals) rather than changes in voice hearing having a direct effect on depression. However, more research is needed using longitudinal designs to understand the role of depression in voice hearing, as psychological interventions may benefit from targeting depression.

Despite trials of psychological interventions for distressing voices having an effect on several variables, the magnitude of these tended to fluctuate across RCTs. Further to that, effect sizes did not only range across trials of different psychological interventions, but this variability was also seen across trials of the same psychological interventions. For instance, trials of group CBT, computer/VR-assisted therapies and CTCH found differing results. This variability might be due to several methodological reasons. First, some studies incorporated a single-blind design (e.g., Craig et al., 2018) whilst others did not (e.g., du Sert et al., 2018). Second, the majority of trials have compared psychological interventions to inactive control groups (e.g., TAU), whilst some others have used active controls (e.g., Supportive Counselling). Thus, the effectiveness of these psychological interventions could have possibly been amplified, as effect sizes tended to be larger when psychological interventions were compared to inactive controls. For example, Shawyer et al. (2012) found smaller effect sizes when TORCH was compared to an active control (Befriending), relative to a comparison between TORCH and a Waiting list control.

Third, there were several underpowered studies including pilot RCTs (e.g., Hayward et al., 2017; Hazell et al., 2018; Shawyer et al., 2012; Trower et al., 2004; van der Gaag et al., 2012) and therefore effect size estimates could have potentially been either inflated or deflated in these trials. For instance, differences in effect sizes of trials of CTCH may be attributed to differences in sample sizes (Trower et al., 2004 n = 38; Birchwood, Michail, et al., 2014 n = 197). A further explanation refers to the fact that several variables were set as primary outcomes in some trials and as secondary in others, leading to differences in the magnitude of effects. That is, trials often implement a range of secondary outcomes, which may result in a false-positive error rate, or they are inadequately powered for the primary outcome but not for secondary outcomes, which may yield false-negative results. Finally, outcomes of these trials have been assessed using a variety of measures, therefore some are likely to be more sensitive than others and thus changes are more readily detected. For example, Kim et al. (2019) found a greater decrease in PSYRATS-AH scores than HPSVQ scores over a period of a year, indicating that the former measure might be more sensitive for detecting changes in voices. Different methods of scoring have also been introduced (item level, subscale scores, total scores), meaning that some of these (e.g., individual items such as amount and intensity of voice-related distress) may lack sensitivity to changes and ultimately complicate attempts to compare findings across trials.

4.2. Strengths and limitations of review

As far as we are aware, this is the first systematic review that specifically examined the course of and heterogeneity in variables of psychological interventions for distressing voices. We followed a rigorous process, in accordance with PRISMA guidelines in order to reduce bias. We have also included a range of study designs (e.g., uncontrolled) rather than only RCTs to form a better understanding of the variables that have been used to assess the longitudinal course and impact in studies of psychological interventions of distressing voices. One limitation is the fact that we have included a range of psychological interventions for distressing voices, and this may have caused heterogeneity, particularly in terms of effect size estimates. A further limitation of this review concerns the trial participants who may not be reflective of the population and therefore results from trials of psychological interventions may not be applicable to routine practice. Despite voice hearing being increasingly considered to be a transdiagnostic phenomenon, the majority of participants met diagnostic criteria for schizophrenia-spectrum or psychotic-spectrum disorders, consequently neglecting different populations that experience distressing voices such as personality disorders and mood disorders. In addition, many of the identified studies have been conducted in Western countries and a high proportion of participants were White. Hence, we cannot be certain of the generalisability of findings and the applicability of these interventions in other ethnicities and cultures.

4.3. Implications and future directions

The findings of this review suggest that a range of broad (e.g., positive symptoms) and indirect variables (e.g., frequency) have been used to measure outcomes within trials, whereas broader concepts of outcome (e.g., functioning) and several associated processes (e.g., self-schemata) have not been consistently examined. Findings also suggest that most studies have focus upon primary outcomes that are not directly impacted by psychological interventions (e.g., voice hearing experience) instead of more relevant outcomes (e.g., voice-related distress). Our review confirms that there are differing views as to what we should be measuring in these psychological interventions and stresses the importance of clarifying outcomes and their measurement (Thomas et al., 2014). Therefore, clarity about the extent to which these psychological interventions are effective could be generated by the consistent use of the most relevant outcomes. Furthermore, psychological interventions do not only need to demonstrate that they work but are also required to show how they work. We therefore suggest that future research trials define primary outcomes that are most directly impacted by psychological interventions such as voice-related distress and to use a balanced number of measures of targeted processes and broader concepts as secondary outcomes. These broader concepts of outcome should also be prioritised by voice hearers, as most of the reviewed variables have been determined by professionals in the field. A qualitative study is currently being conducted that will explore the views and perspectives of voice hearers and practitioners about the outcomes that are prioritised by them (MOTIVE, ref. 21/L0/0257). Reaching a consensus and determining outcomes of psychological interventions for distressing voices will provide clarity to both research and routine clinical practice and subsequently facilitate efforts to compare outcomes across studies and to evaluate the effectiveness of these psychological interventions.

5. Conclusions

The findings of this systematic review show a complex picture that has been created by differing views in the field and methodological differences in trials of psychological interventions. It is evident that symptom-specific psychological interventions have moved away from broad symptomatology measures towards voice-specific measures, which has created the opportunity for a greater focus. Yet, the use of primary outcomes that are not directly impacted by psychological interventions has impeded the comparability of findings. Consequently, refining these outcomes and their measurement is of high importance.

Role of funding sources

This work is supported by the South east Network for Social Sciences (SeNSS) Supervisor-led Collaborative Studentship sustained by the Economic and Social Research Council (ESRC) and Sussex Partnership NHS Foundation Trust. SeNSS has no role in in the design, analysis, interpretation of the data, writing the manuscript or the decision to submit the paper for publication.

Contributors

SL conducted the literature search and the analysis and drafted the
Clinical Psychology Review 97 (2022) 102191
S. Loizou et al.


