Validation of the Hamilton Program for Schizophrenia Voices Questionnaire: associations with emotional distress and wellbeing and invariance across diagnosis and sex

Article (Accepted Version)


This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/96533/

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
Title: Validation of the Hamilton Program for Schizophrenia Voices Questionnaire: Associations with emotional distress and wellbeing and invariance across diagnosis and sex

Authors:
Berry, C. (corresponding)a,b,c, Newcombe, Hb, Strauss, C.b,c, Rammou, A.b, Schlier, B.d,
Lincoln, T.d,& Hayward, M.b,c

aPrimary Care and Public Health, Brighton and Sussex Medical School, University of Brighton, Falmer, Brighton, BN1 9PH.
bSchool of Psychology, University of Sussex, Falmer, Brighton, BN1 9QH.
cResearch & Development, Sussex Partnership NHS Foundation Trust, Sussex Education Centre, Millview Hospital, Nevill Avenue, Hove, BN3 7HZ.
dKlinische Psychologie und Psychotherapie, Institut für Psychologie, Fakultät für Psychologie und Bewegungswissenschaft, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg.

Correspondence to:
Dr Clio Berry,
Primary Care and Public Health,
Watson Building,
Brighton and Sussex Medical School,
University of Brighton,
Falmer,
Brighton,
BN1 9PH.
Email: c.berry@sussex.ac.uk / c.berry@bsms.ac.uk

Word count: 3932
Abstract: 249
Tables: 4
Figures: 1
Abstract

**Background:** Voice-hearing is a transdiagnostic experience with an evident negative impact on patients. Good quality measurement is needed to further elucidate the nature, impact and treatment of voice-hearing experiences across patient groups. The Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) is a brief self-report measure which requires further psychometric evaluation.

**Methods:** Using data from a transdiagnostic sample of 401 adult UK patients, the fit of a conceptual HPSVQ measurement model, proposing a separation between physical and emotional voice-hearing characteristics, was tested. A structural model was examined to test associations between voice-hearing, general emotional distress (depression, anxiety, stress) and wellbeing. The invariance of model parameters was examined across diagnosis and sex.

**Results:** The final measurement model comprised two factors named ‘voice severity’ and ‘voice-related distress’. The former comprised mainly physical voice characteristics and the latter mainly distress and other negative impacts. Structural model results supported voice-related distress as mediating the associations between voice severity and emotional distress and wellbeing. Model parameters were invariant across psychosis versus non-psychosis diagnosis and partially invariant across sex; with females experiencing more severe and distressing voices and a more direct association between voice severity and general anxiety.

**Conclusions:** The HPSVQ is a useful self-report measure of voice-hearing with some scope for further exploration and refinement. Voice-related distress appears a key mechanism by which voice severity predicts general distress and wellbeing. Whilst our data broadly support interventions targeting voice-related distress for all patients, females may benefit especially from interventions targeting voice severity and strategies for responding.
1. Background

Voice-hearing is prevalent amongst people with psychosis, it is reported by 70% of people with a diagnosis of schizophrenia (Larøi et al., 2012). Voice-hearing is also common across other diagnoses, such as Emotionally Unstable Personality Disorder and Bipolar Disorder (Baumeister et al., 2017; Upthegrove et al., 2015). Voice-hearing can be a benign or positive experience (Lorente-Rovira et al., 2020). However, voice-hearing is commonly distressing and is associated with depression (Birchwood et al., 2004), poor functioning (Favrod et al., 2004), and increased suicidality (Kjelby et al., 2015). Even with antipsychotic medication, about 50% of patients report ongoing voice-hearing (González et al., 2006). In recent years, psychological interventions for voice-hearing have been rapidly evolving to meet this treatment demand. New innovations include symptom-specific Cognitive Behaviour Therapies for voices (CBTv) such as: brief therapist-guided self-help CBTv (Hazell et al., 2018b); group mindfulness-based cognitive therapy (Chadwick et al., 2016); relationally-informed CBTv (e.g. Relating Therapy (Hayward, Jones, Bogen-Johnston, Thomas, & Strauss, 2017), AVATAR therapy (Craig et al., 2018); and cognitive therapy for Command Hallucinations (Birchwood et al., 2014).

Psychometrically sound voice-hearing measures are necessary to capture voice-hearing impacts and therapeutic change (Thomas et al., 2014). The most widely used measure is the Psychotic Symptom Rating Scales – Auditory Hallucinations subscale (PSYRATS-AH (Haddock et al., 1999)), a semi-structured observer-rated interview capturing multiple characteristics of voice-hearing, related beliefs and impact. The Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ (Van Lieshout and Goldberg, 2007)) is a promising self-report alternative. The HPSVQ has convergent validity with PSYRATS-AH and covers the same key domains (frequency, loudness, duration, interference with life, distress, negative content), plus it includes three additional items (voice clarity, obey voice commands, and impact on self-appraisal (Ratcliff et al., 2011; Van Lieshout and Goldberg, 2007)). As a self-report measure, the HPSVQ is methodologically flexible and cost-effective for it requires no skilled interviewer (Ratcliff et al., 2011). Nevertheless, the HPSVQ’s psychometric properties are predicated on only two known studies; one involving only 20 patients (Van Lieshout and Goldberg, 2007) and the second using a translated version of the HPSVQ with South-Korean patients (Kim et al., 2010).

Typically, measures such as HPSVQ are used in the form of total scores. Whilst the total HPSVQ score has acceptable internal consistency and test-retest reliability (Ratcliff et al., 2011; Van Lieshout and Goldberg, 2007), it unhelpfully collates indices of voice form and
Voice impact; items which do not necessarily correlate highly (Thomas et al., 2014) and may not reflect a unitary construct. Kim and colleagues (2010) identified two factors with South-Korean patients. The first was termed physical characteristics and comprised frequency, clarity, duration, loudness, and obey commands. The second was labelled emotional characteristics, and comprised distress, negative self-appraisals, negative content and interference with life. Factor explorations of PSYRATS-AH have similarly indicated separation between physical versus emotional indicators (Chang et al., 2009; Haddock et al., 1999; Woodward et al., 2014). However, there are some important limitations of Kim and colleagues’ study (2010). These include the use of a small sample only of people with a schizophrenia diagnosis and limiting factorial validation to principal components analysis which does not allow for significance testing between model iterations. Auditory hallucination-like experiences appear broadly cross-culturally invariant (Siddi et al., 2019), however cultural and language differences may influence both voice phenomenology and responses to self-report measures (Kalra et al., 2012; Kim et al., 2010; Larøi et al., 2014). Thus, the fit of the two-factor South-Korean model must be confirmed with UK transdiagnostic data and compared to the simpler one-factor model analogous to a total measure score.

To measure interventional effectiveness, voice-hearing scales must show criterion validity through evident pre-established associations to change in broader markers of health and wellbeing (Thomas et al., 2014; Van Lieshout and Goldberg, 2007). Presently, we were interested in associations between HPSVQ dimensions and broader emotional distress and wellbeing. Previous research indicates that voice-related distress is centrally important to the impact of voices with respect to influencing key outcomes including mood, functioning and wellbeing (Birchwood and Trower, 2006; McCarthy-Jones et al., 2013; Pappa et al., 2019; Thomas, 2014). Physical voice characteristics apparently impact on patients’ lives less directly but do contribute to voice-related distress (Soppitt and Birchwood, 1997; Thomas, 2014). We therefore expected physical voice characteristics to be associated with broader emotional distress and wellbeing indirectly through the mechanism of voice-related distress.

Within a traditional psychiatric perspective, voice-hearing in the context of non-psychotic diagnoses has typically been considered a “pseudohallucination” (Thomas et al., 2014; Upthegrove et al., 2016). However, theoretical and phenomenological accounts of voice-hearing appear broadly tranidiagnostically applicable (Hazell et al., 2018a; Thomas et al., 2014; Waters and Fernyhough, 2017). Nonetheless, empirical testing of the invariance of voice-hearing and its impacts across psychotic and non-psychotic diagnoses is required (Upthegrove et al., 2016). Moreover, notably little attention has been paid to sex differences
in voice-hearing (Hayward, Slater, Berry, & Perona-Garcélán, 2016). However, women report more common, frequent, and severe auditory hallucinations and appear to have a greater propensity for persecutory or ‘delusional’ interpretation of voices (González et al., 2008; Leung & Chue, 2000; Thorup et al., 2007). More generally, women have greater emotional reactivity, intensity, and expressivity than men, especially in relation to negative emotions (Deng et al., 2016; Gard and Kring, 2007; Skowron and Dendy, 2004). Therefore, it seems likely that women may report greater voice-related distress. Testing the invariance of voice-hearing experiences is needed to ensure any observed diagnostic or sex differences reflect true effects as opposed to artefacts of measurement bias (Meredith, 1993).

In summary, we hypothesised that a two-factor structure for the HPSVQ separating physical and emotional characteristics of voices would adequately fit UK data, with both factors predicting emotional distress and wellbeing. Additionally, we hypothesised that emotional voice characteristics would mediate the association between physical voice characteristics and dependent variables. We predicted measurement and structural invariance across diagnoses but hypothesised that women would report greater severity in physical and emotional voice characteristics and would have greater indirect effects through emotional voice characteristics.

2. Methods

2.1. Sample and participants

This study uses the existing Approve study dataset (Hayward et al., 2020). We established before analysis that the present sample met sample size guidance (Muthén and Muthén, 2002) for a simple two-factor model with nine indicators and ordinal data. We entered dependent variables as observed rather than latent variables to maintain an appropriate case-to-parameter ratio (Bentler and Chou, 1987).

Participants met inclusion criteria for the Approve study; aged 16 years or over; able to read in English; heard voices for at least six months. No diagnostic exclusions were applied. Data from 401 participants were used in the present study (Table 1), with 285 participants coded as having a ‘psychosis’ diagnosis versus 116 ‘non-psychosis’, and 244 participants male versus 151 female.

INSERT TABLE ONE HERE
2.2. Procedure

The Approve study was sponsored by the University of Sussex (024HAY) and approved by the London-South East Research Ethics Committee and Health Research Authority (18/LO/0046). Patients in 14 NHS Trusts across England were referred to Approve by their community adult mental health or Early Intervention in Psychosis service. Patients completed a questionnaire battery including current measures in a clinical setting or in their own home assisted by a researcher.

2.3. Measures

The Hamilton Program for Schizophrenia Voice Questionnaire (HPSVQ (Van Lieshout and Goldberg, 2007)) is a 9-item self-report questionnaire(Van Lieshout and Goldberg, 2007). HPSVQ items are rated on an ordinal 0-4 Likert scale, from least to most severe, with all response points containing individual verbal anchors. HPSVQ items are presented in Table 2.

Emotional distress was measured using the 21-item Depression, Anxiety and Stress Scales (DASS-21 (Lovibond and Lovibond, 1995)). Subtotals for depression, anxiety and stress were used as observed variables, with higher scores reflecting greater severity. Item severity/frequency is rated on a 5-point Likert scale (0 does not apply to 4 applies most of the time). This scale has shown good internal consistency, concurrent validity (Antony et al., 1998) and construct validity (Henry and Crawford, 2005). Depression, anxiety and stress have been found to be strongly correlated but conceptually and statistically separable, and associated with different correlates and outcomes in clinical and non-clinical groups (Clara et al., 2001; Henry and Crawford, 2005; Jovanović et al., 2019; Sinclair et al., 2012).

Wellbeing was measured using the 7-item Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS (Haver et al., 2015; Stewart-Brown et al., 2009)). Items are rated on a 5-point Likert scale (1 none of the time to 5 all of the time). This scale has adequate internal consistency (Haver et al., 2015). The total score was used as an observed variable. Higher scores reflect greater wellbeing.

Patient-reported diagnosis was coded as psychosis, schizophrenia, and schizoaffective disorder (‘psychosis’) and bipolar disorder, depression, anxiety disorder, EUPD, PTSD, ‘other’, no diagnosis (‘non-psychosis’). For sex, data from patients identifying as female or male were used.
2.4. Analytic approach

Using SPSS Version 25.0 (IBM Corp, 2017), the categorical HPSVQ variables showed minor skew but no serious distributional issues, no significant outliers (Cook, 1977), and were appropriate for factor analysis (Yong and Pearce, 2013). The four dependent variables approximated normal distributions with no significant multicollinearity. There was minimal missing data in HPSVQ items (See Table 2) and dependent variables (maximum 1.5% missing). Data were considered missing at random.

Hypothesis tests were conducted in Mplus Version 6.0 (Muthén and Muthén, 2010) using confirmatory factor analysis with mean and variance-adjusted Weighted Least Squares (WLSMV) estimation, a probit link, and THETA parameterization (Li, 2016). Probit coefficients were interpreted as; 0.1 = small, 0.3 = medium, 0.5 = large (Cohen, 1992). Missing data were managed using Full Information Maximum Likelihood (FIML) methods. No adjustment was made for multiple comparisons as corrections which do not adjust for the nonindependence of significance (e.g. Bonferroni) would be overly conservative, whereas less conservative tests (e.g. using a false discovery rate) would render multiple marginal values as significant. Model fit was examined using multiple fit indices: non-significant Chi-square goodness of fit statistic ($\chi^2$) or $\chi^2$/degrees of freedom ratio under 2; Comparative Fit Index (CFI) >.95, a Root Mean Square Error of Approximation (RMSEA) <.06, Tucker Lewis Index (TLI) >.90, and Weighted Root Mean Square Residual (WRMR) <1.0 (DiStefano et al., 2018; Hu and Bentler, 1999; Tabachnick and Fidell, 2007). Nested model iterations were compared using the Mplus DIFFTEST procedure (Muthén and Muthén, 2010). A significant DIFFTEST suggests the more restrictive model provides significantly poorer fit. Further model modifications were guided by modification indices (MIs). MIs reflect the expected reduction in the $\chi^2$ statistic through freeing a fixed-to-zero parameter (Muthén and Muthén, 2010). MIs were examined sequentially, in order of predicted $\chi^2$ reduction. Modifications were only implemented if theoretically grounded, permissible (e.g. would not de-identify the model) and significantly improved model fit (Kline, 2011). MI examination ceased when excellent fit was achieved across all model fit statistics. Mediation was tested using indirect effects and bias-corrected bootstrapped (5000 resamples) 95% confidence intervals.

Measurement and structural model invariance were examined using a multi-group approach (psychosis/non-psychosis and female/male) in which parameters were iteratively constrained to equality and the DIFFTEST procedure applied. A significant DIFFTEST result suggests parameter variance across groups. First, configural invariance was tested (measurement model fit within each group) followed by metric (equality of factor loadings) and scalar
invariance (equality of item thresholds, residual variances, factor means) (Gregorich, 2006; Muthén and Muthén, 2010). A threshold is the expected value of the latent response variable reflected in the individual’s transition from one categorical indicator response to another (Muthén and Muthén, 2010), i.e. analogous to the intercept for a continuous variable. Finally, structural invariance was tested by constraining to factor variances and indirect effects to equality.

3. Results

3.1. Measurement model

INSERT TABLE TWO HERE

HPSVQ descriptive statistics are shown in Table 2. A one-factor solution provided significantly poor fit than the theoretical two-factor model (Kim et al., 2010) according to the $\chi^2$ difference test. The two-factor solution provided excellent fit according to the CFI, TLI and WRMR, but poor fit according to the $\chi^2$ and RMSEA (Table 3). After examining MIs, two alterations were made. First, item 5 (interference with life) was allowed to cross-load onto both factors. The cross-loading significantly improved fit and was retained. Secondly, a residual covariance between item 1 (frequency) and item 4 (duration) was added, which significantly improved model fit and was retained (Table 3). These modifications made respective negligible impact (< .05) on all other standardised indicator loadings. The factors explained a large amount of variance (38.9-85.8%) in their indicators (Figure 1). The final model, the two-factor solution with one cross-loading item and one error covariance, achieved excellent fit across all indices. The first factor was labelled “Voice severity”; comprising frequency, loudness, duration, clarity, obey commands and interference with life, and the second, “Voice-related distress”; comprising distress, negative self-appraisal, negative content, and interference with life.

3.2. Structural model (criterion validity)

Voice-related distress significantly mediated the associations between voice severity and emotional distress and wellbeing. The indirect (mediator) effects via were: between severity and wellbeing $ab = -0.35$, $ab = -1.94 [-3.05, -1.05]$, $p < .001$; between severity and depression $ab = 0.46$, $ab = 2.82 [1.84, 4.02]$, $p < .001$; between severity and anxiety $ab = 0.35$, $ab = 1.89 [1.01, 2.81]$, $p < .001$; between severity and stress $ab = 0.27$, $ab = 1.39 [0.47, 2.41]$, $p = .001$. All total effects were significant and all direct effects, except from
severity to stress, were non-significant. Voice severity explained a large proportion of variance in voice-related distress (68.5%) and the overall model explained a moderate amount of variance in emotional distress and wellbeing (22.3-29.4%).

3.3. Model invariance

There was full measurement (metric and scalar) invariance across diagnoses. With respect to gender, there was partial metric invariance, for constraining factor loadings to equality significantly reduced model fit. MIs suggested items 8 (clarity) and 9 (obey commands) were variant. Freeing these loadings indicated that for females, voice severity was less strongly associated with clarity ($\beta = 0.64, p < .001, R^2 = 42.5\%$) and more strongly with obey commands ($\beta = 0.69, p < .001, R^2 = 45.8\%$) than for males (clarity $\beta = 0.79, p < .001, R^2 = 59.5\%;$ obey commands $\beta = 0.52, p < .001, R^2 = 27.6\%$). There was partial scalar invariance, for constraining thresholds to equality significantly reduced model fit. MIs suggested freeing HPSVQ item 7 (negative self-appraisal) thresholds at 1 (no voices make me feel bad), 2 (a little bit) and 3 (moderately). A smaller proportion of females (11-12%) had selected lower scores, whereas males were very evenly spread across all five response points (18-23%). Table 4 presents configural and final metric and scalar models.

Testing structural invariance revealed equal variability in both factors across diagnoses ($\chi^2(2) = 1.58, p = 0.45$), but a significantly lower factor mean for voice severity in psychosis ($M = -0.28, p = 0.03$) compared to non-psychosis ($M$ constrained to 0). Voice-related distress ($M = -0.23, p = 0.06$) was marginally lower in psychosis. Factor variances were equivalent across sex ($\chi^2(2) = 0.74, p = 0.69$), but females had significantly greater voice severity ($M = 0.34, p = 0.005$) and voice-related distress ($M = 0.28, p = 0.01$).

All indirect effects were equivalent across diagnosis and all but one across sex. The indirect effect from voice severity to anxiety significantly differed (Wald $\chi^2(1) = 6.80, p = 0.01$). There was a significant from voice-related distress to anxiety in males ($\beta = 0.61, p < .001$), but not in females ($\beta = 0.04, p = 0.82$). Therefore, for females, the direct association between voice severity and anxiety was stronger ($c' = 0.40, c' = 2.08, p = 0.01$) and the indirect effect non-significant ($ab = 0.03, ab = 0.16 [-1.30, 1.61], p = 0.82$). Amongst males the reverse was
true; there was a significant indirect ($ab = 0.49$, $ab = 2.61 [1.69, 3.82]$, $p < .001$) but not
direct effect ($c' = -0.08$, $c' = -0.44$, $p = 0.54$). $R^2$ values suggested voice severity explained
similar variance in voice-related distress for females (68.8%) and males (64.7%), but the
structural model explained less variance in emotional distress and wellbeing for females
(18.8-24.1%) versus males (26.8-37.6%).

4. Discussion

We aimed to test the fit of a two-factor HPSVQ model originally established in a South-
Korean sample of people with a psychosis diagnosis (Kim et al., 2010) in a transdiagnostic
UK sample. The two-factor model was significantly better than a simpler one-factor model.
Within the two-factor model, one factor reflected physical characteristics, interference with
life and obey commands (we named this “voice severity”), and the other reflected negative
content and emotional response items (we named this “voice-related distress”). We modified
Kim and colleagues’ model (2010) in cross-loading ‘interference with life’ onto both factors
and correlating errors for frequency and distress. Our findings support the separation of
physical voice severity versus more emotional voice-related distress, with interference with
life relevant to both. Possibly ‘interference with life’ is multidimensional (Ratcliff et al., 2011),
for example, voices may cause physical (e.g. voice loudness undermining attention and
concentration) and emotional interference (e.g. voice-distress undermining emotional
engagement in occupations and relationships). The residual covariance between frequency
and duration likely reflects shared method variance (Bollen, 2019). Frequency and duration
would become indistinguishable at the maximum level, i.e. at their most extreme, both items
would effectively collapse into ‘continuous’. Moreover, difficulties with time perception
associated with psychotic experiences (Ueda et al., 2018) could influence “frequency” and
“duration” responses without affecting other indicators. An ideal model might include no
cross-loadings nor correlated residuals, however, such a model may sacrifice validity for
parsimony (Bollen, 2019; Kline, 2011). A key risk of correlated residuals, the decrease in
factor loadings for corresponding indicators and inflation in remaining indicators (Yuan,
Kouros, Kelley, 2008), did not occur in the present case. Therefore, the two-factor HPSVQ
model provides a fairly simple solution, with criterion validity, high factor loadings and cross-
cultural applicability; albeit with further refinement needed in indicators capturing
interference, frequency and duration.

As we predicted, our cross-sectional data are in line with the theoretical model that voice
severity is associated with emotional distress and wellbeing through voice-related distress.
Indirect effects through voice-related distress to emotional distress and wellbeing were all
medium-sized and significant. Most direct effects from voice severity were non-significant. This model is consistent with a broad CBT approach to voice-hearing, within which voice-related distress is an important mechanism of broader distress and poor wellbeing amongst patients who hear voices and a key treatment target (Birchwood and Trower, 2006; McCarthy-Jones et al., 2013; Thomas, 2014).

Our findings suggest measurement and structural invariance across diagnoses, supporting a commonality in the phenomenology and impacts of voice-hearing for people with and without psychosis (Hazell et al., 2018a; Thomas et al., 2014; Waters et al., 2012; Waters and Fernyhough, 2017). However, at the latent level, voice severity was significantly higher for people without psychosis. Possibly, greater antipsychotic medication use ameliorates voice severity in people with psychosis. By extension, the emotional dampening effects of antipsychotics could have occluded existing differences between participants with and without psychosis in voice impact. Voice impact may therefore be greater in antipsychotic-naïve participants with psychosis compared to voice-hearers without psychosis.

There was partial measurement and structural invariance across sex. Voice severity overall was greater for females, in support of our hypotheses and prior evidence that women have more frequent and severe hallucinations, with a greater propensity for persecutory or ‘delusional’ interpretation of voices (González et al., 2008; Leung & Chue, 2000; Thorup et al., 2007). Obey commands also loaded more strongly onto voice severity for women compared to men. This may be attributable to females perceiving their voices as more powerful and malevolent (Hayward et al., 2016); meaning voice severity exerts a stronger influence upon compliance. Voice-related distress in females was higher overall and seemed to have greater impact on negative self-appraisals, in keeping with women’s greater emotional reactivity, intensity and expressivity, especially regarding negative emotions (Skowron & Dendy 2004; Gard & Kring 2007; Deng et al. 2016). Contrary to our predictions that the mediating role of voice-related distress would be stronger for females than males, most indirect effects were equivalent across sex. However, there was no significant indirect effect between voice severity and anxiety for females. It could be that females’ greater emotionality and anxiety-proneness (Deng et al., 2016; Gard and Kring, 2007; McLean et al., 2011) creates such strong reactions to voice severity that this manifests directly in general anxiety, with voice-related distress making no additional impact; whereas males may suppress the expression of anxiety.

4.1. Limitations
There are several important limitations. First, nine HPSVQ items necessarily constraints the factor solution to a maximum of two to three factors (Kline, 2011). More items may have pragmatically advantageous in minimising participant burden. Secondly, the Likert nature of the HPSVQ constrains participant responses, a continuous response scale is more sensitive and could produce a different factor structure. Thirdly, we recorded only patient-reported diagnosis and thus there may have been issues with patient recall or mis-diagnosis. Moreover, recording and covarying antipsychotic medication usage could have altered the diagnostic invariance model. It is also possible that substance misuse may vary diagnostically and therefore recording and covarying this variable may have altered model parameters. Additionally, with our cross-sectional design we could test whether our data supported a theoretical mediation model but could not test causal relationships. Moreover, our current sample size precluded using the optimal latent bi-factor model for depression, anxiety and stress (Zanon et al., 2020). Therefore, further study is needed to confirm whether voice severity would directly predict stress—and anxiety for women—after parsing out shared general variance across emotional distress dimensions. Finally, as we did not control for multiple comparisons, it is possible that current effects may include type I errors. Nevertheless, the current study offers many strengths. It is the largest validation of the HPSVQ’s factor structure and the only known study using SEM multi-group invariance techniques to examine diagnostic and sex invariance in voice-hearing and associations with emotional distress and wellbeing; outcomes of paramount importance to patients (Byrne et al., 2010).

4.2. Clinical and research implications

The key clinical implication of the current study is the broad applicability of a CBT model of voice-hearing—and interventions which target voice-related distress through changing how one responds and relates to voices—across diagnoses. Our findings also suggest women’s more severe and distressing voices perhaps confer greater treatment need. Moreover, the pattern of direct effects from voice severity points toward a need to consider adapting treatment provision for voice-hearers with markedly high stress and (for females) anxiety. For these populations, interventions that more directly target voice severity may be especially beneficial. For example, treatments could include behavioural approaches which capitalise on behaviours and situations within which voices are less severe, i.e. quieter, less present and impactful, such as Coping Strategy Enhancement (Hayward et al., 2017) and interventions which focus on facilitating responding in a less resistant manner to voices that elicit strong emotional reactions, such as mindfulness-based interventions, (Chadwick et al.,
We suggest that supplementing the HPSVQ measurement of voice-hearing with a measure of delusional beliefs may helpfully inform the wider clinical picture, allowing consideration of how unusual beliefs may impact on voice-related beliefs and responses, and would be appropriate to capture outcomes if voice-hearing is targeted not in a symptom-specific intervention (CBTv), but as part of a broader CBT for psychosis intervention.

There are four key areas for further research. First, we recommend ongoing refinement of the HPSVQ as the two-factor structure may leave important variance still to be explained. Further explorations should consider how additional items may identify further latent dimensions of voice-hearing, such as through conceptualising multiple dimensions of interference with life. Secondly, we recommend testing the structural model fit with data from other cultures including South Korea, as symptom-related distress may differ markedly cross-culturally (Wüsten et al., 2018). Moreover, we recommend recruiting sufficiently large samples to allow replication of the present model with depression, anxiety and stress (DASS-21) using a latent bi-factor emotional distress structure (Zanon et al., 2020) and modern approaches to controlling multiple comparisons (Korthauer et al., 2019). Thirdly, we encourage consideration of putative explanatory mechanisms in how voice severity influences voice-related distress, including interpersonal and social schema, appraisals of voice power and content, traumatic memories and imagery, and coping strategies (Cole et al., 2017; Hardy et al., 2016; Hayward et al., 2017; Hayward & Fuller, 2010; Thomas et al., 2014). Possibly, the importance and effects of such mechanisms differ by diagnosis and sex. Finally, efforts should be made to explore not only dimensions within voice-hearing, but clustering within people and the association of such clusters with diagnosis and sex, i.e. the invariance of latent subtypes of voice-hearing (McCarthy-Jones et al., 2014), and non-recursive interactive temporal symptom networks (Borsboom, 2017; Pappa et al., 2019). Such models may best fit the data and have novel implications for both the origin and treatment of voice-hearing (Smailes et al., 2015).

Contributors:

CB, HN and MH conceived the aims and design of the current study. CB and HN conceived the analytic plan. CB performed the analysis and wrote the first draft of the manuscript. HN, CS, AR, TL, BS and MH provided reviews, written contributions and approval for the
submitted manuscript. MH, CS, TL, and BS conceived of the design and aims of the Approve study which gave rise to the data used in this manuscript. MH co-ordinated the delivery of the Approve study including participant recruitment, data collection, and data management.

Acknowledgement:

We wish to give thanks to all patients who gave their time to participate in the Approve study and all research and health professional staff, carers and other people who supported this participation.

Role of the funding source:

This work was supported by a doctoral studentship award made to Aikaterini Rammou by the Economic and Social Research Council (ES/J500173/1). Neither the funder nor sponsor (University of Sussex) had any role in study design, collection, analysis and interpretation of data, writing of the report or decisions to submit this article for publication.

Conflicts of interest: None.

Data Statement:

No new data were produced in this study. Enquiries about accessing data from the Approve study, which generated the original dataset accessed for the present study, should be directed to Dr Mark Hayward (mark.hayward@sussexpartnership.nhs.uk).

References


Short title: HPSVQ validation, measurement and structural invariance

https://doi.org/10.1177/1403494815588862


IBM Corp, 2017. IBM SPSS Statistics for Windows.


**Table 1: Demographic characteristics of participants**

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M(SD)</td>
<td></td>
</tr>
<tr>
<td>18-28</td>
<td>96 (24.1)</td>
</tr>
</tbody>
</table>
### Gender
- Male: 244 (61.3)
- Female: 151 (37.9)
- Other: 2 (0.5)
- Prefer not to say: 1 (0.3)

### Ethnicity
- White (British): 341 (85.3)
- White Other: 9 (2.3)
- Black/African/Caribbean/Black British: 8 (2.0)
- Asian/Asian British: 16 (4.0)
- Mixed Ethnicity: 11 (2.8)
- Other: 14 (3.5)
- Prefer not to say: 1 (0.3)

### Self-reported diagnosis
- Psychosis/Schizophrenia: 241 (60.1)
- Schizoaffective disorder: 44 (11.0)
- Bipolar disorder: 24 (6.0)
- Depression: 57 (14.2)
- Anxiety disorder: 28 (7.0)
- Emotionally Unstable Personality Disorder: 37 (9.2)
- Post-traumatic Stress Disorder: 20 (5.0)
- Other: 27 (6.7)
- No diagnosis: 45 (11.2)

*Note. Some participants self-reported multiple diagnoses.*

<table>
<thead>
<tr>
<th>Table 2: Descriptive statistics for HPSVQ items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M</strong></td>
</tr>
<tr>
<td>1. How frequently did you hear a voice or voices? <em>(Frequency)</em></td>
</tr>
<tr>
<td>2. How bad are the things the voices say to you? <em>(Negative content)</em></td>
</tr>
</tbody>
</table>
**Table 3: Fit indices of measurement model iterations.**

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\chi^2$/df</th>
<th>RMSEA</th>
<th>TLI</th>
<th>CFI</th>
<th>WRMR</th>
<th>DIFFTEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-factor model</td>
<td>241.94</td>
<td>27</td>
<td>8.96</td>
<td>0.14</td>
<td>[0.13, 0.16]</td>
<td>0.95</td>
<td>0.97</td>
<td>1.41</td>
</tr>
<tr>
<td>Two-factor model</td>
<td>104.88</td>
<td>26</td>
<td>4.03</td>
<td>0.09</td>
<td>[0.70, 0.11]</td>
<td>0.98</td>
<td>0.99</td>
<td>0.85</td>
</tr>
<tr>
<td>Cross-loading item model</td>
<td>81.07</td>
<td>25</td>
<td>3.24</td>
<td>0.08</td>
<td>[0.06, 0.09]</td>
<td>0.99</td>
<td>0.99</td>
<td>0.74</td>
</tr>
<tr>
<td>Cross-loading item and error covariance model</td>
<td>55.26</td>
<td>24</td>
<td>2.30</td>
<td>0.06</td>
<td>[0.04, 0.08]</td>
<td>.099</td>
<td>1.00</td>
<td>0.59</td>
</tr>
</tbody>
</table>

*Note.* $\chi^2$/df = chi-squared/degrees of freedom; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker Lewis Index; CFI = Comparative Fit Index; WRMR = Weighted Root Mean Square

Residual; *a*Theoretical model (Kim et al., 2010); *b*; HPSVQ item 5 (interference with life) cross-loading; *c*HPSVQ item 1 (frequency) and 4 (duration) residual covariance.

---

**Table 4: Measurement invariance across self-reported diagnosis and gender.**
### Diagnostic invariance

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>$\chi^2$/df</th>
<th>RMSEA</th>
<th>TLI</th>
<th>CFI</th>
<th>WRMR</th>
<th>DIFFTEST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Configural model 1</strong></td>
<td>140.46</td>
<td>1.87</td>
<td>0.07</td>
<td>0.99</td>
<td>0.99</td>
<td>1.34</td>
<td>-</td>
</tr>
<tr>
<td><strong>Psychosis</strong></td>
<td>84.33</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Non-psychosis</strong></td>
<td>56.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Metric model 2</td>
<td>149.80</td>
<td>1.80</td>
<td>0.06</td>
<td>0.99</td>
<td>0.99</td>
<td>1.43</td>
<td>12.03(8), $p = 0.15$</td>
</tr>
<tr>
<td>Scalar model 3</td>
<td>146.19</td>
<td>1.62</td>
<td>0.06</td>
<td>0.99</td>
<td>0.99</td>
<td>1.48</td>
<td>8.46(9), $p = 0.48$</td>
</tr>
</tbody>
</table>

### Gender invariance

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>$\chi^2$/df</th>
<th>RMSEA</th>
<th>TLI</th>
<th>CFI</th>
<th>WRMR</th>
<th>DIFFTEST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Configural model 1</strong></td>
<td>187.40</td>
<td>2.50</td>
<td>0.09</td>
<td>0.98</td>
<td>0.98</td>
<td>1.61</td>
<td>-</td>
</tr>
<tr>
<td><strong>Females</strong></td>
<td>116.65</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Males</strong></td>
<td>70.76</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Metric model 2</td>
<td>193.56</td>
<td>2.39</td>
<td>0.08</td>
<td>0.98</td>
<td>0.98</td>
<td>1.64</td>
<td>7.19(6), $p = 0.30$</td>
</tr>
<tr>
<td>Scalar model 3</td>
<td>150.45</td>
<td>1.77</td>
<td>0.06</td>
<td>0.99</td>
<td>0.99</td>
<td>1.13</td>
<td>4.01(6), $p = 0.68$</td>
</tr>
</tbody>
</table>

**Notes:** $\chi^2$/df = chi-squared/degrees of freedom; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker Lewis Index; CFI = Comparative Fit Index; SRMR = Standardised Root Mean Square Residual; *Structural model (Figure 1) fitted to two groups, factor variances and residual variances fixed to 1, factor means fixed to 0, all factor loadings and thresholds estimated; **Final metric invariance model: factor loadings constrained to equal, factor variances fixed to 1 in the non-psychosis group and estimated in the psychosis group, factor means fixed to 0 and residual variances fixed to 1 in both groups, all item thresholds estimated; ***Final scalar invariance model: residual variances constrained to equal, factor loadings and thresholds constrained to equal, factor means fixed to 0, factor variances fixed to 1 in non-psychosis group and estimated in psychosis group; ****Final partial metric invariance model: As model 2b but factor variances fixed to 1 for males and estimated for female, factor loadings for items 8 and 9 freed for partial metric invariance. *****Final partial scalar invariance model: As model 2c but factor thresholds for item 7 freed for partial scalar invariance, factor means fixed to 0, factor variances fixed to 1 for males and estimated for females.
Figure 1. Final structural model.

Notes: All loadings and coefficients depicted are standardised betas. Indirect effects and covariances between observed dependent variables are not included in the diagram.
Short title: HPSVQ validation, measurement and structural invariance