







## Measurement practices in hallucinations research

David Smailes <sup>a</sup>, Ben Alderson-Day <sup>b</sup>, Cassie Hazell <sup>c</sup>, Abigail Wright<sup>d,e</sup> and Peter Moseley <sup>a</sup>

<sup>a</sup>Department of Psychology, Northumbria University, Newcastle upon Tyne, UK; <sup>b</sup>Department of Psychology, Science Laboratories, Durham University, Durham, UK; <sup>c</sup>School of Social Sciences, University of Westminster, London, UK; <sup>d</sup>Center of Excellence for Psychosocial and Systemic Research, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA; <sup>e</sup>Harvard Medical School, Boston, MA, USA

### ABSTRACT

**Introduction:** In several sub-fields of psychology, there has been a renewed focus on measurement practices. As far as we are aware, this has been absent in hallucinations research. Thus, we investigated (a) cross-study variation in how hallucinatory experiences are measured and (b) the reliability of measurements obtained using two tasks that are widely employed in hallucinations research.

**Method:** In Study 1, we investigated to what extent there was variation in how the Launay-Slade Hallucination Scale (LSHS) has been used across 100 studies. In Study 2, we investigated the reliability of the measurements obtained through source monitoring and signal detection tasks, using data from four recent publications. Materials/data are available at doi: 10.17605/osf.io/d3gnk/.

**Results:** In Study 1, we found substantial variation in how hallucinatory experiences were assessed using the LSHS and that descriptions of the LSHS were often incomplete in important ways. In Study 2, we reported a range of reliability estimates for the measurements obtained using source monitoring and signal discrimination tasks. Some measurements obtained using source monitoring tasks had unacceptably low levels of reliability.

**Conclusions:** Our findings suggest that suboptimal measurement practices are common in hallucinations research and we suggest steps researchers could take to improve measurement practices.

### ARTICLE HISTORY

Received 26 March 2021  
Accepted 22 October 2021


### KEYWORDS

Hallucinations; psychosis; measurement; open science

## Introduction

Over the past decade, psychology (like many other disciplines, such as oncology; Begley & Ellis, 2012) has faced a so-called ‘replication crisis’, where many findings have been shown to not be replicable (see Nosek et al., 2015). This has been followed by attempts to address these problems by engaging in methodological practices that increase the

**CONTACT** David Smailes  dave.smailes@gmail.com  Department of Psychology, Northumbria University, Newcastle upon Tyne NE1 8ST, UK

 Supplemental data for this article can be accessed <https://doi.org/10.1080/13546805.2021.1999224>

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

transparency, reproducibility, and replicability of psychological science (Nelson et al., 2018).

Initially, attempts to address the causes of the ‘replication crisis’ focussed on issues such as low statistical power (e.g. Bakker et al., 2012) and flexibility in how data are analysed (e.g. Simmons et al., 2011). More recently, it has been argued that the measurement practices psychology researchers engage in warrant increased attention, as accurately measuring our constructs-of-interest is a foundational part of psychological science (de Beurs & Fried, 2021). That is, there is little value in running a high-powered study, in which data are analysed in a transparent, reproducible manner, if we have not effectively measured the variables we are interested in (Flake & Fried, 2020). The difficulties inherent in measuring hallucinations and hallucinatory experiences – as a result, for example, of the complexity and range of experiences participants report (Woods et al., 2015) – mean that it should be especially valuable for there to be a focus on optimal measurement practices in hallucinations research.

Good measurement practices are typically adopted during the initial development of self-report measures of psychological variables, with researchers providing evidence that these measures are reliable and valid assessments of the construct-of-interest. However, it appears that this good practice often does not extend beyond the initial development of self-report questionnaires (Flake & Fried, 2020). One concern is the “jingle fallacy” (Block, 1995), which refers to situations where scales purporting to measure the same construct actually measure different constructs (e.g. the symptoms measured by different depression scales vary so much that these questionnaires may be measuring different constructs; Fried, 2017). This becomes a problem when trying to synthesise inconsistent findings across apparently similar studies, as it is difficult to discern whether cross-study inconsistency is a result of these studies measuring different psychological constructs, or is a result of other factors (e.g. sampling error).

A second concern is the practice of making modifications to a scale following its publication (e.g. changing a scale’s response options, adding/dropping items, revising scoring procedures; Flake et al., 2017). While there may be good reasons for these modifications (e.g. revising items so they are easier to comprehend), these changes may reduce validity and may allow for a degree of analytic flexibility that using an unmodified scale would have prevented (Flake & Fried, 2020). A final concern is sub-optimal reporting of which measures have been employed and how (e.g. failing to report which version of a questionnaire has been employed). There is evidence that this sub-optimal reporting is quite common (Flake et al., 2017); this is important as it may hide cross-study inconsistency in the scales employed, making it more difficult for readers to be confident that constructs have been measured in a reliable and valid manner.

In contrast to the development of self-report measures, where there is at least an initial focus on good measurement practices, little attention appears to be paid to optimal measurement practices when researchers develop tasks to assess putative cognitive biases. For example, Parsons et al. (2019) have argued that studies that employ novel, unstandardised tasks to assess cognitive biases in clinical samples almost always fail to examine and report the psychometric properties of these tasks. As a result, we often have no knowledge of the reliability of the measurements obtained from these tasks. This is problematic because without knowing how reliably a construct has been measured in a study, readers cannot know how confident they should be in that study’s findings.

More problematically, it is possible that this lack of transparency “hides” a situation where the reliability of most of the measurements obtained from tasks that aim to assess cognitive biases is low. This is a concern because, as shown by Parsons et al. (2019), if the reliability of the measurements we obtain from the tasks we employ is low, the level of statistical power achieved in our studies is substantially reduced. This reduction in power increases the likelihood of “false negative” findings (incorrectly accepting the null hypothesis) and (counter-intuitively) the likelihood of “false positive” findings (incorrectly rejecting the null hypothesis; Bakker et al., 2012). Thus, suboptimal measurement practices around our development and use of tasks can directly contribute to the generation of research findings that fail to replicate.

These problematic or questionable measurement practices have been investigated in sexual behaviour (Kohut et al., 2020), addiction (King et al., 2020), and anxiety research (Waechter & Stolz, 2015), but have not yet been examined in hallucinations research. Attending to measurement practices in hallucinations research is likely to be worthwhile, given (a) the nature of the experiences hallucinations researchers are interested in (i.e. they are often difficult for participants to describe and sometimes have overlapping phenomenology with other types of unusual cognitions, such as intrusive thoughts; Morrison, 2001; Woods et al., 2015), which makes these experiences difficult to measure effectively and (b) the frequent use of tasks to measure cognitive biases that may be related to the presence/frequency of hallucinations. Thus, across two studies, this paper aims to examine measurement practices in hallucinations research.

## Study 1 introduction

A common approach in hallucinations research is to examine factors associated with non-clinical hallucinatory experiences (HEs; Larøi, 2012). Adopting this approach allows researchers to try to understand the factors involved in the development of hallucinations, without the confounding effects that medication use and/or the long-term presence of hallucinations may have when investigating hallucinations in participants with mental health problems (Badcock & Hugdahl, 2012).

Many measures of HEs have been developed, but perhaps the most widely used is the Launay-Slade Hallucination Scale (LSHS; Launay & Slade, 1981). Since its initial development, English-language revisions to this 12-item questionnaire have been published at least four times, by Bentall and Slade (1985), Morrison et al. (2000), Morrison et al. (2002), and McCarthy-Jones and Fernyhough (2011). As can be seen in Table 1, over the course of four revisions, the LSHS has changed in important ways, including the wording of items, the response options presented, and the number of items presented. Some versions (McCarthy-Jones & Fernyhough, 2011) have tried to maintain a narrow focus on HEs, while others (Morrison et al., 2000, 2002) have assessed a broader set of unusual experiences, (e.g. vivid daydreaming). Thus, it is possible that these scales measure different constructs. As a result, when using different versions of the LSHS, hallucinations researchers may commit the jingle fallacy, and this may make it difficult for the hallucinations research community to develop cumulative science. That said, each version of the LSHS appears to have at least reasonable levels of construct validity (e.g. in terms of its correlations with related variables).

**Table 1.** Summary of the key features of five versions of the Launay-Slade Hallucination Scale.

Version of LSHS	Key features/revisions
Launay and Slade (1981)	<ul style="list-style-type: none"> <li>- 12-item scale.</li> <li>- Unidimensional.</li> <li>- Yes/No response options.</li> <li>- Items described as assessing 'vivid thoughts', 'intrusive thoughts', 'vivid daydreams', 'auditory hallucinations' and 'visual hallucinations'.</li> </ul>
Bentall and Slade (1985)	<ul style="list-style-type: none"> <li>- 12-item scale.</li> <li>- Minor revisions to wording of some items.</li> <li>- Response options ranging from 'Certainly does not apply' (1) to 'Certainly applies' (5).</li> </ul>
Morrison et al. (2000)	<ul style="list-style-type: none"> <li>- 13-item scale.</li> <li>- Minor revisions to wording of some items.</li> <li>- Three items deleted.</li> <li>- Four items introduced.</li> <li>- Response options ranging from 'Never' (1) to 'Almost always' (4).</li> <li>- Two factor structure: one factor assessing "auditory or verbal hallucinations/daydreaming" and one factor assessing "visual hallucinations/disturbances".</li> </ul>
Morrison et al. (2002)	<ul style="list-style-type: none"> <li>- 24-item scale.</li> <li>- Two items deleted.</li> <li>- 13 items introduced.</li> <li>- Response options ranging from 'Never' (1) to 'Almost always' (4).</li> <li>- Three factor structure: one factor assessing "vividness of imagination and daydreaming, one factor assessing "visual disturbances and hallucinations", and one factor assessing "auditory hallucinations".</li> </ul>
McCarthy-Jones and Fernyhough (2011)	<ul style="list-style-type: none"> <li>- Nine-item scale.</li> <li>- 15 items deleted.</li> <li>- Response options ranging from 'Never' (1) to 'Almost always' (4).</li> <li>- Two factor structure: one factor assessing "predisposition to auditory hallucinations" and one factor assessing "visual hallucinations and disturbances".</li> </ul>

The first aim of Study 1, then, was to examine how often different versions of the LSHS have been used in 100 recent publications. The second aim of the study was to investigate how these different versions of the LSHS have been used from study-to-study, in terms of the use of subscale scores, rather than full-scale scores. In addition, we aimed to examine how often, and in what ways, these scales have been modified by researchers, and where modifications were made, whether/how these modifications were justified. Finally, we aimed to record how often too little information was provided in a publication for us to fully understand what scale had been used.

## Study 1 materials and method

### *Literature search*

We used five articles (Bentall & Slade, 1985; Launay & Slade, 1981; McCarthy-Jones & Fernyhough, 2011; Morrison et al., 2000; Morrison et al., 2002) that described the development of different versions of the LSHS as "seed papers" for our literature search. We used the electronic database Scopus to identify studies that had cited at

least one of the five seed papers as of 25/SEP/2020 and then merged these five citation lists. The list was ordered chronologically and, beginning with the most recent publication, the abstract or full text of each publication was reviewed (the full-text of a publication was always read, unless it was clear from the abstract that the publication was a review) to examine whether the study had employed a version of the LSHS, until we had identified 100 studies that had employed an English language version of the LSHS. We excluded from this review studies that employed a translated version of the LSHS, as we did not have the expertise to verify how items have been translated (which was relevant to our coding of whether any items had been revised) studies that reported no other analysis than a factor analysis (as we were unable to code these studies in terms of their use, for example, of subscale scores), and studies that reported no inferential tests.

### **Coding of studies**

After refining a draft coding system, all studies were reviewed and coded by one author (DS), with 20 studies also being coded by a second author (AW), to establish inter-rater reliability. Acceptable levels of inter-rater reliability were achieved (see [osf.io/d3gnk/](https://osf.io/d3gnk/)). We coded each study that had employed an English-language version of the LSHS in terms of the 12 variables outlined in [Table 2](#). While several of the variables should have been predictable based on the version of the LSHS that was cited in a study, we anticipated that the way in which the LSHS was used in a study may not always be consistent with the cited version. The coding system is presented in [Table 2](#).

### **Study 1 results and discussion**

Our literature search identified 694 results/studies. To reach our target of 100 studies, we screened the full-texts of 397 studies, with the oldest published in 2010. Most of these studies employed a correlational design and had recruited a non-clinical sample. However, around 25% recruited a clinical, help-seeking, or voice-hearing sample. In many instances, studies did not report a measure of the internal reliability of the version of the LSHS they employed. However, when they did so, internal reliability was almost always acceptable. The full list of search results, reviewed studies, as well as how they were coded is available at [osf.io/d3gnk/](https://osf.io/d3gnk/), with our findings summarised in [Table 3](#).

Several findings are apparent from the 100 studies we reviewed. First, four (Bentall & Slade, 1985; Launay & Slade, 1981; McCarthy-Jones & Fernyhough, 2011; Morrison et al., 2002) of the five versions of the LSHS we used as “seeds” were used quite frequently (24%, 40%, 17%, and 14%, respectively). That is, the field does not appear to have “settled” on a preferred version of the LSHS. The use of multiple versions of the LSHS is important as different versions of this scale appear to measure slightly different constructs. For example, while the Launay and Slade/Bentall and Slade versions include items concerning intrusive thoughts and vivid daydreaming, the McCarthy-Jones and Fernyhough version does not. Second, over and above the use of multiple versions of the LSHS, there was substantial variation in the lengths of the scales employed (14 different item-lengths were reported), ranging from one-item versions to 24-item versions.

**Table 2.** Coding system for assessing variation in the use of the launay-slade hallucination scale (LSHS) across 100 studies.

Code	Notes
Which version of the LSHS was cited?	Initially, where more than one version of the LSHS was cited, we intended to establish which version of the LSHS the scale employed most closely resembled. However, many reviewed studies provided little information about the scale they employed, causing us to adapt our coding system. Instead, where more than one version of the LSHS was cited, we recorded the most recent version cited as the scale that was employed, as we assumed that this was the most likely version used.
How many items does the scale consist of?	Where this was not clearly reported, we coded this as 'unclear'.
How many items were revised?	Where this was not reported, we assumed that no items were revised.
What do the scale response options refer to?	We coded this in terms of level of agreement, how much an item applies to the participant, or frequency. Where this was not clearly reported, we coded this as 'unclear'.
What does the 'lowest' response option refer to?	We coded this variable, as we thought that even when response options may have assessed the same concept (e.g., frequency), they may have employed different response options (e.g., Never versus Very Rarely). Where this was not clearly reported, we coded this as 'unclear'.
What does the 'highest' response option refer to?	We coded this variable, as we thought that even when response options may have assessed the same concept (e.g., frequency), they may have employed different response options (e.g., Almost Always versus Every Day). Where this was not clearly reported, we coded this as 'unclear'.
What was the 'lowest' response option score?	We coded this variable, as we thought that even when response options may have assessed the same concept (e.g., frequency), they may have employed different response options (e.g., Never to Almost Always assessed on a 1-4 Likert scale, or Never to Almost Always on a 0-7 Likert Scale). Where this was not clearly reported/easy to calculate from a table of min-max scores, we coded this as 'unclear'.
What was the 'highest' response option score?	We coded this variable, as we thought that even when response options may have assessed the same concept (e.g., frequency), they may have employed different response options (e.g., Never to Almost Always assessed on a 1-4 Likert scale, or Never to Almost Always on a 0-7 Likert Scale). Where this was not clearly reported/easy to calculate from a table of min-max scores, we coded this as 'unclear'.
How many LSHS scores are used as variables in inferential statistical analyses?	We coded this variable to examine whether, for example, correlations were reported for a full-scale score, as well as two subscale scores. Item-by-item prevalence estimates were not included.
Was a full-scale score used?	We coded this variable in terms of yes/no. A 'yes' code was used if a score based on a 12, 13, 24, or nine-item full-length LSHS was used. When coding this variable, we included analyses reported in supplementary analyses, as well as analyses reported in the full-text of the paper.
How many subscale scores were used?	We coded the number of LSHS scores used that were calculated by summing responses to a subset of items from a full-length LSHS (e.g., the five-item auditory subscale of McCarthy-Jones and Fernyhough's [2011] version of the LSHS). When coding this variable, we included analyses reported in supplementary analyses, as well as analyses reported in the full-text of the paper.
Where revisions have been made to the original scale, is a justification provided for the revision?	We coded this variable in terms of, not applicable, no, partial justification, or yes.

**Table 3.** Summary of variation in use of the launay-slade hallucination scale.

Characteristic	Percentage of studies
<i>Which version of the LSHS was cited?</i>	
Launay and Slade (1981)	24%
Bentall and Slade (1985)	40%
Morrison et al. (2000)	5%
Morrison et al. (2002)	17%
McCarthy-Jones and Fernyhough (2011)	14%
<i>How many items does the scale consist of?</i>	
1	1%
2	3%
3	1%
5	8%
6	1%
9	6%
11	1%
12	44%
13	1%
15	1%
16	3%
20	4%
21	1%
24	4%
Unclear	21%
<i>What do the scale response options refer to?<sup>a</sup></i>	
Yes/No	2%
To what extent the item applies	26%
To what extent you agree with the item	3%
How frequently you have the experience	22%
Unclear	48%
<i>What were the 'lowest' and 'highest' response option scores?<sup>a</sup></i>	
0-1	3%
0-4	29%
0-6	1%
1-4	20%
1-5	7%
Unclear	41%
<i>How many LSHS scores are used as variables in inferential statistical analyses?</i>	
0	2%
1	90%
2	5%
3	1%
5	1%
6	1%
<i>Was a full-scale score used?</i>	
No	20%
Yes	80%
<i>How many subscale scores were used?</i>	
0	78%
1	15%
2	5%
4	1%
5	1%
<i>Where revisions have been made to the original scale, is a justification provided for the revision?</i>	
NA	86%
No	8%
Partial	3%
Yes	3%

<sup>a</sup>Note that the sum of the percentages here exceeds 100 because one study employed response options that referred to frequency and to how much an item applied.



Third, in many studies, too little information was provided for us to code all of the variables we intended (e.g. response options were reported in only 52% of studies). It could be argued that it is unnecessary for authors to provide this information, as long as they cite the version of the LSHS that they have used. However, in some of the studies we reviewed, it appeared that while one version of the LSHS was cited, the response options employed were those used for a different version of the LSHS. For example, in nine of the 11 instances where authors reported using the Launay and Slade (1981) version of the LSHS and reported the response options presented to participants, the response options were not those used in the 1981 version of the LSHS. Meanwhile, in two of the 22 instances where authors reported using the Bentall and Slade (1985) version of the LSHS and reported the response options presented to participants, the response options were not those used in the 1985 version of the LSHS. Thus, for the sake of clarity and transparency, it would be useful for authors to report basic information about the scale they employed such as the response options presented to participants.

These findings are consistent with assessments of measurement practices in other sub-fields of psychological science where suboptimal measurement practices have been reported. This has been seen, for example, in depression (Mew et al., 2020), sexual behaviour (Kohut et al., 2020), and addiction (King et al., 2020) research, where a wide variety of different scales are used to assess the same variables-of-interest, and in social/personality science, where the scales employed in a study are often poorly described (e.g. in terms of the numbers of items and the response options presented to participants; Flake et al., 2017).

Our findings of suboptimal measurement practices and of variation in how HEs are measured have important implications for hallucinations research. Primarily, the combination of suboptimal measurement practices and the variation in how HEs are measured make it difficult to build a cumulative science, where others' research can be reproduced and replicated, and where researchers can easily synthesise data collected across different studies. This may, for example, account for some of the inconsistent findings reported by hallucinations researchers, such as the association between the frequency of hallucinatory experiences and the vividness of mental imagery reported by non-clinical participants (e.g. Aynsworth et al., 2017; Mitrenga et al., 2019). We return to this issue in the General Discussion.

This study had several limitations. First, we limited our sample of reviewed papers to only 100. Clearly, we would be able to generalise our findings more broadly had we sampled a larger number of studies. That said, our aim was to examine current measurement practices in hallucinations research and including more studies in our review (e.g. 250) would have meant reviewing older publications, which would have meant that our analysis would not reflect the field's current practices. Second, we excluded non-English versions of the LSHS from the review, as we were unable to verify whether meaningful changes had been made to the LSHS items as they were translated. It is unclear whether including non-English versions of the LSHS in the review would have increased or reduced the amount of variation in measurement practices we observed. Finally, our focus on the LSHS meant that we primarily reviewed studies involving non-clinical samples. It would be valuable to investigate measurement practices in hallucinations research with predominately clinical participants by, for example, examining variation

in how measures such as the Psychotic Symptoms Rating Scale (Haddock et al., 1999) are employed.

## Study 2 introduction

Cognitive models (e.g. Bentall & Fernyhough, 2008) suggest that people who experience auditory hallucinations have a bias where they mistake internal, self-generated cognitions for external, non-self-generated events. Two main families of tasks have been employed to test this claim – source monitoring paradigms and signal detection paradigms. Across a series of trials, these tasks require participants to judge whether an item was internal and/or self-generated, or was external and/or non-self-generated. In source monitoring tasks, these judgements are made during a testing phase, where participants must remember events from an earlier encoding phase. In signal detection tasks, participants make these judgements “in real time”.

While meta-analytic studies (Brookwell et al., 2013) have reported that there are medium-to-large associations between performance on these tasks and the presence of hallucinations/HEs, findings from individual studies have been inconsistent. The use of small sample sizes presumably plays a role in the inconsistent effects reported across studies. For example, the median sample size in the clinical studies synthesised in Brookwell et al.’s meta-analysis was 30; this sample size would only give us reasonably precise estimates of effect sizes if the “true” association between task performance and presence/frequency of hallucinations/HEs was around  $\rho = .70$  (Schönbrodt & Perugini, 2013). This issue may be compounded by suboptimal levels of reliability of the measures obtained from source monitoring and signal detection tasks. However, at present, it is unclear if this is the case, as we know very little about the reliability of the measurements obtained from source monitoring and signal detection tasks. The aim of Study 2, therefore, was to examine the reliability of the measurements obtained from source monitoring and signal detection tasks, using data from previously published studies. Given that we lack data that would allow us to examine test-retest reliability of the measures obtained from these tasks, we estimated the internal reliability of these measurements by examining their internal consistency.

## Study 2 method

### Datasets

We re-analysed data from four publications (five studies) – Smailes et al. (2015), Garrison et al. (2017), Alderson-Day et al. (2019), and Moseley et al. (2021). We selected these studies because we were able to access trial-level data from the tasks. Trial-level data is required to examine the internal consistency of the measurements obtained from a task (see Analyses section, below), and because these studies were authored/co-authored by at least one of the current study’s authors (BA-D, PM, or DS), we had full access to the datasets. In addition, we selected these studies because they are available as peer-reviewed publications, which allows us to explain their methods briefly here, and to direct readers to the original publications for more detailed information.

Two studies (Moseley et al., 2021; Smailes et al., 2015) employed signal detection tasks, and four employed source monitoring tasks (Garrison et al., 2017, Study 1; Garrison et al., 2017 Study 2; Alderson-Day et al., 2019; Moseley et al., 2021). In the Supplementary Materials we describe the tasks employed in brief (e.g. the number of trials employed), with more detail (e.g. the duration of each trial) available in the original publications.

Across all studies, we followed the approach recommended in Parsons et al. (2019) of reporting reliability estimates for variables that corresponded as closely as possible with the outcomes reported in previous research (e.g. we reported separate reliability estimates for the different groups created in Garrison et al., 2017, Study 1). For each study (or group within a study), we have reported two reliability estimates. For the signal detection tasks, we obtained a reliability estimate for the number of “hits” participants made, and a reliability estimate for the number of “false alarms” participants made. For the source monitoring tasks, we obtained a reliability estimate for the number of “internal misattributions” participants made (where a participant misremembered something that another person had generated or said, as something they had generated, said, or imagined), and a reliability estimate for the number of “external misattributions” participants made (where a participant misremembered something that they had generated, imagined, or said, as something another person had generated or said). In some instances, the number of participants reported in Table 4 differ slightly from the numbers reported in the original publication as the original publication may have only reported data from participants for which we had complete datasets for all of the variables measured in the study (in which case the N reported in Table 4 is larger than in the original publication), or because there were problems re-formatting some participants’ data so that it could be used to calculate a reliability estimate study (in which case the N reported in Table 4 is smaller than in the original publication).

**Table 4.** Reliability estimates across studies, samples, tasks, and outcomes.

Study name	Task/Sample	N	$r_s$ (95% CI) for Hits	$r_s$ (95% CI) for False Alarms
Smailes et al. (2015)	SDT/Non-clinical	139	0.809 (0.783 to 0.836)	0.853 (0.823 to 0.883)
Moseley et al. (2021)	SDT/Non-clinical	594	0.808 (0.797 to 0.819)	0.939 (0.919 to 0.958)
Study Name	Task/Sample	N	$r_s$ (95% CI) for Internal Misattributions	$r_s$ (95% CI) for External Misattributions
Garrison et al. (2017) Study 1	SMT: Imagined vs Perceived/Non-clinical LP Group	22	0.632 (0.578 to 0.686)	0.393 (0.339 to 0.448)
Garrison et al. (2017) Study 1	SMT: Imagined vs Perceived/Non-clinical HP Group	25	0.598 (0.552 to 0.644)	0.781 (0.730 to 0.831)
Garrison et al. (2017) Study 1	SMT: Self- vs Other-Read/Non-clinical LP Group	22	0.569 (0.515 to 0.622)	0.676 (0.615 to 0.737)
Garrison et al. (2017) Study 1	SMT: Self- vs Other-Read/Non-clinical HP Group	25	0.526 (0.476 to 0.575)	0.434 (0.386 to 0.482)
Garrison et al. (2017) Study 2	SMT: Imagine vs. Say/Non-clinical	120	0.374 (0.352 to 0.396)	0.707 (0.684 to 0.729)
Alderson-Day et al. (2019)	SMT: Say vs. Hear/Non-clinical	76	0.588 (0.555 to 0.620)	0.667 (0.647 to 0.688)
Moseley et al. (2021)	SMT: Imagine vs. Hear/Non-clinical	594	0.695 (0.684 to 0.706)	0.679 (0.666 to 0.691)

Notes: SDT = Signal detection task; SMT = Source monitoring task; HP = Hallucination-prone; LP = Low-hallucination-prone;  $r_s$  = Spearman-Brown corrected estimate.

## Analyses

Data (available at doi: 10.17605/osf.io/d3gnk) were analysed in R (R Core Team, 2018) using Sherman's (2015) *multicon* package. This package estimates permutation-based split-half reliabilities for the measures obtained from tasks. The permutation-based split-half reliability is an estimate of a measure's internal reliability, similar to Cronbach's alpha. While Cronbach's alpha can be calculated easily for measures obtained by questionnaires, it cannot often be calculated for tasks, because the order of trials is typically randomised and so varies across participants, and alpha can only be calculated when items are presented in a fixed order (Parsons et al., 2019). Split-half reliability refers to an estimate of reliability where the data from participants are divided in two (e.g. data from odd trials and data from even trials), and the correlation between these two halves is the calculated. This correlation is then used as an estimate of the measure's internal reliability. While these estimates tend to be unstable, this can be addressed by repeatedly, randomly dividing the data in two, calculating the correlation between the two halves, and then finding the average of these correlations. The Spearman-Brown correction is then applied to account for underestimation of reliability that may result from halving the number of trials (Parsons et al., 2019). The permutation-based split-half reliability is this corrected average correlation. Here, 5,000 random splits were performed to obtain each reliability estimate.

## Study 2 results and discussion

As shown in Table 4, a range of reliability estimates were generated for the datasets we analysed. The reliability estimates for measures obtained using signal detection tasks were higher than the estimates for measures obtained using source monitoring tasks, although the confidence intervals around these estimates overlapped in some instances. Parsons et al. (2019) argue that reliability estimates are best treated as continuous variables but note that some categorical conventions of what should be classed as moderate, good, and excellent reliability do exist. According to the thresholds proposed by Koo and Li (2016), 10 of the 18 reliability estimates reported in Table 1 would be classed as moderate, four would be classed as good, and one would be classed as excellent. That said, others (e.g. Barch et al., 2007) have proposed that reliability estimates around .90 are optimal for cognitive tasks, and only one of the reliability estimates we report reached that threshold.

Our findings that the measurements obtained by some of the tasks employed in hallucinations research have suboptimal levels of reliability are consistent with data from other sub-fields of psychological science. For example, analyses of the reliability of measurements obtained by tasks employed in anxiety research (e.g. Kappenman et al., 2014) and in inhibition research (e.g. Hedge et al., 2018) suggest that they may have unacceptably low levels of reliability. That being said, our analyses suggest that the measures obtained using signal detection tasks may be more reliable than those obtained using source monitoring tasks and this is consistent with other data that has suggested that measures obtained using paradigms similar to signal detection tasks have good levels of test-retest reliability (e.g. Huque et al., 2017).

These findings have important implications for hallucinations research. In contexts where researchers are investigating “true effects”, measuring variables with sub-optimal levels of reliability results in the attenuation of “true associations” and so statistical power is reduced (Parsons et al., 2019; Rouder et al., 2019). This is especially unfortunate in hallucinations research where researchers may find it difficult to recruit large samples. In contexts where researchers are not investigating “true effects” the reduction in statistical power caused by using measurements with suboptimal levels of reliability increases the likelihood of false-positive findings (Bakker et al., 2012). Thus, in two ways, it is possible that this issue of sub-optimal levels of reliability has contributed to some findings in our field being difficult to replicate (e.g. inconsistent associations between atypical source monitoring and HEs; Brookwell et al., 2013; Moseley et al., 2021).

This study had several limitations. First, the set of studies we re-analysed data from did not include a clinical sample. The measurements obtained by some neuropsychological tasks are more reliable in clinical than in non-clinical samples (Kopp et al., 2021), and it is possible that this may also be true for measures obtained using source monitoring and signal detection tasks. Future research should examine if this is the case. Second, our analyses focussed on external and internal misattributions (for the source monitoring tasks) and on hits and false alarms (for the signal detection tasks). These are the outcomes we often employ and are the outcomes for which we could calculate split-half reliability estimates for. However, many researchers use signal detection parameters as indices of performance on signal detection and source monitoring tasks, and we cannot comment on the reliability of those measurements. This is because signal detection parameters (e.g.  $d'$ ,  $\beta$ ) are task-level summary scores, whereas split-half reliability estimation relies on trial-level data. That said,  $d'$  and  $\beta$  are calculated using hit-rate and false alarm-rate, and so the reliability of those variables should be of interest to researchers who analyse source monitoring and/or signal detection task performance in terms of signal detection parameters. Nevertheless, it would be valuable for future studies that employ source monitoring and/or signal detection tasks to establish the test-retest reliabilities of signal detection parameters.

## General discussion

Across two studies, we have provided evidence that hallucinations researchers (including some of the authors of this article) engage in suboptimal measurement practices by, for example, using a wide range of different questionnaires to assess the same variable across different studies, modifying validated questionnaires, failing to adequately describe the measures that have been employed, and by using tasks that measure variables-of-interest with low levels of reliability. The primary consequences of these sub-optimal measurement practices are likely to be a reduction in the reproducibility and replicability of the findings hallucinations researchers report. For example, when it is unclear what version of the LSHS has been used in a study, it is harder for that study’s methods to be reproduced by other researchers aiming to replicate the initial study’s findings. Meanwhile, when measurements with low reliability are obtained, statistical power is reduced, and

this increases the likelihood of a study reporting false-negative and false-positive findings. Together, these issues reduce the robustness and credibility of the findings generated by hallucinations researchers.

An implication of these findings is that hallucinations researchers need to attend to issues of measurement more carefully. It should, perhaps, become the norm to report the reliability of the measurements obtained by tasks employed in hallucinations research. Several R packages now exist that allow researchers to estimate the internal reliability of measures obtained from tasks (e.g. *splithalf* and *multicon*; Parsons, 2019; Sherman, 2015), and the Excel-based tool *RELex* (Steinke & Kopp, 2020) can be used by researchers who are unfamiliar with using R. The availability of these packages/tools should facilitate the reporting of reliability estimates. In addition, the scales employed in a study should be described more comprehensively and accurately. Where word-limits prevent researchers from describing the scales employed in detail, the Open Science Framework ([osf.io](https://osf.io)) can be used to post study materials, so that scales which must be reported briefly in an article can be described/presented in full elsewhere. Finally, it seems undesirable to encourage all hallucinations researchers to employ the same measures of hallucinatory experiences (e.g. see Patalay & Fried, 2021, on possible unintended consequences of mandating the use of specific measures for assessing depression and anxiety). Instead, following the recommendation proposed by Flake and Fried (2020), researchers should be encouraged to be more explicit and precise when discussing and describing their construct(s)-of-interest. In terms of hallucinations research, this may involve, for example, avoiding the use of terms such as “hallucination-proneness” in future, as this term may have different meanings for different researchers (e.g. the frequency of experiencing a narrow set of hallucinatory percepts, such as hearing sounds that others do not, versus the frequency of experiencing a broader set of percepts/cognitions that may be related to hallucinations, such as intrusive thoughts or vivid daydreams). Instead, researchers should be encouraged to define their construct-of-interest more precisely and should explain how the scale they have selected measures that construct effectively. Again, where word-limits prevent this kind of reporting in the Method section of a journal article, the Open Science Framework can be used to publish supplementary methodological information.

More broadly, while engaging in better measurement practices would improve the rigour of hallucinations research, it should be considered as only part of a wider set of reforms that the field should engage in. Clinical psychological science has been slow to adopt the reforms that other sub-fields of psychology have engaged in (Tackett et al., 2019), such as the use of pre-registration, open data-sharing, and the use of open-source materials. Given the importance of generating a trustworthy evidence base from which interventions that help people with distressing hallucinations can be developed, it would be extremely valuable if hallucinations researchers engaged in this wider set of methodological reforms, as well as engaging in better measurement practices.

### **Disclosure statement**

No potential conflict of interest was reported by the author(s).



## Funding

This research was funded in whole, or in part, by the Wellcome Trust [Grant number WT108720]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## ORCID

David Smailes  <http://orcid.org/0000-0002-0455-070X>  
 Ben Alderson-Day  <http://orcid.org/0000-0003-0546-8043>  
 Cassie Hazell  <http://orcid.org/0000-0001-5868-9902>  
 Peter Moseley  <http://orcid.org/0000-0002-9284-2509>

## References

- Alderson-Day, B., Smailes, D., Moffatt, J., Mitrenga, K., Moseley, P., & Fernyhough, C. (2019). Intentional inhibition but not source memory is related to hallucination-proneness and intrusive thoughts in a university sample. *Cortex*, *113*, 267–278. <https://doi.org/10.1016/j.cortex.2018.12.020>
- Aynsworth, C., Nemat, N., Collerton, D., Smailes, D., & Dudley, R. (2017). Reality monitoring performance and the role of visual imagery in visual hallucinations. *Behaviour Research and Therapy*, *97*, 115–122. <https://doi.org/10.1016/j.brat.2017.07.012>
- Badcock, J. C., & Hugdahl, K. (2012). Cognitive mechanisms of auditory verbal hallucinations in psychotic and non-psychotic groups. *Neuroscience & Biobehavioral Reviews*, *36*(1), 431–438. <https://doi.org/10.1016/j.neubiorev.2011.07.010>
- Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. <https://doi.org/10.1177/1745691612459060>
- Barch, D. M., Carter, C. S., & Executive Committee, C. N. T. R. I. C. S. (2007). Measurement issues in the use of cognitive neuroscience tasks in drug development for impaired cognition in schizophrenia: A report of the second consensus building conference of the CNTRICS initiative. *Schizophrenia Bulletin*, *34*(4), 613–618. <https://doi.org/10.1093/schbul/sbn037>
- Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483* (7391), 531–533. <https://doi.org/10.1038/483531a>
- Bentall, R. P., & Fernyhough, C. (2008). Social predictors of psychotic experiences: Specificity and psychological mechanisms. *Schizophrenia Bulletin*, *34*(6), 1012–1020. <https://doi.org/10.1093/schbul/sbn103>
- Bentall, R. P., & Slade, P. D. (1985). Reliability of a scale measuring disposition towards hallucination: A brief report. *Personality and Individual Differences*, *6*(4), 527–529. [https://doi.org/10.1016/0191-8869\(85\)90151-5](https://doi.org/10.1016/0191-8869(85)90151-5)
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117*(2), 187–215. <https://doi.org/10.1037/0033-2909.117.2.187>
- Brookwell, M. L., Bentall, R. P., & Varese, F. (2013). Externalizing biases and hallucinations in source-monitoring, self-monitoring and signal detection studies: A meta-analytic review. *Psychological Medicine*, *43*(12), 2465–2475. <https://doi.org/10.1017/S0033291712002760>
- Core Team, R. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- de Beurs, E., & Fried, E. I. (2021). From mandating common measures to mandating common metrics: A plea to harmonize measurement results. *Psyarxiv*, <https://doi.org/10.31234/osf.io/m4qzb>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. <https://doi.org/10.1177/2515245920952393>

- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378. <https://doi.org/10.1177/1948550617693063>
- Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, 208, 191–197. <https://doi.org/10.1016/j.jad.2016.10.019>
- Garrison, J. R., Moseley, P., Alderson-Day, B., Smailes, D., Fernyhough, C., & Simons, J. S. (2017). Testing continuum models of psychosis: No reduction in source monitoring ability in healthy individuals prone to auditory hallucinations. *Cortex*, 91, 197–207. <https://doi.org/10.1016/j.cortex.2016.11.011>
- Haddock, G., McCarron, J., Tarrier, N., & Faragher, E. B. (1999). Scales to measure dimensions of hallucinations and delusions: The psychotic symptom rating scales (PSYRATS). *Psychological Medicine*, 29(4), 879–889. <https://doi.org/10.1017/S0033291799008661>
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50(3), 1166–1186. <https://doi.org/10.3758/s13428-017-0935-1>
- Huque, A. U., Heaney, A., Poliakoff, E., & Brown, R. J. (2017). Development and validation of a voice-hearing task for research on auditory verbal hallucinations and auditory misperception. *Psychosis*, 9(4), 338–346. <https://doi.org/10.1080/17522439.2017.1363275>
- Kappenman, E. S., Farrens, J. L., Luck, S. J., & Proudfit, G. H. (2014). Behavioral and ERP measures of attentional bias to threat in the dot-probe task: Poor reliability and lack of correlation with anxiety. *Frontiers in Psychology*, 5, 1368. <https://doi.org/10.3389/fpsyg.2014.01368>
- King, D. L., Chamberlain, S. R., Carragher, N., Billieux, J., Stein, D., Mueller, K., ... Delfabbro, P. H. (2020). Screening and assessment tools for gaming disorder: A comprehensive systematic review. *Clinical Psychology Review*, 77, 101831. <https://doi.org/10.1016/j.cpr.2020.101831>
- Kohut, T., Balzarini, R. N., Fisher, W. A., Grubbs, J. B., Campbell, L., & Prause, N. (2020). Surveying pornography use: A shaky science resting on poor measurement foundations. *Journal of Sex Research*, 57(6), 722–742. <https://doi.org/10.1080/00224499.2019.1695244>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kopp, B., Lange, F., & Steinke, A. (2021). The reliability of the wisconsin card sorting test in clinical practice. *Assessment*, 28(1), 248–263. <https://doi.org/10.1177/1073191119866257>
- Larøi, F. (2012). How do auditory verbal hallucinations in patients differ from those in non-patients? *Frontiers in Human Neuroscience*, 6, 25. <https://doi.org/10.3389/fnhum.2012.00025>
- Launay, G., & Slade, P. (1981). The measurement of hallucinatory predisposition in male and female prisoners. *Personality and Individual Differences*, 2(3), 221–234. [https://doi.org/10.1016/0191-8869\(81\)90027-1](https://doi.org/10.1016/0191-8869(81)90027-1)
- McCarthy-Jones, S., & Fernyhough, C. (2011). The varieties of inner speech: Links between quality of inner speech and psychopathological variables in a sample of young adults. *Consciousness and Cognition*, 20(4), 1586–1593. <https://doi.org/10.1016/j.concog.2011.08.005>
- Mew, E. J., Monsour, A., Saeed, L., Santos, L., Patel, S., Courtney, D. B., ... Butcher, N. J. (2020). Systematic scoping review identifies heterogeneity in outcomes measured in adolescent depression clinical trials. *Journal of Clinical Epidemiology*, 126, 71–79. <https://doi.org/10.1016/j.jclinepi.2020.06.013>
- Mitrenga, K. J., Alderson-Day, B., May, L., Moffatt, J., Moseley, P., & Fernyhough, C. (2019). Reading characters in voices: Ratings of personality characteristics from voices predict proneness to auditory verbal hallucinations. *PLOS One*, 14(8), e0221127. <https://doi.org/10.1371/journal.pone.0221127>
- Morrison, A. P. (2001). The interpretation of intrusions in psychosis: An integrative cognitive approach to hallucinations and delusions. *Behavioural and Cognitive Psychotherapy*, 29(3), 257–276. <https://doi.org/10.1017/S1352465801003010>



- Morrison, A. P., Wells, A., & Nothard, S. (2000). Cognitive factors in predisposition to auditory and visual hallucinations. *British Journal of Clinical Psychology*, 39(1), 67–78. <https://doi.org/10.1348/014466500163112>
- Morrison, A. P., Wells, A., & Nothard, S. (2002). Cognitive and emotional predictors of predisposition to hallucinations in non-patients. *British Journal of Clinical Psychology*, 41(3), 259–270. <https://doi.org/10.1348/014466502760379127>
- Moseley, P., Aleman, A., Allen, P., Bell, V., Bless, J., Bortolon, C., ... & Fernyhough, C. (2021). Correlates of hallucinatory experiences in the general population: An international multi-site replication study. *Psychological Science*, 32(7), 1024–1037. <https://doi.org/10.1177/0956797620985832>.
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Nosek, B. A., Aarts, A. A., Anderson, J. E., Kappes, H. B., & Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Parsons, S. (2019). *splithalf: Robust estimates of split half reliability* (R package Version 5) [Computer software]. <https://doi.org/10.6084/m9.figshare.5559175.v5>
- Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, 2(4), 378–395. <https://doi.org/10.1177/2515245919879695>
- Patalay, P., & Fried, E. I. (2021). Prescribing measures: Unintended negative consequences of mandating standardized mental health measurement. *Journal of Child Psychology and Psychiatry*, 62(8), 1032–1036. <https://doi.org/10.1111/jcpp.13333>
- Rouder, J., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. *Psyarxiv*. <https://doi.org/10.31234/osf.io/3cjr5>
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Sherman, R. A. (2015). *Multicon: Multivariate Constructs*. R package version 1.6 (R package version 1.6). <https://cran.r-project.org/package=multicon>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Smailes, D., Meins, E., & Fernyhough, C. (2015). Associations between intrusive thoughts, reality discrimination and hallucination-proneness in healthy young adults. *Cognitive Neuropsychiatry*, 20(1), 72–80. <https://doi.org/10.1080/13546805.2014.973487>
- Steinke, A., & Kopp, B. (2020). RELEX: An excel-based software tool for sampling split-half reliability coefficients. *Methods in Psychology*, 2, 100023. <https://doi.org/10.1016/j.metip.2020.100023>
- Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*, 15(1), 579–604. <https://doi.org/10.1146/annurev-clinpsy-050718-095710>
- Waechter, S., & Stolz, J. A. (2015). Trait anxiety, state anxiety, and attentional bias to threat: Assessing the psychometric properties of response time measures. *Cognitive Therapy and Research*, 39(4), 441–458. <https://doi.org/10.1007/s10608-015-9670-z>
- Woods, A., Jones, N., Alderson-Day, B., Callard, F., & Fernyhough, C. (2015). Experiences of hearing voices: Analysis of a novel phenomenological survey. *The Lancet. Psychiatry*, 2(4), 323–331. [https://doi.org/10.1016/S2215-0366\(15\)00006-1](https://doi.org/10.1016/S2215-0366(15)00006-1)