Social Psychological Bulletin

Investigating Lay Perceptions of Psychological Measures: A Registered Report

Joseph Mason[1], Madeleine Pownall[1], Amy Palmer[1], Flavio Azevedo[2]

[1] *School of Psychology, University of Leeds, Leeds, United Kingdom.* [2] *University of Cambridge, Cambridge, United Kingdom.*

Social Psychological Bulletin, 2023, Vol. 18, Article e9383, https://doi.org/10.32872/spb.9383

Study plan received: 2022-04-28 • **Study plan accepted (IPA):** 2022-07-30 • **Full paper received:** 2023-02-22 • **Full paper accepted:** 2023-04-08 • **Published (VoR):** 2023-11-17

Handling Editors: Simine Vazire, Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Australia; Brian Nosek, University of Virginia, Charlottesville, VA, USA

Corresponding Author: Madeleine Pownall, University of Leeds, 4 Lifton Place, Leeds, LS2 9JZ, UK. E-mail: M.V.Pownall@leeds.ac.uk

Related: This article is part of the SPB Special Topic "Is Psychology Self-Correcting? Reflections on the Credibility Revolution in Social and Personality Psychology", Guest Editors: Simine Vazire & Brian Nosek, Social Psychological Bulletin, 18, https://doi.org/10.32872/spb.v18

Supplementary Materials: Materials, Preregistration [see Index of Supplementary Materials]

Abstract

In recent years, the reliability and validity of psychology measurement practices has been called into question, as part of an ongoing reappraisal of the robustness, reproducibility, and transparency of psychological research. While useful progress has been made, to date, the majority of discussions surrounding psychology's measurement crisis have involved technical, quantitative investigations into the validity, reliability, and statistical robustness of psychological measures. This registered report offers a seldom-heard qualitative perspective on these ongoing debates, critically exploring members of the general public's (i.e., non-experts) lay perceptions of widely used measures in psychology. Using a combination of cognitive interviews and a think aloud study protocol, participants (n = 23) completed one of three popular psychology measures. Participants reflected on each of the measures, discussed the contents, and provided perceptions of what the measures are designed to test. Coding of the think aloud protocols showed that participants across the measures had issues in interpreting and responding to items. Thematic analysis of the cognitive interviews identified three dominant themes that each relate to lay perceptions of psychology measurements. These were: (1) participants' grappling with attempting to 'capture their multiple selves' in the questionnaires, (2) participants perceiving the questionnaire method as generally ‘missing nuance and richness’ and (3) exposing the ‘hidden labour of questionnaires’. These findings are discussed in the context of psychology’s measurement reform.

Keywords

measurement crisis, credibility crisis, cognitive interview, think aloud, qualitative research, metascience

Highlights

- In recent years, there have been concerns surrounding the robustness, reliability, and validity of psychological measurement practices.
- This study uses cognitive interviews and a think-aloud protocol with lay (i.e., non-expert) people to explore public perceptions of popular psychology measures.
- Generally, lay people experience issues with completing some of psychology’s most popular measures and find them to lack nuance.
- These results demonstrate the value of adopting qualitative methods in the reappraisal of psychology’s measurement crisis.

In recent years, there have been claims of various crises in psychology, including the credibility crisis, the replication crisis, and, more recently, the measurement crisis (see Lilienfeld & Strother, 2020). This is aligned with concerns surrounding the credibility, robustness, and transparency of research, which forms an integral part of the ‘open scholarship’ or open science movement (Azevedo et al., 2022; Vazire, 2018). Open scholarship refers broadly to the belief that research should be transparent, rigorous, reproducible, replicable, accumulative, and inclusive (Parsons et al., 2022). There have been many useful and accelerating practices to promote this view of research, including tools such as study pre-registration and Registered Reports (e.g., Chambers & Tzavella, 2022; Nosek et al., 2015), open data sharing (Houtkoop et al., 2018), open access publishing (Nosek & Bar-Anan, 2012), and wider adoption of replication studies (Open Science Collaboration, 2015). While these are all useful tools to improve the robustness of psychological research, a consideration of psychological measurement should also be included within these conversations.

A landmark paper ‘Measurement Schmeasurement’ by Flake and Fried (2020) articulated the questionable measurement practices (QMPs) in psychology and critically called into question the reliability and validity of popular psychology measures. In this paper, Flake and Fried (2020) highlighted how measurement validity, reliability, and robustness is a fundamental part of psychological research, which has remained curiously absent from conversations surrounding the improvement of psychological science and psychology’s credibility crisis. As the authors summarise, “a foundational part of science is defining and measuring *what is being studied*” (p. 456, italics added). Therefore, if QMPs plague the literature, the progression of science will continue to be built upon unstable foundations and scientists will not be able to measure the psychological constructs of interest accurately, reliably, or meaningfully. Therefore, it is important that researchers continue to use robust and diverse methodologies to thoroughly appraise psychological measures. The present study contributes to this ongoing discussion, offering a seldom-heard qualitative perspective on the (re)consideration of psychology's measurement crisis.

(Re)appraisals of Psychological Measurement

There has been some useful progress in psychological measurement using *quantitative*, technical, empirical methodologies to explore the statistical robustness of psychological measures and scales. For example, researchers have investigated the statistical utility of self-report measures (e.g., Parry et al., 2021), explored construct validity in measurement approaches (Flake et al., 2017; Kane et al., 2021), highlighted the poor foundations of certain measurement constructs (e.g., Kohut et al., 2020; Satchell et al., 2021; van Hauwaert et al., 2020) and questioned the use of measurements across academic fields (e.g., health education; Barry et al., 2014; and psychopathology; Rodebaugh et al., 2016). Similarly, researchers have also highlighted the reporting biases that exist within self-report scales, which undermine their utility (e.g., Anvari et al., 2022).

More notably, there has also been ongoing discussion surrounding the nuances of statistical *validity* and what constitutes a valid, robust measurement (e.g., see Azevedo & Bolesta, 2022; Flake et al., 2017; Hussey & Hughes, 2020; Wetzel & Roberts, 2020). Such considerations are useful and necessary starting points, without which other efforts to improve psychological research could not progress. However, scholars have also called for researchers to "invest in a deeper examination" of psychological measurement (Wetzel & Roberts, 2020, p. 505). For researchers to capture the more nuanced and in-depth aspects of a measurement's validity, reliability, or robustness, there is value in using alternative and under-used methodologies to conceptualise, measure, and discuss rigour in psychological measurement. We argue that qualitative approaches may offer a more nuanced, richer account of how the contents, wording, and structure of psychological measures are interpreted and understood, ensuring our measures are credible and robust. This may provide richer insights into the measures we employ in psychology, thus contributing to the reappraisal of validity within psychological measurement.

In the present study, we propose an alternative perspective on conversations surrounding psychological measurement, by offering a qualitative appraisal based on lay perceptions of measurement. It is important to note that this somewhat critical perspective is inherently situated within the ongoing discourses of 'methodological legitimacy' in psychology and beyond. Lewis (2021) describes methodology legitimacy as the consideration of which approach *counts* as a valid method for generating knowledge about thoughts, feelings, and behaviours. Such discussions can pave the way for more epistemologically diverse approaches to the measurement crisis, including a consideration of *how*, *what*, and *why* we measure constructs across psychological research (Steger et al., 2022).

Indeed, to date, many discussions surrounding *how* psychology's measurement crisis should be tackled have involved technical, quantitative investigations into the validity, reliability, and statistical robustness of psychological measures, often neglecting the utility of qualitative methods. While qualitative methods are often useful in assessing the robustness of various self-report measures (e.g., Belzer et al., 2013; Murphy et al., 2018), papers such as '*Best Practices for Developing and Validating Scales for Health, Social, and Behavioural Research*' (Boateng et al., 2018) neglect the use of qualitative methods in assessing the validity of self-report measures beyond initial item generation. We thus argue for a greater emphasis on *qualitative* perspectives in reconsiderations of self-report measurements, aligning these scales with the population they are ultimately designed *for*, i.e., members of the public who typically do not have insights into the typically homogenous expert assumptions of psychological constructs. For example, evidence from conducting cognitive interviews with the Freiburg Mindfulness Inventory, showed that authors and participants held distinctly different accounts of what certain items were measuring (Belzer et al., 2013). Put simply, if non-experts do not understand, relate to, or accurately interpret the contents of psychological measures in a similar manner as they are believed to by experts (i.e., researchers), then psychology risks that any theories and models derived from these measurements may not accurately reflect the human mind. In this paper, we suggest there is value in informing the structure, contents, and wording of scales on lay (or "non-expert") perspectives in conversations surrounding the validity, reliability, and usability of psychological measures.

Our approach derives from the social constructionist epistemological position, which views psychological constructs as concepts that are actively constructed to have specific meanings through individual, subjective lenses (Goodson, 1990), allowing us to offer an alternative epistemological consideration on the 'measurement crisis'. To date, considerations of psychological measurement have stemmed almost exclusively from an idealist-positivist approach, favouring 'objective' empirical investigations (Azevedo, 2023), however, these tend to neglect the impact an individual's subjective experience of a construct can have on the measurement of it. In-turn, these empirical investigations would assume that there is one 'true' experience humans can have of a given construct, with each scale item in a self-report measure being an accurate reflection of this experience. However, there is no one unified human experience and, thus, all experiences are unique. Therefore, in the present study, we adopted a social-constructionist epistemology, using qualitative methods to gain insights into how questionnaire items may be interpreted differently by members of the public (i.e., non-experts). As we conceptualise it, there is a notable gap in the current understanding of psychological measurement and an *overreliance* on quantitative appraisals of measures as a proxy for validity and reliability.

Think Aloud and Cognitive Interviewing

The present study will use think aloud and cognitive interviews to explore lay perceptions of psychological measurements. Foddy (1993) previously offered a suggestion of testing self-report measures using qualitative methods by, first, having participants think aloud while completing the questionnaire, followed by the interviewer probing to ensure that the questions are understood. Following on from this, Wolcott and Lobczowski (2021) suggested that while think aloud protocols are often implemented in problem solving tasks (e.g., Hoppmann, 2009), they may be equally used in assessing the validity of questionnaire items when combined with a retrospective cognitive interview.

Think aloud protocols (TAPs) involve participants continuously verbalising their thoughts while completing each item in a questionnaire in such a manner that they do not disrupt their own thought process (Ericsson & Simon, 1998; Gilhooly & Green, 1996). Think aloud has previously been used to uncover a wide array of issues with self-report questionnaires; from participants misunderstanding terminology used in some self-report health questionnaires (e.g., Willis et al., 1991), outdated and complicated question wording (e.g., Schmidt, Brandt, et al., 2022), specific difficulties with various sub-constructs (e.g., van Oort et al., 2011), and how question wording can bias choices towards middle values (e.g., Darker & French, 2009). Research demonstrates that participants tend to experience four major processes when answering items on a questionnaire: question comprehension, retrieval of information relevant to the question, decision making, and generation of their response (Tourangeau et al., 2000). Therefore, TAPs can be a useful way to tap into these processes during questionnaire completion.

Cognitive interviews, on the other hand, involve probing participants with questions to gain an insight into their comprehension of a given task and are thus more able to gain an understanding of participants' thought processes during the completion of a scale by being able to directly probe participants on specific scale items (Wolcott & Lobczowski, 2021). They have also been used to find various issues with self-report measures, including participants making use of completely different information than that which the scale designers assumed they would (Belzer et al., 2013), issues with question interpretation when changing question wording from adjective to traditional (Murphy et al., 2018), and problems with double barrelled and open items (Hilton, 2017).

Study Aims

The present study offers a qualitative perspective on ongoing debates surrounding psychology's measurement crisis; exploring how members of the public (i.e., non-experts) perceive popular and widely used measures in psychology. Qualitative perspectives on metascience discussions are valuable but are yet to focus on reflexively appraising (i.e., re-appraising) popular psychological measurements, largely due to epistemological considerations and the dominance of quantitative, experimental, empirical investigations across metascience (Branney et al., 2023) and the main use of qualitative methods that are widely considered isolated from the *development* of measures (e.g., Boateng et al., 2018). This study thus aimed to address concerns with the reliability and validity of popular measures across psychology and thus constitutes a metascience project. In doing so, this research aimed to ‘give a voice’ to public psychology (see Lewis, 2021).

Research Questions

Our core research question was: “How do non-experts understand the contents, wording, and structure of popular psychology measures?”, which was approached using the think aloud method, followed by cognitive interviewing, with both of these methods being used in previous research (e.g., Belzer et al., 2013; Darker & French, 2009; and Schmidt, Brandt, et al., 2022). The present study is important as it is, to the best of our knowledge, the first to make use of Wolcott and Lobczowski’s (2021) notion of the utility of TAPs *and* CIs for reappraising psychological measurement. While we had *a priori* research questions, we approached the study in an exploratory way with no predetermined ideas about what might be generated from participants.

Method

Participants

Participants were “non-experts” (i.e., lay-people) who had not encountered the three measures before and who had no formal research training. Our sample aimed to be as diverse as possible, with recruitment intentionally conducted away from our university campus to capture as wide an age range and socioeconomic background as possible. While generalisability is not a pressing concern of qualitative methods (e.g., see Smith, 2018), this was done to ensure that the present study’s sample was not taken up by university undergraduates, as we aimed to ‘give a voice’ to *public* psychology and not confine our research to the realm of academia (as per Lewis, 2021). We initially interviewed 25 people. However, one transcript was corrupted when we exported it from Teams and one participant’s internet signal prevented a meaningful conversation. Therefore, after removing one corrupted and one unintelligible transcript, our final sample comprised a total of 23 participants, whose ages ranged from 20 to 68 years old (M^{age} = 45.29, *SD* = 12.03). Participants were recruited through the following avenues: social media (namely Twitter and Facebook), Prolific Academic, and survey share platforms. Participants came from a wide and diverse range of occupations, including civil servants, magistrates, teachers, stay-at-home parents, and administrative assistants (see Supplementary Materials in the Open Science Framework [OSF] for an exhaustive account). Participants were excluded if they were unable to ‘think aloud’, which was self-defined. Study materials, including interview protocol and think aloud script, can be accessed openly via the Supplementary Materials. Ethical approval was obtained from the University of Leeds School of Psychology Ethics Committee on 27th April 2022, Reference: PSYC-523. Data collection took place between 23rd August and 2nd September 2022.

Sample Size Justification

Nielsen (1994) suggests that five participants are necessary for think aloud interviews, while Willis (2005) recommends between 7–10 participants when conducting cognitive interviews. We aimed to recruit 10 participants per measure, as per Nielsen's (1994) recommendations and in order to recruit as large and insightful a sample as possible, (total target n = 30). However, due to resources, time, and feasibility concerns, we also considered 8 participants per measure to be the minimum sample required (n = 24). Our final sample was 23 participants, after removing two corrupted and uninterpretable transcripts; given the richness of the data, we considered this a satisfactory sample.

Materials

The present study used the following measurements: the Rosenberg Self-Esteem Scale (RSE; Rosenberg, 1965), the Need for Cognition Scale (NCS; Cacioppo et al., 1984), and the Short-Form Big Five Inventory (BFI-10; Rammstedt & John, 2007). These measures were chosen because they are among the most widely cited scales in psychology, are all direct measures, are used widely in research settings, are thought to have good internal consistency (Hussey & Hughes, 2020; although see Wetzel & Roberts, 2020) and, importantly, are designed to assess a general (rather than a clinical) population. They were also selected because they each measure distinct psychological constructs: self-esteem, need for cognition, and personality. They are thus suitably diverse in their aims and scope, but each follow a relatively similar responding pattern (i.e., items with Likert-style scales of agreement). They are also short enough as to not become tedious for participants while thinking aloud, as well as relatively easy to discuss in our method of immediate retrospective probing (*item Ns* = 10, 18, 10 for each scale, respectively).

Each of these scales are also typically used in research on psychological constructs, rather than in diagnostic or clinical settings. Importantly, each of these scales have been involved in conversations surrounding measurement robustness. The RSE, for example, has been found to possess some reliability issues surrounding its negatively worded items (Wongpakaran & Wongpakaran, 2012). Additionally, it has been suggested that the BFI-10 has weak reliability (Brown & Sotardi, 2019), a high risk of producing both Type-1 and Type-2 errors compared to its longer versions (Credé et al., 2012), possesses different levels of validity when administered through different mediums (Lang et al., 2011), and may not be a valid method to assess personality in isolation (Balgiu, 2018). However, due to the present study making use of qualitative methods to explore self-report measures, we explicitly decided to include a highly statistically robust measure in our measures.

The NCS has shown robust test-retest reliability (Sadowski & Gulgoz, 1992), both convergent *and* discriminant validity (Lilienfeld & Strother, 2020; Osberg, 1987; which has been suggested as necessary in determining construct validity), as well as being invariant across age groups (Soubelet & Salthouse, 2017). We will now detail each scale in turn.

The Rosenberg Self-Esteem Scale

The first measure to be selected was the Rosenberg Self-Esteem Scale (RSE; Rosenberg, 1965). This 10-item scale is designed to measure one's overall sense of self-esteem. The original 1965 paper has been cited 1,236 times on Google Scholar and is a widely used measure for self-esteem across the literature. Participants answer ten items on a 1 (*Strongly disagree*) to 4 (*Strongly agree*) Likert scale, including some that are reverse scored (e.g., "At times I think I am no good at all"). Rosenberg (1979) reported that the RSE has strong test-retest reliability (with correlations of .85 and .88), and concurrent, predictive, and construct validity.

The Need for Cognition Scale

The second item used was the Need for Cognition Scale (NCS; Cacioppo et al., 1984). This scale consists of 18 items (including nine reverse-scored) which measure an individual's tendency to engage in and enjoy tasks that require cognitive effort (Cacioppo & Petty, 1982). This includes items such as "I would prefer complex to simple problems" and "I prefer my life to be filled with puzzles that I must solve" which are measured on a 1 (*Extremely uncharacteristic*) to 5 (E*xtremely characteristic*) Likert scale. This scale has been cited 4,176 times on Google Scholar and is a widely used measure of need for cognition, with good internal consistency (e.g., see Schmidt, Buchanan, et al., 2022; Tobin & Guadagno, 2022). Need for cognition is also associated with, but crucially distinct from, the Big Five dimensions of personality (Tuten & Bosnjak, 2001).

The Short-Form Big Five Inventory

The final measure presented to participants was the Short-Form Big Five Inventory (BF1-10; Rammstedt & John, 2007). At the time of writing, this 10-item version of the Big Five has been cited 3,910 times according to Google Scholar (as of April 2022). The BFI-10 scales have good levels of reliability, test-retest stability, structural validity, and convergent validity (see Rammstedt & John, 2007) and is designed to measure five dimensions of personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Participants are asked to rate the extent to which "I see myself as someone who…" on a 1 *(Disagree strongly*) to 5 (*Agree strongly*) Likert-scale, including items such as "tends to be lazy" (conscientiousness, reverse-scored) and "gets nervous easily" (neuroticism). There are two items for each of the Big Five personality dimensions and half are reverse scored. We elected to use the short form of the Big Five inventory so as not to overwhelm participants with overly long or complex scales to discuss.

Procedure

Due to participants likely becoming fatigued if they were to complete all three of the scales over one extended think aloud interview, participants were assigned to one of three groups, each completing either the BFI-10 (n = 10), the RSE (n = 7), or the NCS (n = 6). Note that of the two transcripts that were removed due to corruption or unintelligibility (see *Participants*), one participant was assigned to the RSE and one to the NCS, which led to an uneven distribution of participants across groups. However, the richness of the data was comparable across the measures, so this was not considered to be problematic. Once participants had signed up and read the information sheet, online 1-to-1 meetings were set up with participants over MS-Teams. The interviews were conducted by the second author, who is an undergraduate psychology student at the School of Psychology, University of Leeds, who had received training in qualitative research and cognitive interviewing and interviews were supervised by the first author. Participants were first welcomed and reminded of the study aims, structure, and their right to withdraw. The interviewer's script on instructing participants on how to think aloud as well as for cognitive interviewing was based on the example scripts provided by Wolcott and Lobczowski (2021), with participants each completing one scale. Each interview lasted an average of 42 minutes (Range 23–64 minutes) and were recorded using MS-Teams's built-in software.

Think Aloud Interviews

To date, most studies have made use of *either* cognitive interviews or think aloud to assess scale validity; however, the use of *both* methods has been suggested as a viable (if not recommended) option to increase the richness of the data (Foddy, 1993; Wolcott & Lobczowski, 2021). The present study used a think aloud protocol during the completion of each scale to uncover participants' cognition, followed by a retrospective cognitive interview (CI) after the completion of the scale to probe participants' comprehension of specific scale items.

In line with best practice recommendations (e.g., see Gilhooly & Green, 1996; Wolcott & Lobczowski, 2021), a 'warm up' task was completed before each questionnaire was administered. The warmup task consisted of a single question, designed to have a similar format and wording to the questionnaire they were presented before, however, its content was not related to any of the present study's measures (see Results in the Supplementary Materials). The purpose of each warmup task was to allow the researcher to check whether the participant was able to think aloud and to clarify requirements of the participant. If participants struggled with the first example question, they had an example answer read out to them, and they were then presented with a second warmup question (see Results in the Supplementary Materials). While participants completed each questionnaire they were not interrupted, unless they remained silent for more than 10 seconds, in which case they were instructed to 'please keep talking'. While the interview was conducted, the interviewer's camera was turned off to not distract the participant with potential non-verbal cues.

Cognitive Interviewing

After the thinking aloud phase, participants were asked to complete a semi-structured interview. This involved retrospective probing using questions and prompts (see Table 1), which aimed to explore different types of validity and reliability, with a focus on face validity (i.e., that the scale appears to measure what it intends to measure). For example, participants were asked about how they interpreted scale items, or the scale as a whole, whether the way in which the scale was written may have impacted how they answered, and if they would make changes to the structure (see Table 1). Many of these prompts were adapted from Wolcott and Lobczowski's (2021) cognitive interviewing script. At the end of each interview, participants were told what each scale was titled and what it intended to measure and were asked to reflect on whether their responses were a true reflection of each scale's intended measure.

Table 1

Cognitive Interview Questions and Example Prompts

Type of validity/ reliability	Semi-structured questions	Example prompts	Elaborative prompts
Face validity	What knowledge or ability do you think this scale is testing and why?	Could it be measuring something else?	Could you elaborate on that, please?
	How would you go about completing this scale?	How could you complete this scale accurately?	Which items, specifically?
	What changes you would make to how this is structured?	Do you have any wider thoughts on the structure?	How could it be improved (if at all)?
	What, if any, experiences did this question make you think of while answering it?	What memories did this question make you think of when measuring it?	Could you elaborate on that, please?
	If you were completing this question in a different setting, would that perhaps impact your response?	Was there wording about this question that has influenced your response?	Could you elaborate on that, please?

Type of validity/ reliability	Semi-structured questions	Example prompts	Elaborative prompts
Construct Validity	Can you think of any other concepts that are relevant to what this scale is measuring?	What might these be?	Could you elaborate on that, please?
	If you had to retitle this scale, what would you change it to?	How, if at all, would having this as the scale's title have impacted your responses?	Could you elaborate on that, please?
Content Validity	How could this scale expand upon the idea of [*self-esteem, need for cognition, personality*]?	Are there other questions that should be asked in this scale?	Could you elaborate on that, please?
Internal consistency	Do you think any of the items are measuring something different from the rest?	Do they all seem to be measuring the same concept?	Can you elaborate on that, please?

Note. Adapted from Wolcott and Lobczowski (2021).

A Note on Researcher Positionality

An explicit acknowledgment of research positionality is an important facet of all research (Jamieson et al., 2023), particularly when dealing with potentially contentious or epistemologically complex issues. Therefore, we actively centred our perspectives as researchers in the research process. For context, the supervisor of this project (M.P) is an early-career Lecturer, social psychologist, and psychology educator. She identifies as a researcher who is gently (and increasingly) critical of the open science movement and has previously called into question the accessibility and inclusivity of open science spaces (Pownall et al., 2021). The lead member of the research team (J. M.) is an undergraduate psychology student whose concern with potential issues surrounding how current psychological research is conducted has led to a keen, albeit relatively new, interest in the open science movement. The third author (A. P.) is a PhD researcher with expertise in cognitive interviewing and think aloud protocols. The final author (F. A.) is a political psychologist with an interest in broadening epistemological considerations within conversations of psychology's measurement concerns and is an advocate for the integration of open scholarship principles into higher education as well as the advancement of research transparency, reproducibility, rigour, and ethics through pedagogical reform. We aimed to centre our various lenses throughout our exploratory analyses and thus our qualitative findings must be contextualised within what we 'bring' to the research itself.

Methods of Analysis

Data Preparation

After the interviews were complete, each of their respective recordings were then transcribed verbatim by one member of the research team (J. M.) and checked thoroughly by one other member of the team (M. P.) We transcribed exhaustively; that is, we included all details of the interviews, irrespective of participants stop-starting or their vocal disfluencies. This style of transcription was chosen to ensure that analysis captured the data in as accurate a way as possible. Transcripts were written and checked in Microsoft Word and we did not use any software packages (e.g., NVivo) to analyse the data.

Analytical Approach

The think aloud aspect of the protocol was analysed using an established coding framework (as per van Oort et al., 2011), and the cognitive interviews were analysed using a richer, qualitative thematic analysis approach (Braun & Clarke, 2006), due to the open-ended nature of our research question.

Think Aloud — As per other research using TAPs, we adopted a coding framework process to analyse the TAP aspect of the present study. This process of codifying data into categories has been used by other measurement analyses in the literature (e.g., van Oort et al., 2011). Participants’ responses to each questionnaire item were coded on a 1 (“No issues with an item”) to 5 (“Misinterpretation of an item”) category scale, using the framework in Table 2. This was achieved by copying the TAP aspect of the interviews into an Excel sheet, whereby each row represented a participant’s full response. Codes were represented in columns on the spreadsheet (1–5) and each response was assigned a code. All responses were coded by the first author. To establish inter-rater reliability, 50% of responses were coded blind by the fourth author. Initial inter-rater agreement was 61.11% (k = 0.284; McHugh, 2012), indicating a “fair” level of agreement in accordance with Landis and Koch (1977). Note that although this kappa statistic is low, it does not indicate overall poor inter-rater agreement and is likely due to the homogeneity of the data; for example, there was a high number of ‘no issues with an item’ and participants’ misinterpreting/comprehending an item’ in the initial coding, resulting in the ‘base rate problem’ (Morris et al., 2008). Coding disagreements were resolved through active and regular discussion between J. M. and A. P. and revised accordingly.

Table 2

Coding Categories for Analysis of Think-Aloud Interviews

Response Category	Description
1. No issues with an item	No significant problems with the questions were identified
2. Insufficient thinking aloud while completing an item	Participants were unable to report sufficient information to be coded
3. Issues with reading an item	Participants would make an effort to re-read the question or stumbled (i.e., stammered or stuttered due to a misreading) while attempting to read it aloud. While not necessarily reflective of misunderstanding, multiple participants stammering or stuttering on an individual item may indicate some issues
4. Issues in understanding an item	Participants may have questioned the content of the item, stated they would need more information before answering, or appeared confused by the question
5. Misinterpretation/comprehension of an item	Answered a different question from the one that was asked, or gave reasoning which was inconsistent with the question

Cognitive Interviews — To add richness to the findings, we then conducted a thematic analysis to analyse the data for the immediate retrospective cognitive interviews. This type of analysis was considered appropriate, given the quantity and depth of our data and our interest in broad perceptions that reflect participants' nuanced and complex viewpoints. We followed Braun and Clarke's (2006, 2019) six steps for thematic analysis to interpret the qualitative data. These are: (1) data familiarisation, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, (5) defining and naming themes, (6) producing the final research report. As part of the first step, we first familiarised ourselves with the data by reading and rereading each transcript, making annotated notes down the margins of the Microsoft Word transcripts to highlight interesting aspects about the data. Given that our research question was relatively open-ended, we adopted a data-driven, iterative process and attempted to capture the contents of the interviews in an inductive way, rather than arriving at the data with a predetermined set of deductive ideas. However, to relate the analysis of the interviews to the ongoing discussions surrounding psychological measurement, we were attentive to formal types of validity and reliability discussed in the literature (e.g., construct validity, internal consistency, convergent validity etc.) to provide a broad, non-exhaustive theoretical framework to keep in mind (e.g., Flake et al., 2022; Steger et al., 2022). These constructs were not 'applied' to the data, but rather served as a useful vocabulary for us to draw upon while engaging in our thematic analysis.

We aimed to remain as attentive as possible to participants' perceptions of the scales. Our analysis was guided by our core research question: "How do non-experts understand the contents, wording, and structure of popular psychology measures?" After we had familiarised ourselves with the transcripts, we then began generating initial codes before searching for broader themes that connected codes within the data. We generated initial codes individually and engaged in reflexive discussion to translate these codes into broader themes, using regular in-person meetings between the research team to facilitate discussion. Finally, we then worked openly and collaboratively to review and finalise themes, keeping our research question in mind throughout this process. This was achieved using both (a) research team meetings, working through the themes and codes with transcripts, (b) asynchronous commenting on the transcripts and responding to each other's comments.

Results

Think Aloud

Table 3 shows the breakdown of the five think aloud codes assigned to each of the scales in total. Note that in the Stage 1 manuscript, we expected that we would report percentages by calculating % of overall segments; the total number of segments is the total number of items in that scale x *N* of participants that completed that scale (e.g., for the Rosenberg Self-Esteem Scale; 10**N*). However, given that there were instances where participants satisfied *two* think aloud codes in one response (e.g., they had issues reading an item but then misinterpreted it), we now report percentages of the total number of codes applied to each scale. No item was completely unproblematic and some were more problematic than others. The Need for Cognition scale had the highest percentage of issues (35%), followed by the Big Five Inventory (29%), and the Rosenberg Self-Esteem Scale (24%). The most common problem coded between the three scales was 4) issues understanding an item (47/87 problems) with a similar number of items being insufficiently thought aloud (41/287 total codes). The least frequent issue was 'issues reading an item' (19/87 problems). The Supplementary Information (Results in the Supplementary Materials) provides the total codes and percentages for each of the individual items within each scale as well as all verbalizations to exemplify our coding.

Table 3

Breakdown of Total Think Aloud Codes per Scale in Total (N and %)

Measure	**No issues with an item (1)**	**Insufficient thinking aloud while completing an item (2)**	**Issues with reading an item (3)**	**Issues in understanding an item (4)**	**Misinterpretation or comprehension of an item (5)**	**Total *N* of problem-related codes (3, 4, or 5) applied to each scale**	**Total *N* think aloud answers that were coded**	**Total *N* of codes applied to each scale**
RSE (*n* items = 10; *n* participants = 7)	46	9	4	6	7	17	70	72
NCS (*n* items = 18; *n* participants = 6).	56	17	9	26	5	40	108	113
BFI-10 (*n* items = 10; *n* participants = 10)	58	15	6	15	9	30	100	103
Total	160	39	19	47	21	87	278	288

NCS

The most problematic item between all three scales was item 16 of the NCS, '*I feel relief rather than satisfaction after completing a task that required a lot of mental effort*', with 6/6 (100%) of participants being coded as having at least one problem with the item. Item 15, '*The notion of thinking abstractly appeals to me'* was also the second most problematic item between all three scales; displayed by 4/6 (66%) of participants having difficulties answering the item. The least problematic items were item 9 '*I like tasks that require little thought once I've learnt them'* and item 18 '*I usually end up deliberating about issues even when they do not affect me personally'* with only one participant struggling to think aloud on each of these questions with no codable problems identified. Remaining items were problematic in 2–3/6 (33–50%) of participants.

RSE

The most problematic item in the RSE scale was item 2 '*At times I think I am no good at all*', with 3/8 (38%) of participants being coded as having at least one problem with the item. The least problematic item was item 5 '*I feel I do not have much to be proud of',* with only one participant struggling to think aloud on the question with no codable problems identified. Remaining questions were problematic in 1–2/8 (13–25%) of participants.

BFI-10

The most problematic item in the BFI-10 was item 1 'I see myself as someone who is reserved', with 5/10 (50%) of participants being coded as having at least one problem with the item. The BFI-10 also boasted the least problematic between all three scales with item 6, '*I see myself as someone who is outgoing, sociable',* with only 1/10 (10%) of participants struggling to think aloud while answering it with all others having no issues. Remaining questions were problematic in 1–4/10 (10–40%) of participants.

Thematic Analysis

Our thematic analysis of the interview transcripts across the measures generated three dominant themes. These were: (1) participants' grappling with attempting to capture 'multiple selves' and temporality in the questionnaires, (2) participants perceiving the questionnaire method as generally 'missing nuance and richness' and (3) exposing the 'hidden labour of questionnaires'. Participant quotes are used, indicated by italics, to illustrate examples.

Theme 1. Attempting to Capture Multiple Selves

Across the interviews, participants spoke in-depth about their own lived experiences, to make sense of and relate to the items within the questionnaires. For example, participants spoke about their families, careers, living circumstances, and other life events when asked to reflect upon the questionnaires. This demonstrated the thought that goes into answering each of the questionnaire items. Across the dataset, participants' responses were rarely linear, and often drew upon vast and complex personal histories and lived experiences to understand and answer each item on the questionnaires. Some participants did this more than others. For example, one participant changed their mind when answering questions in the Big Five Inventory, and later explained this thought process:

> *"the 'finding fault in others' [item in the questionnaire] made me think about my family this week. I've got a couple of family members who have been extremely critical, extremely critical of this week and I found fault in them and that that you know, again, just being honest with you made me think about how I thought and how I've sort of acted this week with both in my own mind and then verbalizing that to my wife. Uh, so yeah, it made me—made me think about that."*

Related to the idea of participants connecting lived experience with the questionnaires was an underlying notion of participants occupying 'multiple selves' and drawing upon these different selves when answering questionnaires. For example, participants often referred to the existence of a "work self" and positioned this as being different from their 'other selves' (e.g., "*I sort of felt myself going into work mode and, and some of it I felt myself going into personal history mode*"). Beyond different self-aspects, participants also referred to the differences across anticipated feedback, which is another component of McConnell's model. In this sense, participants struggled to understand some of the questionnaire items because the source of the feedback was not clear, for example:

> "*It depends who's making that judgement. Is it in work? Is it in social? Is it my boss? And is it my family? Is it whatever the circumstances require, so I'm not really sure.*"

Similar to the issue of multiple selves, participants also grappled with *temporality* in their responses. Some participants spoke about their questionnaire responses as unchanging and fixed (e.g., "*at a gut level, everybody knows where they are")* and did not perceive a potential change in responses due to delivery modality of the questionnaire, time, or context. However, other participants discussed how differing contexts and time affects their perceptions of the questionnaire items. For example, several participants were keen to caveat their responses by acknowledging how their responses would change depending on their day (e.g., "*well, it depends on that scenario and on my mood*"). One participant explained this in length, when completing the Rosenberg Self-Esteem Scale, and took particular issue with the Likert-scale anchors:

> "*Some days you feel better than others. I know it says 'at times' and 'on the whole', but I think it will be better if the questionnaire took like*

> *a "how do you feel today" approach…so that you could maybe gauge that if someone's having like a really rubbish day that that would impact on like an on [sic] the whole questions, if that makes sense."*

Throughout the data, participants discussed how time, place, or context will alter their responses and were keen to convey this to the interviewer. Participants drew upon a range of examples, including those related to context-specific examples:

> *"There might be occasions when I'm in the middle of a DIY task and I can't get it to do it. Then I'd say I feel useless then because I can't accomplish a specific task. But on the whole I rarely feel useless."*

As well as more temporal qualities that were perceived to influence the integrity of their questionnaire responses:

> *"And also the times in which this is being done. I'm a morning person. This is perfect for me but by the late afternoon, I hate everything. I hate my job. I hate interacting."*

Beyond temporality between days, shifting moods, and experiences, other participants in this theme also reflected upon how their life stage and age can impact their responses to questionnaires, such that their responses have changed over time longitudinally. For example, participants often reflected on how different people at different life stages would respond to questions in different ways. For example, participants reflected upon how younger people may relate to the questions differently: "*if you are sort of younger then it may be you, you may be more self-critical of yourself because you're still sort of growing*".

In response to these concerns about missing the multidimensionality of participants' full character, some participants then provided the interviewers with advice on how to improve the measures. For example, one participant said:

> *"It might be good to pre-screen before somebody goes into this, to sort of like know the emotions of how they're feeling that day. So how are you feeling today? happy? Sad?... Have you had sleep? Have you experienced anxiety the past couple of days, blah blah blah. Because that could also impact the data for what they're actually trying to gather with the real study."*

Overall, this theme highlights how participants often find it difficult to limit their full lived experiences to one response on a questionnaire item and have an awareness of this when completing studies that use these tools.

Theme 2. Questionnaires Perceived as Missing Richness and Nuance

Beyond participants' grappling with attempting to capture multiple aspects of themselves in the interview, participants also discussed their perceptions of the limits of the questionnaire method, more generally. Often this centred around the limits of scales and items, but this also included a broader perception that questionnaire measures do not capture the richness of participants' fully realised responses. For some participants, questionnaires were perceived to be limited as a methodology because the cross-sectional nature meant that nuance in participants' responses was often missing. For some participants, this nuance was an important part of their 'story' and, when completing the questionnaires, they were keen to 'fill in the gaps' for the interviewer, to provide the full context of these more complete 'stories'. For example, when completing the need for cognition scale, one participant told a story about how their broader political and personal experiences affect their responses:

> *"What's immediately springing to mind is my husband's in the coast guards, and he's had lots and lots of emergency call outs and particularly over the summer where we live by the sea. Very sadly, recently he had to deal with the death of somebody and I spent a long time thinking both about the impact on him personally, but he was fine. I [also] think about people in Ukraine, I think about...the state of the economy and people who in the future won't be able to afford the increase in fuel bills."*

More broadly, some participants spoke of perceived discrepancies between the perception of their 'real' selves and their 'perceived' selves, in a way that was not captured by the questionnaires.

> *"This questionnaire is not gonna tell you much about someone's personality. Well, no, no, it will. I mean, it's gonna tell you more about how they see their personality than their actual personality... which is probably a different thing entirely".*

Other participants agreed, and felt that the complexity of phenomena such as self-esteem and personality simply cannot be captured accurately by a scale (e.g., "*I don't think you can limit people's personality to questions*" and*"there's no sort of nuance.").* Interestingly, some participants spoke about the researchers' role in this perception of limits and were conscious of the researcher's function in the questions. For example, one participant referred to questionnaires as "*there's somebody's trying to assess what way you think*" which made them "*always aware of when I'm answering stuff like that*". Another participant expressed frustration at the end of the study:

> *"So what is the outcome? ... if a person scale scores highly on self-esteem or low on self-esteem, what within the field of what this questionnaire is used for? What does that actually say about a person?"*

Other participants also showed frustration at the limits of questionnaires to capture nuance and complexity (e.g., *"I think it's the subtleties that are missing in questionnaires like this").* One participant, for example, spoke at length about how these feelings of inadequacy of questionnaire measures may be particularly problematic for people from marginalised groups, and found it difficult to know how to answer some of the quantitative questions:

> *"Different demographics of people and marginalised groups have far more complicated identities that they have to navigate, construct, and engage with. A questionnaire such as this... does not have the space in order to look at different ways that we have to do that as marginalised groups."*

Overall, this theme encapsulates participants' feelings that quantitative questionnaire measures themselves may not be considered an accurate or adequate tool, from the perspective of the lay audience, to capture authentic, nuanced, and rich accounts of the human experience.

Theme 3. Hidden Labour of Questionnaires

Beyond attempting to capture multiple selves and concerns with the ability of questionnaires to recognise nuance and richness, the third theme that was observed in the dataset was a wider issue of the hidden labour associated with completing questionnaires. While some participants did not exert much effort and time into completing the questionnaire (*"I am a very simple kind of person", "I don't like to overthink"),* more commonly, participants discussed in the cognitive interviews the labour involved with completing each measure. Across the transcripts, participants referred to the questionnaires as "*head games*" and "*so complex*", which prompted challenging thinking (e.g., "*thoughts are actually flying around my head*" and "*this is very complex for me to do it. You know what I mean?*"). Among participants, there was a concern for 'correctly' or accurately completing the questionnaire:

> *"I mean, have I thought about my personality before? Do I know who I am? How I thought about different aspects of my personality myself or I mean the one thing that I, I think this questionnaire is measuring."*

Participants often reflected on the authenticity of their responses. Some participants reported absolute honesty in their questionnaire responses; for example, one participant who reported that they have "*actually done a lot of surveys like this as part of other studies, like on prolific*" discussed the authenticity and honesty of questionnaires and concluded that "*there's no reason not to be honest… I'm really talking about situations in the real world*". However, some participants described active consideration of "*what the researcher wants*", in a way that goes beyond the standard consideration of demand characteristics. Importantly, this was often done in a way that was well-meaning and prosocial, with the researchers in mind. For example, one participant perceived the Rosenberg Self-Esteem Scale as creating a "*depression score*" and wanted to make it clear that they are not depressed, to help the researcher (e.g., "*you can kind of guess what the questions are are are aiming at...well, I'm not a depressed person.*") Participants ranged in their explicitness of this concern. One participant, for example, interrupted the interview on multiple occasions to check-in with the interviewer to ensure that their answers were useful ("*I'm trying to think, [Interviewer's Name], how is this helping you? How will this help you down the line?*")

This represents a perspective on questionnaire research that is often overlooked; that is, participants themselves wish to be well attuned with the researcher's needs from the research and adopt their answers to aid this. This can lead, in some cases, to hidden labour or *"overthinking loads"* on the participants' side. For example, one participant in our study contacted us after the interview to provide more insights and reflections on their experiences of completing questionnaires:

> *"...the other day I responded to a follow up survey on how UK consumers think about meat. It was unfortunate that the survey came when it did. Because of the cost-of-living crisis, I haven't been eating meat as much. Consequently, my responses made it look as though I had really made an effort to cut down on meat because of the environment! So my behaviour could be viewed by the researcher in a way that wasn't correct. I sent an email explaining why I responded the way I did."*

This demonstrates the vast hidden labour, cognition, and effort that participants often employ when completing questionnaire research, which should be acknowledged more explicitly in conversations surrounding measurement validity and reliability. This also represents a more nuanced picture of measurement validity, as it demonstrates the complex nature of completing survey items from a lay perspective.

Discussion

Overall, in this study we were interested in understanding lay perceptions of psychology measures. Taken together, the think aloud analysis and thematic analysis of the interview transcripts demonstrate that participants across the measures experienced issues in interpreting and responding to items. We identified three dominant themes that each relate to lay perceptions of psychology measurements, which explores this further. These were: (1) participants' grappling with attempting to capture their 'multiple selves' and temporality in the questionnaires, (2) participants perceiving the questionnaire method as generally ‘missing nuance and richness’ and (3) participants exposing the ‘hidden labour of questionnaires’. This study has provided a comprehensive understanding of the lived experience of lay people completing psychology measurements to inform the conversation about psychology’s measurement crisis.

The themes identified by our thematic analysis of the cognitive interviews corroborate some of the ongoing meta-scientific work in psychology measurement, and thus serves to add nuance and richness to these conversations. For example, the first theme of participants attempting to capture multiple selves through questionnaire items echoes experimental research which shows the impact of social context on questionnaire responses, such that social identity salience can influence questionnaire items (e.g., Cinnirella, 1998). This theme may also be best understood through the lens of McConnell’s (2011) multiple self-aspects framework, which demonstrates the link between self-representation and context. As McConnell (2011) posits, constructions of ‘the self’ comprise multiple different self-aspects, which are context dependent. Secondly, the second theme of ‘missing nuance and richness’ corroborates the notion of *multiple* validities in questionnaire research (see Vazire et al., 2022). Finally, the theme of ‘hidden labour of questionnaires’ also aligns well with longstanding concerns surrounding response fatigue (Rolstad et al., 2011). Therefore, while the methods adopted here are relatively novel in this context, the findings largely extend and corroborate existing concerns within the more quantitative appraisal of psychological measurement.

Implications

Overall, this work has revealed that there is more to completing questionnaires than meets the academic eye; participants are not *just* answering questions, but instead engage in a high level of cognitive effort to understand measures, interpret questions, and, in some cases, attempt to pre-empt the researcher’s needs to be ‘helpful’ participants in psychological research. This is an important facet to measurement practice that has received less attention in conversations surrounding measurement practice in psychology, specifically when considering the context of reappraising scales rather than just initial validation. Indeed, qualitative approaches that centre lay perceptions of measures have historically been used well in the *development* and initial validation of scales; for example, Boateng et al. (2018) highlight how the process of ‘evaluation by target population’ constitutes best practice in scale design (see also Haynes et al., 1995). However, beyond initial scale development, more meta-scientific investigations into the robustness and appropriateness of psychological measures typically focus on quantitative psychometric outcomes or statistical evaluations of robustness (see Flake & Fried, 2020 for a summary). Therefore, our findings highlight that this may miss important intricacies and nuances of the instrument’s functioning (i.e., response processes, scores interpretability, user comprehensibility of item content) and thus more qualitative approaches, such as those adopted here, may be valuable to complement and extend existing efforts to improve the state of psychological measures.

Taken together, our think aloud findings and thematic analysis suggest that qualitative approaches to validity offer real opportunities to improve latent instruments in psychology, and indeed the social sciences more generally. Whereas the majority of the existing literature focuses on increasingly sophisticated statistical methods and to attend to response biases and inattentive responding, we argue that scholars can also learn crucial insights from participants' experience, particularly given how our findings here offer richness to existing concerns within measurement practice. We also advocate for the centring of lay perceptions in metascience, which has seen promising work in this area recently (e.g., see Bottesini et al.'s, 2022 analysis of lay perceptions of research practices). By integrating social constructivist epistemologies to their toolset and applying it to their measurement practices, scholars incrementally build a path towards a more comprehensive, participant-informed assessment of latent construct validity.

Future Directions

Without centring lived experience of measurement, psychologists risk limiting the breadth and depth of insights during the validation process, jeopardising not only the degree to which the content of the instrument is an adequate reflection of the measured construct but also the verity of its inferences. To improve the robustness of questionnaires, we thus argue that psychologists not only adopt qualitative, lay-person centred approaches to item construction and instrument development, but also to investigations of existing measures' appropriateness and rigour. Moreover, scale validation should be programmatic and dynamic—i.e., a continuous, iterative process rather than a static one. Therefore, journal editors should invest and support efforts to continuously validate, update, and test the robustness of scales by publishing well-designed studies that leverage both qualitative and quantitative approaches to exploring the validity, reliability of measures.

To extend this work, there is now potential for future research, such as comparing how non-expert and expert (i.e., researcher) participants interpret items on popular psychology measures, similar to that of studies such as Belzer et al. (2013), Boness and Sher (2020), Peterson et al. (2017), Schildmann et al. (2016), and Wright et al. (2021) uncovering the hidden benefits of cognitive interviewing for survey item validation and scale development.

Limitations

It is important to note here that we appreciate that this methodology does not allow us to comment *directly* on the questionable research practices that Flake and Fried (2020) outline. That is, in investigating lay perceptions of psychology measures, this does not speak to any problematic or questionable practices within the measurement development themselves. Rather, our study contributes to the broader *reappraisal* of measurement practices across psychology, zeroing in on the significance of integrating qualitative approaches to strengthen the robustness, validity, and credibility of quantitative inferences.

Conclusion

Social scientists have referred to a looming measurement crisis in psychology—and beyond—because latent instruments manifest gender- and race-based biases (Azevedo et al., 2023; Barabas et al., 2014; Pérez & Hetherington, 2014), are non-invariant across socio demographics (Davis et al., 2016; Harper & Rhodes, 2021; Hussey & Hughes, 2020; Pietryka & MacIntosh, 2022), display poor psychometric validity properties (e.g., van Hauwaert et al., 2020), and instruments attempting to measure the same construct seldom do so successfully (e.g., Azevedo & Bolesta, 2022; Weidman et al., 2017; Weidmann et al., 2023). This is not to say measurement is doomed to fail but rather suggestive that over relying on the positivist quantitative approach to the validation of latent instruments can be limiting, providing only a part of a larger picture. If the basic measurement properties of widely used constructs are flawed, it is likely that insights from research will be biased. Valid, invariant, and theoretically derived instruments are urgently needed for the reliable accumulation of knowledge in psychological science, and qualitative approaches provide useful tools towards these goals.

Funding: This project was funded by a British Psychological Society Undergraduate Research Assistantship awarded to Dr Madeleine Pownall and Joseph Mason.

Acknowledgments: We are grateful to the participants who were interviewed for this project, who were generous in sharing their insights.

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: *Joseph Mason*—Idea, conceptualization | Design planning | Data collection | Data management (storage, curation, processing, etc.) | Writing | Feedback, revisions. *Madeleine Pownall*—Idea, conceptualization | Design planning | Data collection | Data management (storage, curation, processing, etc.) | Writing | Feedback, revisions | Supervision, mentoring | Funding to conduct the work. *Amy Palmer*—Data management (storage, curation, processing, etc.) | Feedback, revisions. *Flavio Azevedo*—Writing | Feedback, revisions | Supervision, mentoring.

Ethics Statement: Ethical approval was obtained from the University of Leeds School of Psychology Ethics Committee on 27th April 2022, Reference: PSYC-523.

Supplementary Materials

The Supplementary Materials contain all study materials and coding framework (see Pownall & Mason, 2022), as well as the registration of the study plan (Stage 1 manuscript) after in-principle acceptance (IPA) of the current Registered Report (see Pownall et al., 2022).

Index of Supplementary Materials

Pownall, M., & Mason, J. (2022). *Investigating lay perceptions of psychological measures using a think aloud and cognitive interview protocol: A registered report* [Study materials]. OSF. https://osf.io/th9w4/

Pownall, M., Mason, J., Palmer, A., & Azevedo, F. (2022). *Investigating lay perceptions of psychological measures using a think aloud and cognitive interview protocol: A registered report* [Preregistration]. OSF Registries. https://doi.org/10.17605/osf.io/58ek4

References

Anvari, F., Efendic, E., Olsen, J., Arslan, R. C., Elson, M., & Schneider, I. K. (2022, April 12). *Bias in self-reports: An initial elevation phenomenon.* PsyArXiv. https://doi.org/10.31234/osf.io/udwrg

Azevedo, F. (2023). *Ideologies, ideological asymmetries, and the roots of political behavior* [Unpublished PhD Thesis]. Cologne University.

Azevedo, F., & Bolesta, D. (2022). *Measuring ideology: Current practices, consequences, and recommendations.* https://measuring.ideology.flavioazevedo.com

Azevedo, F., Liu, M., Pennington, C. R., Pownall, M., Evans, T. R., Parsons, S., Elsherif, M. M., Micheli, L., Westwood, S. J., & Framework for Open, Reproducible Research Training (FORRT). (2022). Towards a culture of open scholarship: the role of pedagogical communities. *BMC Research Notes, 15*(1), Article 75. https://doi.org/10.1186/s13104-022-05944-1

Azevedo, F., Micheli, L., & Bolesta, D. S. (2023). Does stereotype threat contribute to the political knowledge gender gap? A preregistered replication study of Ihme and Tausendpfund (2018). *Journal of Experimental Political Science.* Advance online publication. https://doi.org/10.1017/XPS.2022.35

Balgiu, B. A. (2018). The psychometric properties of the Big Five inventory-10 (BFI-10) including correlations with subjective and psychological well-being. *Global Journal of Psychology Research: New Trends and Issues, 8*(2), 61–69. https://doi.org/10.18844/gjpr.v8i2.3434

Barabas, J., Jerit, J., Pollock, W., & Rainey, C. (2014). The question (s) of political knowledge. *American Political Science Review, 108*(4), 840–55. https://doi.org/10.1017/S0003055414000392

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior, 41*, 12–18. https://doi.org/10.1177/1090198113483139

Belzer, F., Schmidt, S., Lucius-Hoene, G., Schneider, J. F., Orellana-Rios, C. L., & Sauer, S. (2013). Challenging the construct validity of mindfulness assessment—A cognitive interview study of the Freiburg mindfulness inventory. *Mindfulness, 4*(1), 33–44. https://doi.org/10.1007/s12671-012-0165-7

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Boness, C. L., & Sher, K. J. (2020). The case for cognitive interviewing in survey item validation: A useful approach for improving the measurement and assessment of substance use disorders. *Journal of studies on alcohol and drugs, 81*(4), 401–404. https://doi.org/10.15288/jsad.2020.81.401

Bottesini, J. G., Rhemtulla, M., & Vazire, S. (2022). What do participants think of our research practices? An examination of behavioural psychology participants' preferences. *Royal Society Open Science, 9*(4), Article 200048. https://doi.org/10.1098/rsos.200048

Branney, P. E., Brooks, J., Kilby, L., Newman, K., Norris, E., Pownall, M., Talbot, C. V., Treharne, G. J., & Whitaker, C. M. (2023). Three steps to open science for qualitative research in psychology. *Social and Personality Psychology Compass, 17*(4), Article e12728. https://doi.org/10.1111/spc3.12728

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Brown, G., & Sotardi, V. (2019, March). *Evaluation of the BFI10 Personality Inventory with New Zealand university students: Failing psychometric tests.* Poster presented at the 2019 International Congress of Psychological Science (ICPS), Paris, France. https://doi.org/10.17608/k6.auckland.7813472

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour 6*, 29–42. https://doi.org/10.1038/s41562-021-01193-7

Cinnirella, M. (1998). Manipulating stereotype rating tasks: Understanding questionnaire context effects on measures of attitudes, social identity and stereotypes. *Journal of Community & Applied Social Psychology, 8*(5), 345–362. https://doi.org/10.1002/(SICI)1099-1298(1998090)8:5<345::AID-CASP441>3.0.CO;2-F

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102*(4), 874–888. https://doi.org/10.1037/a0027403

Darker, C. D., & French, D. P. (2009). What sense do people make of a theory of planned behaviour questionnaire? A think-aloud study. *Journal of Health Psychology, 14*(7), 861–871. https://doi.org/10.1177/1359105309340983

Davis, D. E., Rice, K., van Tongeren, D. R., Hook, J. N., DeBlaere, C., Worthington, E. L., Jr., & Choe, E. (2016). The moral foundations hypothesis does not replicate well in Black samples. *Journal of Personality and Social Psychology, 110*(4), Article e23. https://doi.org/10.1037/pspp0000056

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist, 77*(4), 576–588. https://doi.org/10.1037/amp0001006

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

Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research.* Cambridge University Press. https://doi.org/10.1017/CBO9780511518201

Gilhooly, K. J., & Green, C. (1996). Protocol analysis: Theoretical background. In J. T. E. Richardson (Ed.), *Handbook of qualitative research methods for psychology and the social sciences* (pp. 140–150). BPS Books.

Goodson, I. F. (1990). Studying curriculum: Towards a social constructionist perspective. *Journal of Curriculum Studies, 22*(4), 299–312. https://doi.org/10.1080/0022027900220401

Harper, C. A., & Rhodes, D. (2021). Reanalysing the factor structure of the moral foundations questionnaire. *British Journal of Social Psychology, 60*(4), 1303–1329. https://doi.org/10.1111/bjso.12452

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hilton, C. E. (2017). The importance of pretesting questionnaires: a field research example of cognitive pretesting the Exercise referral Quality of Life Scale (ER-QLS). *International Journal of Social Research Methodology, 20*(1), 21–34. https://doi.org/10.1080/13645579.2015.1091640

Hoppmann, T. K. (2009). Examining the ‘point of frustration’. The think-aloud method applied to online search tasks. *Quality & Quantity, 43*(2), 211–224. https://doi.org/10.1007/s11135-007-9116-0

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E. J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Jamieson, M. K., Govaart, G. H., & Pownall, M. (2023). Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass, 17*(4). https://doi.org/10.1111/spc3.12735

Kane, M. J., Smeekens, B. A., Meier, M. E., Welhaf, M. S., & Phillips, N. E. (2021). Testing the construct validity of competing measurement approaches to probed mind-wandering reports. *Behavior Research Methods, 53*(6), 2372–2411. https://doi.org/10.3758/s13428-021-01557-x

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. *The Journal of Sex Research, 57*(6), 722–742. https://doi.org/10.1080/00224499.2019.1695244

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363–374. https://doi.org/10.2307/2529786

Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior Research Methods, 43*(2), 548–567. https://doi.org/10.3758/s13428-011-0066-z

Lewis, N. A., Jr. (2021). What counts as good science? How the battle for methodological legitimacy affects public psychology. *American Psychologist, 76*(8), 1323–1333. https://doi.org/10.1037/amp0000870

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

McConnell, A. R. (2011). The multiple self-aspects framework: Self-concept representation and its implications. *Personality and Social Psychology Review, 15*(1), 3–27. https://doi.org/10.1177/1088868310371101

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Morris, R., MacNeela, P., Scott, A., Treacy, P., Hyde, A., O'Brien, J., Lehwaldt, D., Byrne, A., & Drennan, J. (2008). Ambiguities and conflicting results: The limitations of the kappa statistic in establishing the interrater reliability of the Irish nursing minimum data set for mental health: A discussion paper. *International Journal of Nursing Studies, 45*(4), 645–647. https://doi.org/10.1016/j.ijnurstu.2007.07.005

Murphy, M., Hollinghurst, S., & Salisbury, C. (2018). Qualitative assessment of the primary care outcomes questionnaire: A cognitive interview study. *BMC Health Services Research, 18*(1), Article 79. https://doi.org/10.1186/s12913-018-2867-6

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies, 41*(3), 385–397. https://doi.org/10.1006/ijhc.1994.1065

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ...Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23*(3), 217–243. https://doi.org/10.1080/1047840X.2012.692215

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Osberg, T. M. (1987). The convergent and discriminant validity of the need for cognition scale. *Journal of Personality Assessment, 51*(3), 441–450. https://doi.org/10.1207/s15327752jpa5103_11

Parry, D. A., Davidson, B. I., Sewall, C. J., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour, 5*(11), 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., ...Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour, 6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Pérez, E. O., & Hetherington, M. J. (2014). Authoritarianism in black and white: Testing the cross-racial validity of the child rearing scale. *Political Analysis, 22*(3), 398–412. https://doi.org/10.1093/pan/mpu002

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development, 50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Pietryka, M. T., & MacIntosh, R. C. (2022). ANES scales often do not measure what you think they measure. *The Journal of Politics, 84*(2), 1074–1090. https://doi.org/10.1086/715251

Pownall, M., Talbot, C. V., Henschel, A., Lautarescu, A., Lloyd, K. E., Hartmann, H., Darda, K. M., Tang, K. T. Y., Carmichael-Murphy, P., & Siegel, J. A. (2021). Navigating open science as early career feminist researchers. *Psychology of Women Quarterly, 45*(4), 526–539. https://doi.org/10.1177/03616843211029255

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality, 41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology, 125*(6), 840–851. https://doi.org/10.1037/abn0000184

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health, 14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and Commitment Therapy: Measures Package, 61*(52), 18–20.

Rosenberg, M. (1979). *Conceiving the self.* Basic Books.

Sadowski, C. J., & Gulgoz, S. (1992). Internal consistency and test-retest reliability of the Need for Cognition Scale. *Perceptual and Motor Skills, 74*(2), 610. https://doi.org/10.2466/pms.1992.74.2.610

Satchell, L. P., Fido, D., Harper, C. A., Shaw, H., Davidson, B., Ellis, D. A., Hart, C. M., Jalil, R., Bartoli, A. J., Kaye, L. K., Lancaster, G. L. J., & Pavetich, M. (2021). Development of an Offline-Friend Addiction Questionnaire (O-FAQ): Are most people really social addicts? *Behavior research methods, 53*(3), 1097–1106. https://doi.org/10.3758/s13428-020-01462-9

Schildmann, E. K., Groeneveld, E. I., Denzel, J., Brown, A., Bernhardt, F., Bailey, K., Guo, P., Ramsenthaler, C., Lovell, N., Higginson, I. J., Bausewein, C., & Murtagh, F. E. M. (2016). Discovering the hidden benefits of cognitive interviewing in two languages: The first phase of a validation study of the Integrated Palliative Care Outcome Scale. *Palliative Medicine, 30*(6), 599–610. https://doi.org/10.1177/0269216315608348

Schmidt, H., Brandt, D., Bischof, A., Heidbrink, S., Bischof, G., Borgwardt, S., & Rumpf, H. J. (2022). Think-aloud analysis of commonly used screening instruments for Internet use disorders: The CIUS, the IGDT-10, and the BSMAS. *Journal of Behavioral Addictions.* https://doi.org/10.1556/2006.2022.00034

Schmidt, K., Buchanan, E. M., & Hall, B. F. (2022). Registered report: Moderators of the relationship between implicit and explicit measures of evaluation and identification. *Journal of Experimental Social Psychology, 100*, Article 104280. https://doi.org/10.1016/j.jesp.2021.104280

Smith, B. (2018). Generalizability in qualitative research: Misunderstandings, opportunities and recommendations for the sport and exercise sciences. *Qualitative Research in Sport, Exercise and Health, 10*(1), 137–149. https://doi.org/10.1080/2159676X.2017.1393221

Soubelet, A., & Salthouse, T. A. (2017). Does need for cognition have the same meaning at different ages? *Assessment, 24*(8), 987–998. https://doi.org/10.1177/1073191116636449

Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2022, March 23). The road to hell is paved with good intentions: How common practices in scale construction hurt validity. *PsyArXiv.* https://doi.org/10.31234/osf.io/p3zxa

Tobin, S. J., & Guadagno, R. E. (2022). Why people listen: Motivations and outcomes of podcast listening. *Plos One, 17*(4), Article e0265806. https://doi.org/10.1371/journal.pone.0265806

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Tuten, T. L., & Bosnjak, M. (2001). Understanding differences in web usage: The role of need for cognition and the five factor model of personality. *Social Behavior and Personality: An International Journal, 29*(4), 391–398. https://doi.org/10.2224/sbp.2001.29.4.391

van Hauwaert, S. M., Schimpf, C. H., & Azevedo, F. (2020). The measurement of populist attitudes: Testing cross-national scales using item response theory. *Politics, 40*(1), 3–21. https://doi.org/10.1177/0263395719859306

van Oort, L., Schröder, C., & French, D. P. (2011). What do people think about when they answer the Brief Illness Perception Questionnaire? A 'think-aloud' study. *British Journal of Health Psychology, 16*(2), 231–245. https://doi.org/10.1348/135910710X500819

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science, 31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17*(2), 267–295. https://doi.org/10.1037/emo0000226

Weidmann, R., Chopik, W. J., Ackerman, R. A., Allroggen, M., Bianchi, E. C., Brecheen, C., Campbell, W. K., Gerlach, T. M., Geukes, K., Grijalva, E., Grossmann, I., Hopwood, C. J., Hutteman, R., Konrath, S., Küfner, A. C. P., Leckelt, M., Miller, J. D., Penke, L., Pincus, A. L., ... Back, M. D. (2023). Age differences in narcissism: A comprehensive study across eight measures and over 250,000 participants. *PsyArXiv.* https://doi.org/10.31234/osf.io/5ygtc

Wetzel, E., & Roberts, B. W. (2020). Commentary on Hussey and Hughes (2020): Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(4), 505–508. https://doi.org/10.1177/2515245920957618

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Sage.

Willis, G. B., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology, 5*(3), 251–267. https://doi.org/10.1002/acp.2350050307

Wolcott, M. D., & Lobczowski, N. G. (2021). Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning, 13*(2), 181–188. https://doi.org/10.1016/j.cptl.2020.09.005

Wongpakaran, T., & Wongpakaran, N. (2012). A comparison of reliability and construct validity between the original and revised versions of the Rosenberg Self-Esteem Scale. *Psychiatry Investigation, 9*(1), 54–58. https://doi.org/10.4306/pi.2012.9.1.54

Wright, J., Moghaddam, N., & Dawson, D. L. (2021). Cognitive interviewing in patient-reported outcome measures: A systematic review of methodological processes. *Qualitative Psychology, 8*(1), 2–29. https://doi.org/10.1037/qup0000145

Social Psychological Bulletin (SPB) is an official journal of the Polish Social Psychological Society (PSPS).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.