journal of cognition

Qualitative Individual Differences are Useful, but Reliability Should be Assessed and Not Assumed

SPECIAL COLLECTION: THEORETICAL REVIEW WITH COMMENTARIES: QUALITATIVE INDIVIDUAL DIFFERENCES

COMMENTARY

CRAIG HEDGE

ubiquity press

ABSTRACT

Rouder and Haaf (2021) propose that studying qualitative individual differences would be a useful tool for researchers. I agree with their central message. I use this commentary to highlight examples from the literature where similar questions have been asked, and how researchers have addressed them with existing tools. I also observe that while the hierarchical Bayesian framework is a useful tool for studying individual differences, it does not relieve us of the requirement to evaluate the forms of reliability that are critical to our research questions.

CORRESPONDING AUTHOR:
Craig Hedge
Aston University, School of Psychology, College of Health & Life Sciences, Birmingham, UK; Cardiff University, School of Psychology, Cardiff, UK
c.hedge@aston.ac.uk

KEYWORDS:
Statistical analysis; Cognitive Control; Mathematical modelling

TO CITE THIS ARTICLE:
Hedge, C. (2021). Qualitative Individual Differences are Useful, but Reliability Should be Assessed and Not Assumed. *Journal of Cognition*, 4(1): 48, pp. 1–4. DOI: *https://doi.org/10.5334/joc.169*

Rouder and Haaf (2021) outline an approach in which qualitative variations in experimental effects can be a valuable tool to researchers, by constraining theory and leading us to questions about individual differences that we might not reach otherwise. I am keen to see this approach embraced and hope this commentary will contribute to the discussion on where and how it can be best used.

DO RESEARCHERS THINK ABOUT QUALITATIVE INDIVIDUAL DIFFERENCES?

Rouder and Haaf ask whether the identification of qualitative individual differences speaks to researchers. Indeed, I think the 'does everybody…' question formalises intuitions that I and colleagues have had about our data. This prompted me to consider how I would previously have tackled these kinds of questions.

In the discussion of a recent paper examining individual differences in strategic changes in the speed-accuracy trade-off (SAT), I reflected on whether every individual responds to the instruction to prioritise speed over accuracy in the same way (Hedge et al., 2019). We typically assume that the SAT reflects a change in how much evidence an individual requires to make a decision, though it may also involve changes in how evidence is processed (Rae et al., 2014). From this I speculated '*does everybody* change in more than just their strategy?'. I ran an analysis for one dataset here to illustrate what I would do with my existing toolkit.[1] I fit multiple variants of a cognitive model to each participant's data separately. I then used model comparisons to determine the most parsimonious account for each person. One model assumed that participants only changed their threshold for how much evidence is required, which was sufficient to explain the data of 10 out of 81 participants. A second model assumed that the duration of perceptual and motor processing changed in addition to their threshold (47/81). A third further assumed that the efficiency of evidence processing was affected (24/81). Based on this, I would conclude that there are qualitative individual differences in the speed-accuracy trade-off.

I also see qualitative individual differences assumed or implied in the literature. Responder analyses are common in clinical studies and are sometimes used in other areas (e.g. working memory training, see Tidwell et al., 2014). There, researchers define a cut-off that represents a clinically significant effect, and then may ask what covaries with the presence or absence of that effect. Finally, while Haaf and Rouder (2017) have shown that "everybody Stroops", perhaps not everybody spider-Stroops (Watts et al., 1986). Using a mixed ANOVA, Watts et al. show that spider related words interfere with colour naming for individuals with a spider phobia to a greater extent than controls. They also show that the distribution of observed interference scores is centred close to zero for controls.

The questions in these examples do not take the same form as the ones posed by Rouder and Haaf. The model comparisons in the speed-accuracy trade-off example do not ask about the direction of changes or formally test whether qualitative individual differences are present at the sample level. In the spider-Stroop example, testing for an interaction in an ANOVA does not ask whether some individuals show null or negative (true) effects. However, these existing approaches may be a barrier to the uptake of new tools if they scratch the same theoretical itch for researchers.

ARE ESTIMATED TRUE EFFECTS RELIABLE/STABLE?

A desirable property of the hierarchical Bayesian approach proposed by Rouder and Haaf is the ability to estimate 'true' effects from observed effects by incorporating assumptions about the magnitude of trial noise. There has been recent concern that traditionally measured experimental effects taken from widely used tasks are not as reliable as we would like them to be for asking questions about individual differences (e.g. Hedge et al., 2018; Parsons et al., 2019). The same concern applies to qualitative individual differences in principle – it is difficult to identify factors that covary with the presence or absence of an effect if an individual sometimes shows an effect and sometimes does not. Do hierarchical models relieve us of this concern?

[1] Data are the random-dot motion task data from Hedge et al. (2019). The script and model outputs for this new analysis are on the OSF (***https://osf.io/yh4ak/***).

The idea of true scores is associated with classical test theory (Novick, 1966), where it is assumed that observed measures reflect variation in people's true values on the dimension of interest plus some measurement error. When we assess reliability, we estimate the ratio of signal (variation in true scores) to noise (error) by examining the consistency of participants' scores over some form of repeated measurement. It has been stated that reliability is not a property of tasks or procedures, but rather a property of a set of scores obtained from a given population (Wilkinson, 1999). This is because the magnitude of the signal and the noise are context dependent (Cronbach et al., 1963). For example, there may be more variance in true Stroop effects in a clinical population than in a healthy population. Measurement error is also potentially comprised of multiple sources. I agree with Rouder and Haaf that trial noise is a large component in reaction time-based effects (see Supplementary Material D of Hedge et al., 2018), though there can be additional sources of error in a test-retest reliability context (e.g. fluctuations in mood or health). Until we have assessed the test-retest reliability of a task in our population, we do not know if our qualitative or quantitative differences reflect stable characteristics of those individuals.

As an illustration, I applied Rouder and Haaf's quid() function to test-retest reliability data for the spatial-numeric association of response codes (SNARC) effect from Hedge et al. (2018).[2] I chose this task because the effect in mean reaction times was relatively small (15ms and 8ms in sessions one and two respectively), so it is likely that there are qualitative individual differences. I applied the analysis to session one (***Figure 1***; black) and session two (red) separately, to highlight the conclusion we would draw if we only had data from a single session. We would reach the same conclusion about whether qualitative individual differences are present from both sessions – Bayes factors favour the unconstrained model. However, if we were to ask "what kind of person shows a positive/null/negative SNARC effect", then we might select different individuals at different time points depending on how we identify them. Twelve out of forty participants have numerically positive effects in one session and negative effects in the other. Further, the 95% credible interval for twelve participants includes zero in one session and not the other.

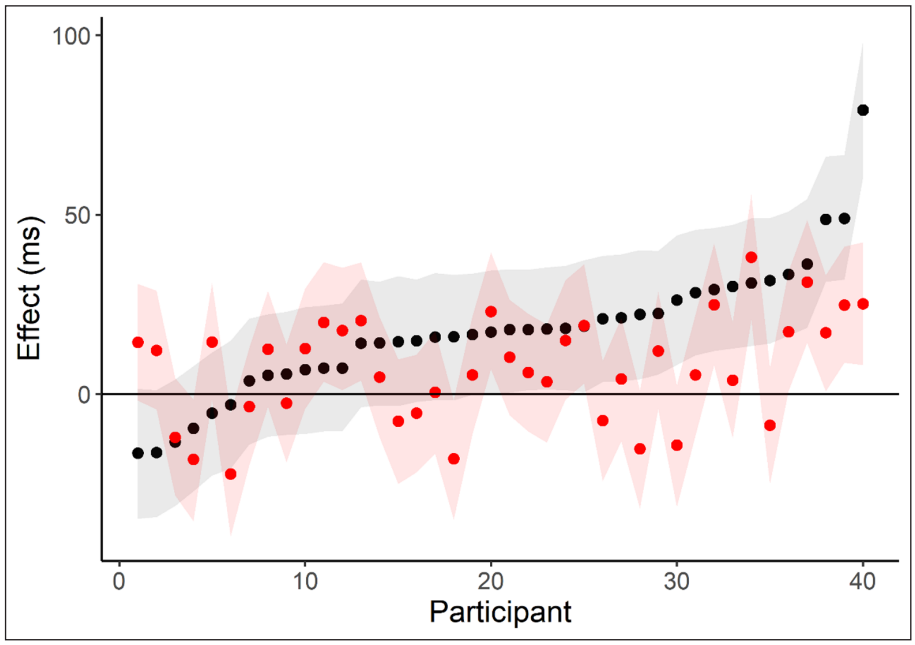

Figure 1 Estimated SNARC effects from session 1 (black) and session 2 (red) from Hedge et al., (2018). Shaded regions are 95% credible intervals. Estimates are sorted by their size in session 1.

Several papers have shown that hierarchical Bayesian models can improve our estimates of individual differences (Brown et al., 2020; Rouder & Haaf, 2019; Wiecki et al., 2013), and the illustration above does not contradict this. The key point is that we still need to evaluate the forms of reliability that are important to our research questions to be able to make appropriate generalisations about qualitative individual differences.

[2] R code for this analysis can be found on the Open Science Framework.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

Craig Hedge *orcid.org/0000-0001-6145-3319*
Aston University, School of Psychology, College of Health & Life Sciences, Birmingham, UK; Cardiff University, School of Psychology, Cardiff, UK

REFERENCES

Brown, V. M., Chen, J., Gillan, C. M., & **Price, R. B.** (2020). Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. DOI: *https://doi.org/10.1016/j.bpsc.2019.12.019*

Cronbach, L. J., Rajaratnam, N., & **Gleser, G. C.** (1963). Theory of Generalizability: a Liberalization of Reliability Theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. DOI: *https://doi.org/10.1111/j.2044-8317.1963.tb00206.x*

Haaf, J. M., & **Rouder, J. N.** (2017). Developing Constraint in Bayesian Mixed Models. *Psychological Methods*, *22*(4), 779–798. DOI: *https://doi.org/10.1037/met0000156*

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. DOI: *https://doi.org/10.3758/s13428-017-0935-1*

Hedge, C., Vivian-Griffiths, S., Powell, G., Bompas, A., & **Sumner, P.** (2019). Slow and steady? Strategic adjustments in response caution are moderately reliable and correlate across tasks. *Consciousness and Cognition*, *75*. DOI: *https://doi.org/10.1016/j.concog.2019.102797*

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18. DOI: *https://doi.org/10.1016/0022-2496(66)90002-2*

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. DOI: *https://doi.org/10.1177/2515245919879695*

Rae, B., Heathcote, A., Donkin, C., Averell, L., & **Brown, S.** (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(5), 1226–1243. DOI: *https://doi.org/10.1037/a0036801*

Rouder, J. N., & **Haaf, J. M.** (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, *26*(2), 452–467. DOI: *https://doi.org/10.3758/s13423-018-1558-y*

Rouder, J. N., & **Haaf, J. M.** (2021). Are There Reliable Qualitative Individual Difference in Cognition? *Journal of Cognition*, 4(1): 46. 1–16. DOI: *https://doi.org/10.31234/osf.io/3ezmw*

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & **Mendoza, J. L.** (2014). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin and Review*. DOI: *https://doi.org/10.3758/s13423-013-0560-7*

Watts, F. N., McKenna, F. P., Sharrock, R., & **Trezise, L.** (1986). Colour naming of phobia‐related words. *British Journal of Psychology*, *77*(1), 97–108. DOI: *https://doi.org/10.1111/j.2044-8295.1986.tb01985.x*

Wiecki, T. V., Sofer, I., & **Frank, M. J.** (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. DOI: *https://doi.org/10.3389/fninf.2013.00014*

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. DOI: *https://doi.org/10.1037/0003-066X.54.8.594*

TO CITE THIS ARTICLE:
Hedge, C. (2021). Qualitative Individual Differences are Useful, but Reliability Should be Assessed and Not Assumed. *Journal of Cognition*, 4(1): 48, pp. 1–4. DOI: *https://doi.org/10.5334/joc.169*

Submitted: 29 April 2021
Accepted: 15 May 2021
Published: 27 August 2021

Journal of Cognition is a peer-reviewed open access journal published by Ubiquity Press.