nature research

https://doi.org/10.1038/s41562-020-00976-8

Supplementary information

Calibrating the experimental measurement of psychological attributes

In the format provided by the authors and unedited

Supplementary material for Bach, Melinscak, Fleming, Voelkle (2020) Nature Human Behaviour: Calibrating the experimental measurement of psychological attributes

September 18, 2020

Contents

1 Supplementary Methods

1.1 Retrodiction model: derivation

In this section, we derive the model on the sample level; the next section then develops a probabilistic perspective. We use the formalism of linear regression models, but we do not seek to invoke any implicit assumptions about the identifiability of parameters, or about error distributions; all assumptions are explicitly stated.

We take a repeated measurement under $m > 1$ levels, $j = 1, ..., m$, of the experimental manipulation, in a sample of n subjects $i = 1, ..., n$. Let e_{ij} be the intended change of the psychological attribute, t_{ij} the true score of the attribute, and y_{ij} the estimated (i.e. measured) true score, in subject i and condition j.

First, we write t as a function of e. We are only interested in score differences between conditions, and not in baseline values, and so include an intercept term in our model. For within-subjects designs, this can be a subject-specific baseline; for between-subjects designs it is a group-level baseline:

$$t_{ij} = t_{0i} + f(e_{ij}) + \omega_{ij}, \tag{1}$$

where, t_{0i} is an (unknown) baseline value of the attribute, f the (unknown and potentially non-linear) mapping from e to changes in t, and ω_{ij} an (unknown) imprecision term that captures variations in f between subjects as well as variation in the effectiveness of the experimental manipulation.

Since t is on an arbitrary scale, the ideal relationship between e and t is not generally an identity mapping but a linear mapping, and we decompose eq. (1) into a linear part and a residual (R) non-linearity f_R:

$$t_{ij} = \bar{t}_i + \gamma e_{ij} + f_R\left(e_{ij}\right) + \omega_{ij}, \tag{2}$$

where $\gamma > 0$ a scalar. For within-subjects designs, $\bar{t}_i$ is the (unknown) within-subject expected value of t across levels of e, and for between-subjects designs, $\bar{t}_i \coloneqq \bar{t}$ is the expected value of t across levels of e and over subjects.

Since we cannot empirically separate these aberration terms, we combine them into a total (T) aberration $\omega_{Tij} \coloneqq f_R(e_{ij}) + \omega_{ij}$. Thus, our final linear model is:

$$t_{ij} = \bar{t}_i + \gamma e_{ij} + \omega_{Tij}. \tag{3}$$

For a fixed value of e and a fixed subject, we assume that ω_T is independent of other subjects and of other experimental levels. However, because it includes systematic aberration, the expected value may be non-zero for a fixed value of e. Notably, we are not interested in estimating the coefficient γ since the scaling of psychological attributes is arbitrary. In the next section, we will uniquely specify γ and ω_{Tij} on the population level.

Next, we write the estimated score y as a function of the true score t. Again, we assume a potential systematic misspecification g in the measurement model, that is, a lack of trueness, and a subject-specific error ε_{ij}, that is, an imprecision. Then we can write

$$y_{ij} = y_{0i} + g(t_{ij}) + \varepsilon_{ij}, \tag{4}$$

where y_{0i} is a baseline. Because the ideal relationship between t and y is linear, we again rewrite this as the sum of a linear term and a residual non-linearity, noting that we cannot generally assume $g_R(t_{ij}) = 0$:

$$y_{ij} = \overline{y}_i + \beta(t_{ij} - \bar{t}_i) + g_R(t_{ij}) + \varepsilon_{ij}. \tag{5}$$

Here, $\beta > 0$ a scalar. For within-subject designs, $\overline{y}_i$ is the within-subject expected value of t across levels of e, and for between-subjects designs, $\overline{y}_i \coloneqq \overline{y}$ is the expected value of y across levels of e and over subjects. We denote the total error $\varepsilon_{Tij} \coloneqq g_R(t_{ij}) + \varepsilon_{ij}$ such that our final true score model is:

$$y_{ij} = \overline{y}_i + \beta(t_{ij} - \bar{t}_{ij}) + \varepsilon_{Tij}. \tag{6}$$

1.2 Retrodiction model: probabilistic analysis

In this section, we consider the retrodiction model on a probabilistic level and expand it in tutorial style. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an arbitrary probability space. Denote expectation and variance with $\mathbb{E}$ and $\mathbb{V}$, respectively.

Let $E : \Omega \to \{e_{11}, .., e_{mn}\}$ be a discrete random variable that takes the a priori defined intended values. The scaling of E is arbitrary and hence we can set, without loss of generality:

$$\mathbb{E}(E) \coloneqq 0, \qquad \mathbb{E}\left(E^2\right) = \mathbb{V}(E) \coloneqq 1. \tag{7}$$

Let $T : \Omega \to \mathbb{R}$ be a random variable that takes values $t_{ij} - \bar{t}_i$, and $Y : \Omega \to \mathbb{R}$ a random variable that takes values $y_{ij} - \overline{y}_i$. Let $\omega_T : \Omega \to \mathbb{R}$ and $\varepsilon_T : \Omega \to \mathbb{R}$ be random variables that will be defined later. Assume that $Cov(E, T) > 0$ and $Cov(T, Y) > 0$.

Now we can rewrite eq. (3) and (6) as:

$$T = \gamma E + \omega_T \tag{8}$$

and

$$Y = \beta T + \varepsilon_T \tag{9}$$

From the definition of T and Y, it follows that

$$\mathbb{E}(T) = \mathbb{E}(Y) = \mathbb{E}(\omega_T) = \mathbb{E}(\varepsilon_T) = 0. \tag{10}$$

Up to now, the values of the coefficients γ, β and the random variables ω_T, ε_T are unspecified. As in ordinary least squares regression, we now define γ such that $\mathbb{V}(\omega_T)$ takes its minimum, by setting

$$\frac{\partial \mathbb{V}(\omega_T)}{\partial \gamma} = \frac{\partial \mathbb{V}(T - \gamma E)}{\partial \gamma} = \frac{\partial \left(\mathbb{V}(T) + \gamma^2 \mathbb{V}(E) - 2\gamma Cov(E, T)\right)}{\partial \gamma} = 0, \tag{11}$$

which leads us to

$$2\gamma \mathbb{V}(E) - 2Cov(E, T) = 0 \tag{12}$$

such that

$$\gamma \coloneqq \frac{Cov(E, T)}{\mathbb{V}(E)} = Cov(E, T). \tag{13}$$

This choice of γ ensures that

$$Cov(E, \omega_T) = \mathbb{E}(E(T - \gamma E))$$

$$= \mathbb{E}(ET) - \gamma \mathbb{E}\left(E^2\right)$$

$$= Cov\left(E, T\right) - \gamma = 0.$$

Similarly, let

$$\beta \coloneqq \frac{Cov\left(T, Y\right)}{\mathbb{V}\left(T\right)}, \tag{14}$$

such that $\mathbb{V}\left(\varepsilon_T\right)$ is minimised and $Cov\left(T, \varepsilon_T\right) = 0$. Now all terms in our model are defined.

To simplify the sequel, and to make our main result independent of arbitrary rescalings of Y, we now define the standardised experimental aberration

$$\omega \coloneqq \frac{\omega_T}{\gamma}, \tag{15}$$

and the standardised measurement error

$$\varepsilon \coloneqq \frac{\varepsilon_T}{\beta\gamma}, \tag{16}$$

such that

$$T = \gamma E + \gamma\omega \tag{17}$$

and

$$Y = \beta T + \beta\gamma\varepsilon = \beta\gamma E + \beta\gamma\omega + \beta\gamma\varepsilon. \tag{18}$$

To prepare the derivation of our main result, we note the following identities:

$$\mathbb{V}\left(T\right) = \mathbb{V}\left(\gamma E + \gamma\omega\right) = \gamma^2\left(1 + \mathbb{V}\left(\omega\right)\right) \tag{19}$$

$$\mathbb{V}\left(Y\right) = \mathbb{V}\left(\beta T + \beta\gamma\varepsilon\right) = \beta^2\gamma^2\left(1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)\right) \tag{20}$$

$$Cov\left(E, \varepsilon\right) = \mathbb{E}\left(\left(E + \omega\right)\varepsilon - \omega\varepsilon\right) \tag{21}$$

$$= \mathbb{E}\left(\frac{1}{\gamma}T\varepsilon\right) - \mathbb{E}\left(\omega\varepsilon\right) \tag{22}$$

$$= -Cov\left(\omega, \varepsilon\right). \tag{23}$$

Using the definition of β and expressions (19, 20), we can expand the correlation between true and measured score:

$$Cor\left(T, Y\right) = \frac{Cov\left(T, Y\right)}{\sqrt{\mathbb{V}\left(T\right)\mathbb{V}\left(Y\right)}} \tag{24}$$

$$= \frac{\beta\mathbb{V}\left(T\right)}{\gamma\sqrt{\left(1 + \mathbb{V}\left(\omega\right)\right)}\beta\gamma\sqrt{\left(1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)\right)}} \tag{25}$$

$$= \frac{\beta\gamma^2 \left(1 + \mathbb{V}\left(\omega\right)\right)}{\beta\gamma^2 \sqrt{1 + \mathbb{V}\left(\omega\right)} \sqrt{1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)}} \tag{26}$$

$$= \frac{\sqrt{\left(1 + \mathbb{V}\left(\omega\right)\right)}}{\sqrt{1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)}}. \tag{27}$$

After all this preparation, we can now finally expand retrodictive validity:

$$Cor\left(E, Y\right) = \frac{Cov\left(E, Y\right)}{\sqrt{\mathbb{V}\left(E\right)\mathbb{V}\left(Y\right)}} \tag{28}$$

$$= \frac{\mathbb{E}\left(E\left(\beta\gamma E + \beta\gamma\omega + \beta\gamma\varepsilon\right)\right)}{\beta\gamma\sqrt{1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)}} \tag{29}$$

$$= \frac{\beta\gamma\mathbb{E}\left(E^2 + E\omega + E\varepsilon\right)}{\beta\gamma\sqrt{1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)}} \tag{30}$$

$$= \frac{1 - Cov\left(\omega, \varepsilon\right)}{\sqrt{1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)}} \tag{31}$$

$$= \frac{1 - \sqrt{\mathbb{V}\left(\varepsilon\right)\mathbb{V}\left(\omega\right)} Cor\left(\omega, \varepsilon\right)}{\sqrt{1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)}}. \tag{32}$$

If $Cor\left(\omega, \varepsilon\right) = 0$ then

$$Cor\left(E, Y\right) = \frac{1}{\sqrt{1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)}} \tag{33}$$

and we can see that

$$Cor\left(T, Y\right) = \sqrt{1 + \mathbb{V}\left(\omega\right)} Cor\left(E, Y\right). \tag{34}$$

To see how changes in $Cor\left(\omega, \varepsilon\right) \neq 0$ and in $\mathbb{V}\left(\varepsilon\right)$ influence $Cor\left(E, Y\right)$, we differentiate eq. (32) with respect to these two variables.

First, if all variances are strictly positive, we see that for all values of $\mathbb{V}\left(\varepsilon\right)$

$$\frac{\partial Cor\left(E, Y\right)}{\partial Cor\left(\omega, \varepsilon\right)} < 0. \tag{35}$$

In words, as $Cor\left(\omega, \varepsilon\right)$ decreases, retrodictive validity increases.

Next, we analyse the impact of changes in $\mathbb{V}\left(\varepsilon\right)$. We have:

$$\frac{\partial Cor\left(E, Y\right)}{\partial \mathbb{V}\left(\varepsilon\right)} = -\frac{\sqrt{\mathbb{V}\left(\omega\right)\mathbb{V}\left(\varepsilon\right)} + \left(\mathbb{V}\left(\omega\right)^2 + \mathbb{V}\left(\omega\right)\right) Cor\left(\omega, \varepsilon\right)}{2\sqrt{\mathbb{V}\left(\omega\right)\mathbb{V}\left(\varepsilon\right)}\left(1 + \mathbb{V}\left(\omega\right) + \mathbb{V}\left(\varepsilon\right)\right)^{\frac{3}{2}}}. \tag{36}$$

If all variances are strictly positive, this expression is negative if

$$Cor\left(\omega, \varepsilon\right) > -\frac{\sqrt{\mathbb{V}\left(\omega\right)\mathbb{V}\left(\varepsilon\right)}}{\mathbb{V}\left(\omega\right)^2 + \mathbb{V}\left(\omega\right)} \tag{37}$$

$$= -\frac{\sqrt{\mathbb{V}(\varepsilon)}}{\sqrt{\mathbb{V}(\omega)}\left(\mathbb{V}(\omega)+1\right)}. \tag{38}$$

In words, if $Cor(\omega, \varepsilon)$ is larger than the bound stated above, then decreasing $\mathbb{V}(\varepsilon)$ increases retrodictive validity. Otherwise, increasing $\mathbb{V}(\varepsilon)$ increases retrodictive validity.

2 Supplementary Results

2.1 Main Result

Theorem 1. *Assume a calibration experiment with intended values* E, $\mathbb{E}(E) = 0$, $\mathbb{V}(E) = 1$. *Let* T *be the true score with associated standardised aberration*

$$\omega = \frac{T}{Cov(E,T)} - E,$$

which does not depend on the measurement method. Let Y *be a measured score with associated standardised error*

$$\varepsilon = \frac{1}{Cov(E,T)}\left(\frac{\mathbb{V}(Y)}{Cov(T,Y)}Y - T\right).$$

Similarly define, for the same experiment, Y_1, Y_2, ε_1 *and* ε_2.

(1) If $Cor(\omega, \varepsilon_1) = Cor(\omega, \varepsilon_2) = 0$, *then* $Cor(E, Y_2) > Cor(E, Y_1) \implies \mathbb{V}(\varepsilon_2) < \mathbb{V}(\varepsilon_1)$.

(2) If $Cor(\omega, \varepsilon) = 0$, *then* $Cor(T, Y) = \sqrt{1 + \mathbb{V}(\omega)}Cor(E, Y)$.

(3) If $Cor(E, Y_2) > Cor(E, Y_1)$ *then at least one of the following three statements is true:*

(a) $Cor(T, Y_2) > Cor(T, Y_1)$ and $Cor(\omega, \varepsilon_i) > -\frac{\sqrt{\mathbb{V}(\omega)\mathbb{V}(\varepsilon_i)}}{\mathbb{V}(\omega)^2+\mathbb{V}(\omega)}$, $i = 1, 2$.

(b) $Cor(\omega, \varepsilon_2) < Cor(\omega, \varepsilon_1)$.

(c) $Cor(\omega, \varepsilon_2) \leq -\frac{\sqrt{\mathbb{V}(\omega)\mathbb{V}(\varepsilon_2)}}{\mathbb{V}(\omega)^2+\mathbb{V}(\omega)}$.

Proof. (1) follows directly from eq. (33). (2) follows from eq. (34). (3a-c) follow from eqs. (32) and (38). □

Without proof, we note that an analogous theorem holds in a finite sample, as can be demonstrated geometrically (see version 1 of this paper's pre-print: [1]).

In the following, we explain this theorem and give an intuition about how it can be used. In general it is reasonable to assume $Cor(\omega, \varepsilon) = 0$. In this case, selecting Y to increase retrodictive validity $Cor(E, Y)$ also reduces $\mathbb{V}(\varepsilon)$, the standardised measurement error, and thus increases measurement accuracy, $Cor(T, Y)$. Otherwise, if $Cor(\omega, \varepsilon)$ is positive, or if the standardised measurement error is large compared to the experimental aberration, then selecting Y with large retrodictive validity may still either ensure large accuracy, but could

also decrease $Cor(\omega, \varepsilon)$ (i.e. make aberration and error more anticorrelated). If $Cor(\omega, \varepsilon)$ is already negative and below the bound stated above (implying that the measurement error is small compared to the experimental aberration), then increasing retrodictive validity may actually reduce accuracy.

Below, we explore situations where the assumption $Cor(\omega, \varepsilon) = 0$ is violated, but we note that we regard these scenarios as edge cases. Notably, the theorem makes assumptions about the relation between aberration and error, but not on the size of the aberration as long as $Cor(E, T) > 0$. Even a weakly effective experimental manipulation can serve to evaluate accuracy of a measurement method. Although the estimates of retrodictive validity will be more uncertain in such cases, this can be mitigated by larger sample sizes. Crucially, in cases where theory only predicts ordinal differences between two conditions, we need to make additional assumptions to use the method – namely that the achieved difference in true score is constant over participants. This additional assumption however will increase aberration but will not impact on the ranking of measurement methods, as long as aberration and error are uncorrelated.

2.2 Examples

Scenario 1. No noise correlation (calibrating a reward learning measure)

A research team seeks to optimise a measure of subjective value estimates in operant reward conditioning. In a 2-alternative forced choice task, subjects decide on which (reward-predicting) cue they want to obtain. Researchers use these choices as observables, together with a measurement model that takes the form of a sigmoid curve (e.g., logistic regression), to infer subjective values. However, they have noticed a large variability in reaction times, and developed a drift-diffusion model that takes account of this variability in reaction times in order to estimate subjective value. As a retrodiction experiment, they use an experimental manipulation with 2 values (low reward vs. high reward) and overtrain subjects in this task. It is likely that this overtraining procedure minimises the imprecision component of experimental aberration. Because there are only two levels of the experimental manipulation, there is no systematic aberration term. Finally, it seems unlikely that the experimental aberration and the measurement error are correlated: there is no plausible reason why the measurement model would attenuate the estimated value difference for subjects with higher true value differences, and amplify estimated value difference for subjects with lower true value differences. No hidden stable confounds are known that impact on subjective value and behavioural preference in opposing directions.

Scenario 2. Estimating measurement uncertainty

A research team investigates presurgical epilepsy patients and has discovered orbitofrontal neurons with firing that closely corresponds to intended subjective

values in the calibration experiment from scenario 1 - much more closely than the estimate derived from observable behaviour. The residual unexplained variance in the relation between intended values and neural firing can thus serve to estimate an upper bound on ω, which yields an upper bound on $Cor\,(T,Y) = \sqrt{1+\mathbb{V}\,(\omega)}Cor\,(E,Y)$.

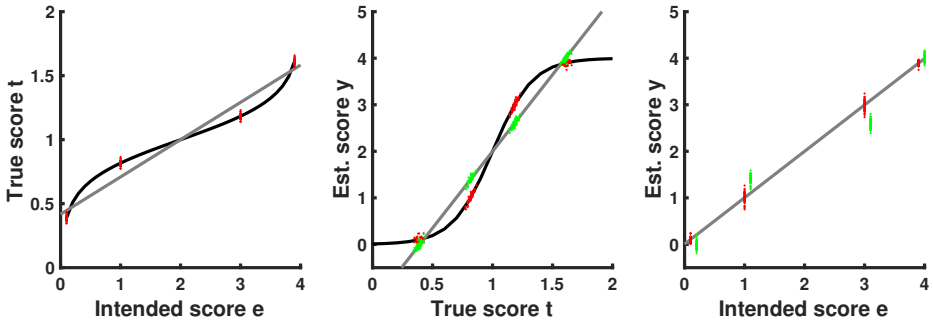
Scenario 3. Anticorrelated imprecision (calibrating a pain measure)

A research team is interested in assessing subjective pain intensity in subjects who cannot communicate. To this end, they use pupil size as the observable, together with a biophysical model of the pupil response. As a retrodiction experiment, they use 2 levels of pain (low versus high) in healthy subjects. It is known that there is biological variability in pain perception, introducing variability in the difference between low and high subjective pain. What the researchers do not know in this (entirely hypothetical) scenario is that because of genetic linkage, subjects with high pain sensitivity tend to have smaller pupils, thus limiting the dynamic range for pain-induced pupil dilation. Thus, subjects with higher true subjective pain difference will tend to have smaller pupil dilation differences and thus smaller estimated subjective pain difference. In this case, aberration and measurement error due to imprecision are anticorrelated. The researchers have also asked people for their subjective pain perception. This measure has the same measurement error variance, but here the error is uncorrelated with the variability in pain sensitivity, i.e. aberration. Because of this, the pupil-based measure has higher retrodictive validity and is erroneously preferred.

Proposed solution: (a) Report known predictors of between-subjects differences in the effectiveness of the experimental manipulation, and in the measurement. (b) Explore, report, and if possible include in the measurement model, multiplicative scalings and limits in the dynamic range of a measure.

Scenario 4. Anticorrelated inaccuracy (validating a measure on itself; Figure S1)

A research team seeks to validate a novel method of measuring subjective valence of a stimulus. As a retrodiction experiment, they take a database of pictures that have been previously rated by a large sample in terms of their valence. They use 4 distinct valence levels and show the pictures to a new sample of subjects, in which they assess both their subjective rating (red dots in Figure S1) and the novel measure (green dots in Figure S1). To define intended score differences e, they take advantage of the previous picture ratings. However, in this hypothetical example, subjects' reported valence follows their true (i.e. actually experienced) subjective valence by a sigmoid function (Figure S1 left). The relation between the previous ratings (intended scores) and the actually induced true scores is thus the perfect inverse of this function (Figure S1 middle). Hence, the relation between intended valence, and valence measured by self-report, will be almost perfectly linear (Figure S1 right, grey dots),



Supplementary Figure 1: Scenario 4, validating a measure on itself. The systematic aberration in the experimental model is exactly the inverse of systematic error in the measurement model. The resulting estimated scores (red dots) are linearly related to intended scores. In comparison, a measurement model without systematic error (green dots) appears to have a larger error. In this case, aberration and error are perfectly anti-correlated, such that reducing the measurement error ε would counterintuitively reduce retrodictive validity. This edge case arises from an invalid calibration procedure.

while the novel measure has a lower correlation with e. At the request of a reviewer, the researchers repeat their analysis by using pairs of levels of e. This removes (almost) all systematic aberration from the relation between e and t, and consequently the two measures showed equally good retrodictive validity in this analysis. To the researchers (who don't know about the true relationship between e, t, and y) this means that there must be an systematic aberration in the specification of e or a systematic error in the measurement of y.

Proposed solution: For multiple values of e, always perform an auxiliary analysis on pairs of levels, thus removing anticorrelated systematic aberration and trueness. If this auxiliary analysis consistently yields conclusions that are different from the primary analysis, then further investigation is required.

3 Supplementary Discussion

Here, we highlight some substantive questions associated with selection of calibration experiments. Our criterion guarantees sensitivity but not specificity - as common in experimental psychology, specificity must be guaranteed by the experimental procedures that are used for calibration and for substantive experiments. Therefore, the first question is whether a manipulation can be entirely specific. For example, some theories of Pavlovian aversive conditioning posit that at least two independent forms of aversive memory are established concurrently: implicit and declarative memory [2, 6]. In this case, a fear conditioning experiment would affect both attributes, with possibly different experimental

aberration. If we use a fear conditioning experiment to select between different aversive memory measurement methods, then maximising retrodictive validity may unintentionally prioritise the contribution of the lower-aberration attribute (e.g. declarative memory) to the estimated score, over the higher-aberration one (e.g. implicit memory). It appears unlikely that this could happen when comparing two transformation methods for the same observable. However, it has been suggested that different observables - skin conductance and startle eyeblink, for example - are influenced to a different degree by implicit and declarative memory [7]. In this case, basing the choice of observables on retrodictive validity may be problematic. Of course, if psychological attributes are generally indistinguishable by observation, then it makes limited sense to separate them theoretically. A second question is the specificity of the estimated score. For example, skin conductance responses are influenced not only by aversive memory but by a range of other psychological attributes [3]. We can carefully ensure these other attributes are not affected in the retrodiction experiment. However, if they are not held constant in a substantive experiment, then inference on the psychological attribute is limited. Notably, both of these issues limit measurement methods independent of which approach we select to validate them. Here, we suggest making these issues explicit by specifying validity conditions for a retrodiction experiment. Furthermore, we suggest reporting all empirical data and previous knowledge that may suggest the presence or absence of (anti)correlated experimental aberration and measurement error. For example, this may include known stable predictors of interindividual differences in the experimental effect on the observable, or known predictors of the dynamic range of the observable. In the case of more than two experimental levels, we suggest auxiliary analysis with pairs of levels to confirm the conclusions.

One limitation of our criterion is that it does not separately assess trueness and precision of measurement. We note that if two measures of T have exactly the same retrodictive validity, then trueness and precision can be decomposed by assessing reliability, which is affected only by precision and not by trueness [4]: the estimate with higher reliability will have higher precision and lower trueness, and vice versa. Assessing reliability in experimental contexts is however not without challenges. First, the number of observables is usually limited, often to one observable at a time, such that stability over observables cannot be assessed. Secondly, because we are concerned with volatile attributes, assessing stability of measurement requires ensuring that the attribute itself is stable over measurements, which requires a calibration approach. Third, as [4] illustrate, under constant measurement precision, reliability scales with interindividual variability in the psychological attribute. However, experimental manipulations have often evolved to minimise, rather than maximise, individual differences in the psychological attribute [5], which minimises metrics of reliability. Without strong assurance of trueness, reliability can be erroneously increased by inferring a different attribute that has higher interindividual variability. Fourth, when interindividual variability and temporal volatility of the attribute are on the same order of magnitude, stable hidden confounds with high interindividual variability become important. As an example, systematic differences in skin

composition can multiplicatively scale skin conductance responses, and thereby can have an impact on the measures scores. This interindividual variability is presumably much more stable than interindividual variability in the psychological attribute, which will likely result in lower reliability once such confounds are accounted for.

Supplementary References

[1] Dominik R Bach, Filip Melinscak, Stephen M Fleming, and Manuel Voelkle. Calibrating the experimental measurement of psychological attributes. *preprint available at http://https://psyarxiv.com/bhdez*.

[2] A Bechara, D Tranel, H Damasio, R Adolphs, C Rockland, and A R Damasio. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, 269(5227):1115–8, Aug 1995.

[3] Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.

[4] Andreas M Brandmaier, Elisabeth Wenger, Nils C Bodammer, Simone Kühn, Naftali Raz, and Ulman Lindenberger. Assessing reliability in neuroimaging research through intra-class effect decomposition (iced). *Elife*, 7, 07 2018.

[5] Craig Hedge, Georgina Powell, and Petroc Sumner. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50(3):1166–1186, 2018.

[6] Peter F Lovibond and David R Shanks. The role of awareness in pavlovian conditioning: empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, 28(1):3, 2002.

[7] Marieke Soeter and Merel Kindt. Dissociating response systems: erasing fear from memory. *Neurobiol Learn Mem*, 94(1):30–41, Jul 2010.