Calibrating the experimental measurement of psychological attributes

Dominik R Bach[1-3*], Filip Melinščak[3], Stephen M Fleming[1,2,4], Manuel C Voelkle[5,6]

[1]*Wellcome Centre for Human Neuroimaging, University College London, United Kingdom*

[2]*Max Planck UCL Centre for Computational Psychiatry and Aging Research, University College London, United Kingdom*

[3]*Computational Psychiatry Research, Department of Psychiatry, Psychotherapy, and Psychosomatics, Psychiatric Hospital, University of Zurich, Switzerland*

[4]*Department of Experimental Psychology, University College London, United Kingdom*

[5]*Psychological Research Methods, Humboldt University, Berlin, Germany*

[6]*Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany*

**Corresponding author (*d.bach@ucl.ac.uk)*

Abstract

Behavioural researchers often seek to experimentally manipulate, measure, and analyse latent psychological attributes, such as memory, confidence, or attention. The best measurement strategy is often difficult to intuit. Classical psychometric theory, mostly focused on individual differences in stable attributes, offers little guidance. Hence, measurement methods in experimental research are often based on tradition and differ between communities. Here, we propose a criterion, which we term retrodictive validity, that provides a relative numerical estimate of the accuracy of any given measurement approach. It is determined by performing calibration experiments to manipulate a latent attribute, and assessing the correlation between intended and measured attribute values. Our approach facilitates optimising measurement strategies, and quantifying uncertainty in the measurement. Thus, it allows power analyses to define minimally required sample sizes. Taken together, our approach provides a metrological perspective on measurement practice in experimental research that complements classical psychometrics.

Introduction

When planning behavioural experiments, researchers must decide which observables to collect (observation), and how to pre-process them (transformation), before performing statistical analyses. In many fields of behavioural science and psychology, there are no hard criteria to make these decisions, although they can have a drastic impact on the conclusions from a given study [1-3]. Often they are based on common laboratory practice or expert consensus (e.g. [4, 5]), under the implicit assumption that tradition and expertise have evolved to approximate the best method. However, recent research has highlighted a wide variability in observation [6] and transformation [2, 3, 7, 8] methods within different fields of psychology. In this paper, we develop a quantitative criterion for evaluating measurement methods in the context of experimental research. We ground our approach in classical validity theory and seek to surmount its shortcomings by integrating metrological concepts from technology.

Experimental measurement in psychology

We constrain our focus to the experimental study of the human mind, which includes many fields of psychology. As the mind is not directly observable, its attributes are assessed from observable behaviour, such as verbal expressions, motor responses, or physiological processes. Thus, the psychological inverse problem is how to infer a latent psychological attribute from an observation [9], a process often termed measurement.

Across sciences, there are at least two questions associated with measurement: whether it is meaningful, and whether it is accurate. The first question is addressed by measurement theory, concerned with the formal representation of empirical observations as numbers, and the rules that can be applied to these numbers [10]. For example, a majority of psychologists

represent observations such as response times with real numbers and treat them as if they were on an interval scale, i.e. additive [11]. Measurement theory prescribes fundamental axioms any representation must obey to be truly additive [10]. These axioms can be empirically tested. For example, two equal weights combined must weigh as much as the sum of the individual weights, an operation termed concatenation. Because one cannot concatenate psychological attributes in this way, representational measurement theory provides alternative tests of additivity [10]. Measurement theory operates on idealised empirical observations. For example, the measurement of the same weight with the same instrument is regarded as invariant [10], which does not account for the measurement error present in even the most precise weight measurements [12]. For weight measurement, this error is relatively small, and can be "averaged out" by repeated measurement. This situation is rather different in psychology, where measurement error can be on the same order of magnitude as differences between experimental conditions. This makes any test of measurement axioms challenging – and indeed they have only been investigated in the subdisciplines of psychophysics, item-response theory, and behavioural economics [10].

The second question is addressed by metrology, which is concerned with the quantification of measurement error through calibration, and its reduction by suitable technology [13]. A related field in psychology is psychometrics. Metrology assumes a true attribute score (without any realist claims on its existence outside measurement), and an (often probabilistic) measurement model that describes how this true score relates to the observation. Measurement can be cast as inference on the true score [12]. The quality of a measurement is judged by its accuracy. Given hypothetical repetitions of the measurement, accuracy can be decomposed into two components: low variability of the inferred attribute under constant true scores (precision, i.e. low random measurement error, also termed

variance), and low average distance from the true scores (trueness, i.e. low systematic measurement error, also termed bias) [14]. We note that "trueness" alone is sometimes referred to as "accuracy" in the wider literature; here we use metrological conventions.

Our proposal is grounded in this second, metrological perspective and aims at reducing measurement error. In doing so, we hope to advance the first perspective as well, by facilitating empirical tests of measurement axioms.

Classical psychometric concepts: construct validity and reliability

According to a psychometric perspective, measurement methods should be valid and reliable [15]. These crucial concepts were developed to evaluate the measurement of stable attributes for which the true scores are unknown [16]. To evaluate the measured score, the unknown true score is surrogated with a known variable, termed criterion: a concurrent measurement related to the attribute in question (concurrent validity), a process or observation that is influenced by the attribute (predictive validity), or properties of the measurement instrument itself (content validity) [17]. However, because there is usually no singular criterion, researchers form a nomological net that defines how the studied attribute, in theory, relates to other attributes or observables. A measurement of the attribute is considered to have construct validity if it occupies the same place in the nomological net as the attribute itself [16]. Because there is no method to combine the observed correlations within the nomological net into a single number [16], and because the predicted correlations are usually specified in loose terms rather than as precise coefficients [16, 18], the concept of construct validity cannot serve to quantify *trueness* and *precision*.

Classical reliability, on the other hand, assesses how interindividual differences in the measurement are stable across repetitions - over time or over test items. This addresses measurement precision but not trueness [16]. Indeed, improving reliability may even reduce trueness. For example, if one replaces a standard intelligence test score with a measurement of index finger length, the inferred attribute will be very reliable, but is unlikely to have a strong relation with actual intelligence. Thus, interpreting reliability metrics requires a criterion to guarantee trueness [16].

Retrodictive validity

Classical validity theory is built on the premise that the true score is unknown, and that there is no observable variable (outside the measurement to be evaluated) that captures all relevant variance in the true score. Therefore, classical validity theory cannot provide a single criterion for validity assessment. However, in experimental research on volatile attributes, the true score can be influenced by experimental manipulation. This creates an opportunity to apply the metrological concept of calibration, which is based on measurement in a standardized experiment. We propose that intended values of the true score in such a calibration experiment can provide a singular criterion to assess accuracy (Figure 1A). We term this type of criterion validity "retrodictive validity", since the aim is to retrodict the (experimentally induced) values of the psychological attribute. Note that we have previously used the term "predictive validity" [19, 20], which confusingly refers to a different concept in classical validity theory and as such we have dropped it in more recent publications [21]. We illustrate this approach with a worked example before discussing the general conditions under which this framework will yield improved accuracy. Table 1

provides an exemplary and non-exhaustive list of further application examples across different subfields of psychology.

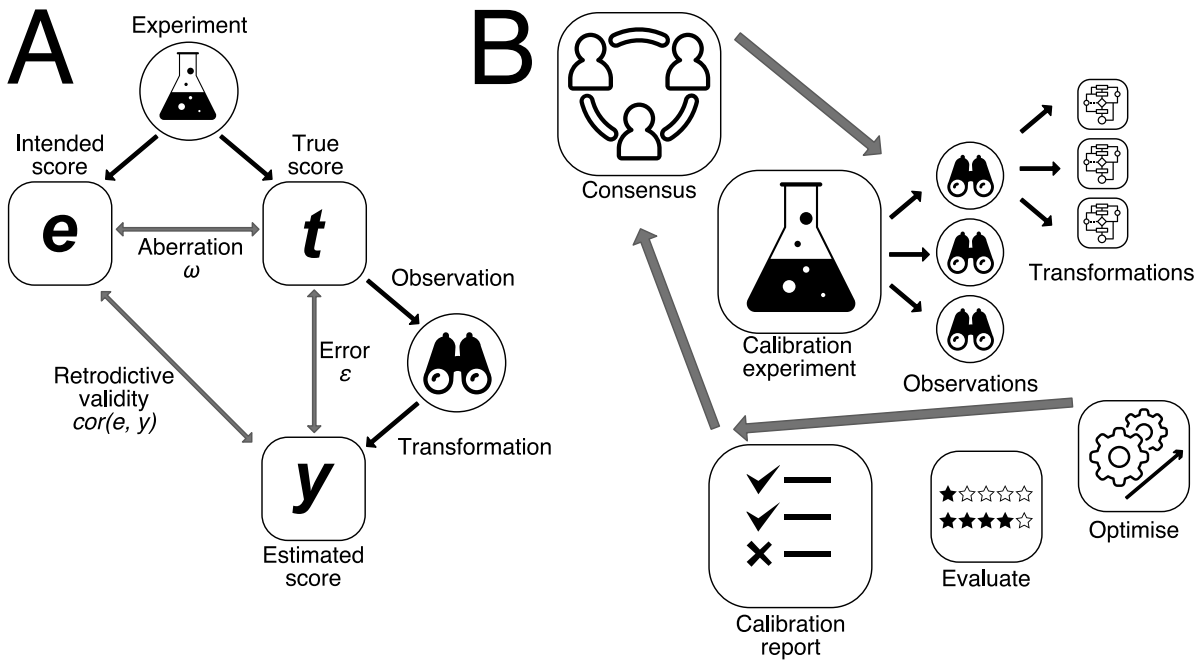

Figure 1. Retrodiction and calibration. A: A standardised experiment with intended attribute scores *e* generates true scores *t*. The measured attribute, *y*, is generated by transforming some observed data. Retrodictive validity denotes the observable correlation between e and *y*, and is influenced by the measurement error as well as by the correlation between experimental aberration and measurement error, *Cor(ω, ε)*. B: The calibration process. Expert consensus defines calibration experiments. Different observables and transformations can be optimised and evaluated. The calibration report is fed back to the community and inspires refined calibration experiments, observables, and measurement models.

Table 1: Example latent attributes from different subfields of psychology for which calibration experiments appear feasible

Subfield	**Latent attribute**	**Possible calibration experiments**	**Specification of intended values per theory**	**Application outside calibration**
Perception	Perceived stimulus property (e.g. length)	Manipulation of true stimulus property	Interval scale with arbitrarily many levels (true stimulus property)	Investigating Bayesian integration of prior expectation
Learning	Stimulus-stimulus or stimulus-response association	Pavlovian conditioning, operant conditioning	Interval scale with 3 or more levels (associative learning theory)	Evaluation of learning interventions
Memory	Declarative memory	Number of repetitions in word lists	Interval scale with 3 or more levels (retrieved context theory)	Measuring clinical memory impairments
Cognition	Spatial attention	Spatial cueing task	Ordered levels	Investigating influence of spatial attention on evidence accumulation in value-based decision-making
Decision-making	Utility	Food-deprived vs. satiated state	Ordered levels	Comparing theories of economic choice
Metacognition	Decision confidence	High vs. low noise in perceptual decision	Ordered levels	Comparing metacognition across domains
Emotion	Subjective feeling of 'disgust'	Disgust-eliciting video exposure vs. neutral video	Ordered levels	Investigating the role of disgust in trauma-related disorders
Social psychology	Physical attraction	Exposure to photos of attractive physiques of preferred vs. non-preferred sex	Ordered levels	Investigating the dynamics of emerging social media platforms

Worked example 1: quantifying implicit learning

We consider a group of clinical psychologists who have proposed a novel technique to reduce trauma memory. To evaluate their intervention in healthy individuals, they experimentally create aversive associations and seek to reduce them with their novel method. To this end, they conduct an experiment in which a person associates a geometric cue with an electric shock (CS+) and another cue with no shock (CS-), a procedure often termed fear conditioning. They want to measure the ensuing associative memory after the subsequent intervention, compared to a control group with no intervention. They record each person's skin conductance response to the geometric symbols, which is known to be influenced by implicit memory for the electric shock. Then they need to find the best possible transformation for quantifying the attribute 'implicit associative threat memory' from the observed skin conductance responses. A related question is whether a different observation (such as cardiac responses) may provide an even better measurement.

In the absence of any memory intervention, a plethora of research has demonstrated in healthy individuals and using various measurement methods that CS+ is more strongly associated with electric shock than CS-. We can transform this ordinal prediction into real-valued intended values, which we denote with *e*: CS+ is assumed to instil a higher level of aversive memory (e = 1) than CS- (e = 0). One could also create more than two levels of *e* by leveraging classical associative learning theory. Here, one prediction is that the difference from CS- aversive memory for a third cue C has half the size of that for CS+ if an association was established with compound cue CX (e = 0.5).

Our proposal is to perform an independent pilot experiment, without the experimental intervention, and measure skin conductance. One can then select the data transformation

(pre-processing) method that yields the highest correlation between intended associative memory values *e*, and measured associative memory values *y*, i.e. the highest retrodictive validity. We term this a calibration experiment. In our example, the calibration procedure can be identical to control group in the planned substantive experiment, just without the planned intervention, which additionally allows power analyses (see below). The formal calibration process now proceeds in three steps.

1. Defining the measurand: The procedure that is used to create fear memory for calibration includes specifying the CS (e.g. triangles with specific size and colour), the US (e.g. electric shock with defined strength), the reinforcement schedule, the CS-US interval, the inter trial interval, the number of trials, the instructions, the preparation of the participant, and so on.

2. Validity conditions: These are the measurement conditions under which the optimised measurement method is assumed to be optimal. For example, fear memory-induced skin conductance responses occur with some latency after CS presentation. This latency is influenced by the duration and regularity of the CS-US interval, and so the CS-US interval is an important validity condition. In a future experiment with deviating CS-US interval, the optimised measurement method from the calibration experiment may not be optimal anymore. In contrast, discriminability of CS+ and CS- colour is not among validity conditions. Discriminability is suspected to influence the effectiveness of the experimental procedure, that is, the variability in true scores between participants. This impacts on retrodictive validity but is independent from any specific measurement method and is not known to influence measurement error. The next section clarifies the relation between variability in true scores (which we term experimental aberration), and measurement error.

3. Reporting the relationship: In the simple case of discriminative fear conditioning, researchers will report Cohen's *d* or Hedge's *g* for the CS+/CS- difference across participants. They will compare several methods in one sample and report the ranking of the methods.

The planned memory-editing experiment consists of a control group that receives the same treatment as in the calibration experiment, and an intervention group in which this treatment is followed by the memory-editing intervention. In this situation, we can assume that both the experimental aberration and the measurement error in the control group are the same as in the calibration experiment. This situation allows performing a power analysis for the planned experiment (Figure 2, see ref [22] for an example). Imagine that in the calibration experiment, the method with highest retrodictive validity achieved an effect size of (Cohen's) $d = 1.2$ for the within-subject CS+/CS- difference. If the intervention itself has no variation across participants (which is a best-case assumption), then it will simply shift this distribution towards zero in the intervention group. The researchers want to be able to detect a reduction in fear memory of 50% or more, with 80% power in a one-tailed *t*-test at $p < .05$. The difference between a control group that is similar to the calibration experiment, and an intervention group with 50% less fear memory, corresponds to an effect size of Cohen's $d = 0.6$, resulting in $N = 72$ participants. Any variation in the effectiveness of the intervention would increase the experimental aberration in the experimental group and further increase the required sample size.

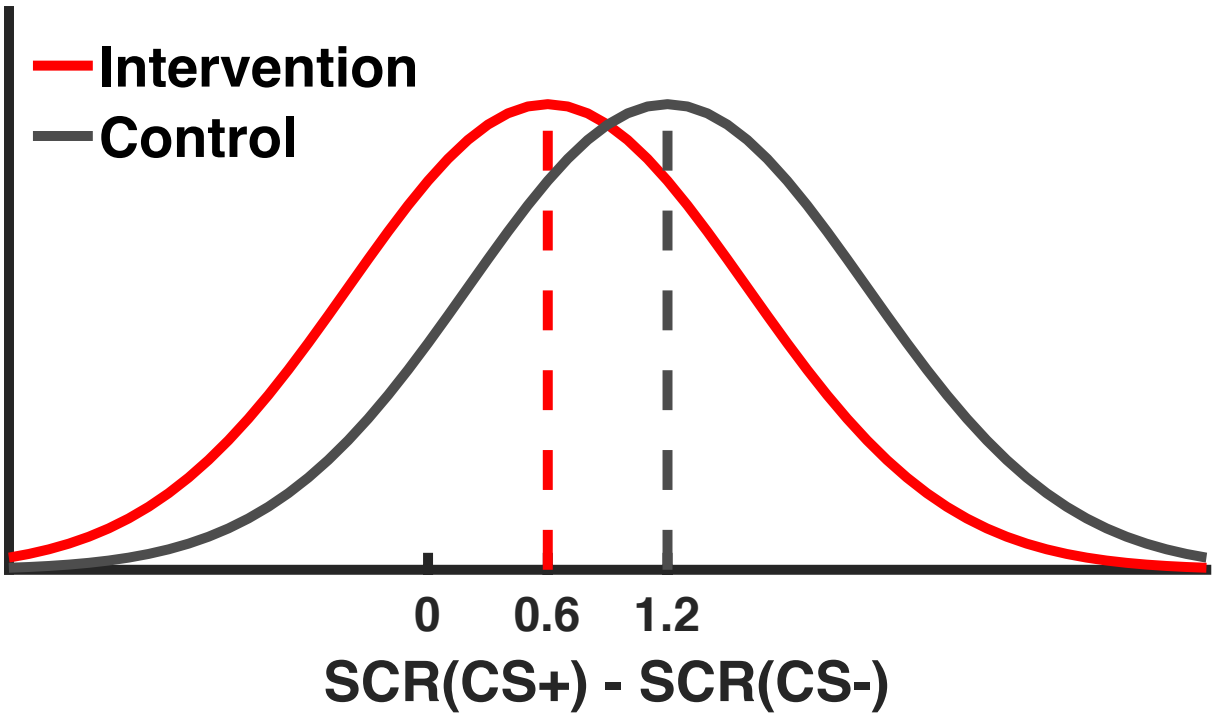

Figure 2: Power analysis. Standard normally distributed scores *y* in a calibration experiment (black) are affected by measurement error and experimental aberration. In this example, an experimental treatment is composed of the same manipulation as in the calibration experiment and either an additional intervention (red lines), or no intervention (control, black lines). In the best-case scenario of no intervention variability, the distribution of measured scores in the intervention group will be the same as in the control group with shifted mean. In this example, $d = 1.2$ in the calibration experiment, and a 50% fear memory reduction in the intervention corresponds to a between-group effect size of $d = 0.6$, resulting in $N = 72$ participants to measure this fear memory reduction with 80% power at $p < .05$ in a one-tailed t-test.

Retrodictive validity and measurement accuracy

For a formal treatment, we now define key terms (see Figure 1A for illustration and supplemental material for mathematical detail). As in classical test theory and other true score theories [23, 24], we assume the existence of real-valued *true scores* of a psychological attribute, which we denote *t*. We assume a priori that they are measurable (in a measurement-theoretic sense [10]) and on interval scale. With our (within- or between-subjects) experimental manipulation, we seek to achieve *intended differences* in *t*; we denote these experimentally intended values with *e.* We note that psychological theories differ in how quantitative their predictions are. Some theories, such as associative learning theory or perceptual decision theory, prescribe the intended values on several levels of an interval scale. Other theories may make only ordinal predictions for two levels of the attribute. In such cases, we specify *e* by assuming a fixed average difference in intended true score, which brings *e* on an interval scale. This additional assumption will usually not affect accuracy assessment, as we will see later.

We are interested in an error-free measurement of the true score from some observable quantity. We make no assumption on the measurement model that is used to transform the observable. We denote the resulting *estimate of the true score* with *y* and assume it is on an interval scale. Thus, when we evaluate the measurement method that generates *y*, we evaluate the observation method together with a measurement model or transformation method.

In the ideal case of an error-free measurement, since psychological attributes have no natural scale, there is an arbitrary linear mapping between *e*, *t*, and *y*. Any non-linearity in the mapping between these variables constitutes a misspecification of the intended values

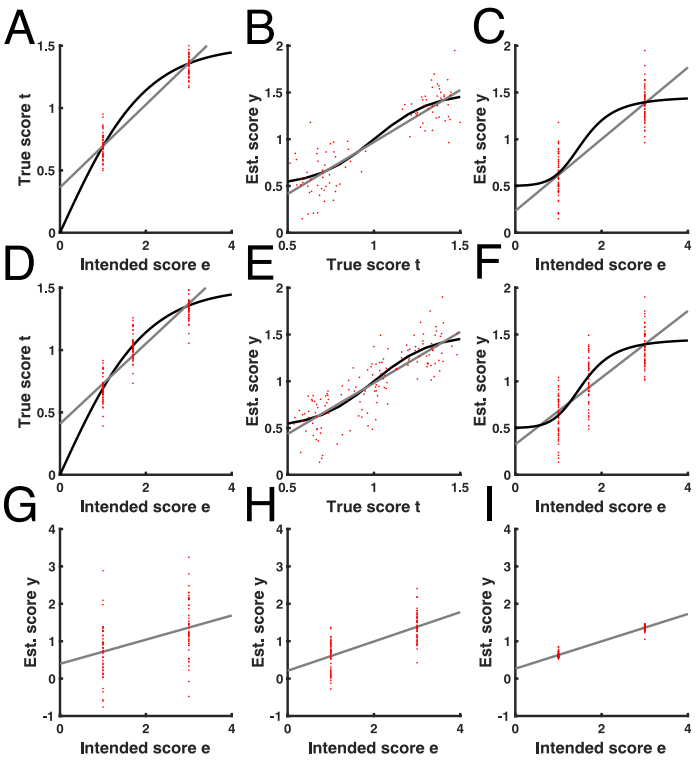

Figure 3: The retrodiction approach. A: The ideal relation between intended and true scores is a linear mapping with arbitrary coefficients (grey line), but the true relation is possibly non-linear (systematic aberration, black line) and imprecise (distribution of red dots). Because there are only two values of experimental manipulation in this example, the systematic aberration does not influence the correlation between e and t. B: Similarly, the relation of true scores and measured scores includes systematic error and imprecision. C: Resulting mapping from intended to measured scores is assessed by their correlation, that is, a linear mapping (grey line), but the true relation may be non-linear (black line, composition of the two non-linear functions in A and B), and imprecise (distribution of data points). D-F: Same model but with three (not equidistant) intended scores. Here, the systematic aberration impacts the resulting error in F. G-I: Correlation between e and y under three different levels of measurement error ε. In I, ε=0, but experimental aberration renders the resulting error non-zero.

in the underlying theory, or a misspecification of the measurement model, and so we regard it as an error term.

Our goal is to evaluate trueness and precision of *y*. If we have several measurements (e.g. participants) per level of *e*, the total measurement error is jointly influenced by trueness and precision. Our goal is to minimize the total measurement error.

First, we consider the mapping from *e* to *t*. In an ideal experiment, this would be a non-stochastic linear mapping. Any deviation from this situation constitutes *experimental aberration* ω. Aberration can be decomposed into two terms. The first is non-linearity, a systematic (i.e. across subjects) misspecification of *e*, which reduces the trueness of the experimental model. This is illustrated in Figure 3A where the black line denotes the actual non-linear dependency between *e* and *t*, which is contrasted to a linear relationship illustrated by the grey line. If there are just two levels of *e*, then this systematic aberration vanishes at the considered levels of *e* and therefore becomes irrelevant, but this is not the case for more than two levels of *e* (Figure 3D). The second component is stochastic variation in the effectiveness of the manipulation, such that for the same value of *e*, *t* takes different values in different subjects or repetitions of the experiment. This means the model of our experimental manipulation is *imprecise*. This is illustrated by the distribution of red dots in Figure 3A which depict the true score differences under a constant value of *e*.

Next, we consider the mapping from *t* to *y* (see Figure 3B). Again, we assume a potential systematic misspecification in the measurement model, that is, a lack of trueness, and stochastic error, that is, an imprecision. Together, they constitute the measurement error ε.

In the supplementary material, we mathematically derive the conditions under which maximising the (observable) correlation between intended and estimated scores, Cor(*e, y*),

minimizes measurement error. The main result is that these conditions are defined by the correlation between experimental aberration and measurement error, $\mathrm{Cor}(\omega, \varepsilon)$.

Because the experimental manipulation is usually distinct from the measurement method, it is generally reasonable to assume $\mathrm{Cor}(\omega, \varepsilon) = 0$. In this case, increasing Cor(*e, y*) is guaranteed to increase measurement accuracy. Additionally, for any fixed measurement method, Cor(*e, y*) prescribes a lower bound on measurement accuracy. This is a standard case and will apply in most circumstances. In other cases, discussed in the supplementary material, increasing Cor(*e, y*) may still increase measurement accuracy, but this not guaranteed. However, we argue that these are identifiable edge cases.

The only assumption the model makes is that the correlations between *e*, *t*, and *y*, are strictly positive – but they can be small. Thus, one can use weak theories or calibration experiments to improve measurement. In particular, the transformation of an ordinal theory into an interval-scaled variable *e* does not diminish the viability of the approach.

Calibration

Calibration is the evaluation of a measurement method under controlled circumstances, and can be broken up into several parts [13].

Defining the measurand

What is being measured in the calibration process [13] is known as the measurand – the true values of the measured attribute in our case. We need to define how they are created. We suggest using an experimental manipulation that has a relatively specific impact on the psychological attribute in question, and precisely defining the procedure by which e is

manipulated. Details will depend on the substantive research field and will generally include a definition of the population from which the test sample is drawn.

Validity conditions

The calibration results are only valid under the specified validity conditions [13]. These are conditions known to impact on the measurement method. Conditions known to impact on the experimental aberration are less important here, as they do not speak to future use of the measurement method in other experimental contexts.

Reporting the relationship

In metrology, the relationship between measured and reference values are usually reported separately as a trueness and precision [13]. Because of the presumably large aberration in psychology, these two terms cannot be separated and are jointly minimized. Because aberration influences observed retrodictive validity, we would expect that retrodictive validity *rankings* of different methods will be more generalizable than the actual effect sizes. Therefore, we suggest comparing several measurement methods in the same calibration experiment.

Iteration

Sample size of calibration studies should be reasonably large, to avoid overfitting a method to particular data sets. Often, the goal is to compare different measurement models (or transformation methods) which can be applied retrospectively to previously acquired data sets. To facilitate this in an iterative process (see Figure 1B), we suggest compiling and sharing data from calibration experiments across laboratories in standardized format (for an example see ref [25]). Current developments in data management automation could possibly enable fully automated benchmark testing as soon as a new calibration data set is published.

Further application

Besides the main goal of improving measurement accuracy, retrodictive validity allows further applications. First, by specifying measurement uncertainty [26], it allows power analyses. Often, the true size of a hypothesized effect is not known a priori, and published effect sizes tend to overestimate the true effect size [27]. In many cases, retrodictive validity can determine the maximum achievable effect size (see Figure 2 and worked Example 1). This will often render it possible to compute minimum sample sizes, required under the best-case assumptions that an experimental manipulation has no variation. This also provides a direct route to compare financial costs associated with different measurement methods.

Next, when the measurement method is kept constant, retrodictive validity is only influenced by experimental aberration, which can depend on laboratory standards and staff training. For example, testing in noisy rooms with many participants may result in lower retrodictive validity than testing the same measurement method in a quiet room. Retrodictive validity could enable quality control, by comparing different laboratories or trainees in standardised experiments. We note that current scientific practices implicitly incentivize large effect sizes in hypothesis tests [28]. Replacing these incentives with success in calibration experiments could potentially improve research culture.

Finally, one can use the retrodiction model to optimise experimental manipulations. Maximising retrodictive validity will then yield the experimental manipulation with lowest combined aberration and measurement error. This can aid experimental design. As an example, we have used this approach to empirically find the optimal number of trials to measure fear memory recall. Here, more trials mean less measurement error but at the

same time reduction of the true effect due to extinction (i.e. increased aberration). The optimal balance is difficult to intuit but can be found empirically [29].

Worked example 2: measuring decision confidence

To see how the framework can be applied in diverse research settings, we here give another concrete example. A research team seeks to characterise the influence of social conformity on decision confidence. They plan to use a perceptual decision-making task and provide social information before measuring participants' confidence. They further plan to record explicit confidence ratings, reaction time of the ratings, and key stroke force. Their goal is to identify the most precise method for integrating these observables into a confidence measure.

It is well known from decision-making research that the quality of perceptual evidence influences one's decision confidence. As calibration experiment, the researchers can thus use a random dot motion task with high and low coherence, and predict that decision confidence is higher in the high coherence condition ($e = 1$) than in the low coherence condition ($e = 0$). Using data from this experiment, they can now compute y under various different measurement models, for example a model only taking into account the explicit ratings, or multiple regression models that also incorporate reaction times and/or key force [30]. They will finally select the method with highest retrodictive validity.

The researchers can then set up their substantive experiment, perhaps using only a single, staircased level of random dot motion coherence, and test their hypothesis about the effect of social conformity on confidence in such a setting. For instance, different conditions of the experiment may provide the participant with helpful or unhelpful advice about the correct

decision on each trial. Importantly, despite the experiment no longer containing variation in coherence, the researchers can be sure that, due to selecting a confidence measure based on its high retrodictive validity, they have chosen the most accurate metric of perceptual decision confidence against which to evaluate their hypothesis.

Discussion

Retrodictive validity corresponds to accuracy of inference on a true score. It provides a framework for rational selection between, and optimisation of, measurement methods, and can be established and exploited in a calibration process. We note that this approach also applies to non-behavioural measures, such as inferring a psychological attribute (e.g. pain) from neuroimaging data [31].

As anticipated by classical validity theory [16], the method does not allow separating trueness and precision, but jointly improves both. Assessment of reliability can help in disentangling these two, as it depends on precision alone (see supplementary discussion for details). Our method is guaranteed improve accuracy as long as the experimental aberration is uncorrelated with the measurement error. It is difficult to come up with plausible cases where this condition is violated, but if substantive research reveals circumstantial evidence for any such violations then the proposed method should be used with caution.

Tradition remains the mainstay of justification for data collection and pre-processing methods in many subfields of psychology, but this comes with a range of theoretical, statistical, and practical problems, including low reproducibility. Widespread researcher degrees of freedom have been criticized [7], and there are increasing calls to plan and pre-register data pre-processing before a study is being conducted [32, 33]. This leaves research

practitioners in the uncomfortable situation of having to select between methods without good reason. Collecting huge samples increases reproducibility but imposes a heavy cost if the method itself is not optimised. Here, we propose a generic solution that can be applied across different branches of psychology and may alleviate several challenges experimental psychology is currently confronted with.

References

1. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspect Psychol Sci* **11**, 702-712 (2016).
2. Silberzahn, R.*, et al.* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* **1**, 337-356 (2018).
3. Lonsdorf, T.B.*, et al.* Navigating the garden of forking paths for data exclusions in fear conditioning research. *Elife* **8** (2019).
4. Boucsein, W.*, et al.* Publication recommendations for electrodermal measurements. *Psychophysiology* **49**, 1017-1034 (2012).
5. Blumenthal, T.D.*, et al.* Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology* **42**, 1-15 (2005).
6. Ojala, K.E. & Bach, D.R. Measuring learning in human classical threat conditioning: Translational, cognitive and methodological considerations. *Neuroscience and biobehavioral reviews* **114**, 96-112 (2020).
7. Simmons, J.P., Nelson, L.D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* **22**, 1359-1366 (2011).
8. Lonsdorf, T.B., Merz, C.J. & Fullana, M.A. Fear Extinction Retention: Is It What We Think It Is? *Biological psychiatry* **85**, 1074-1082 (2019).
9. Houwer, J.D. Why the Cognitive Approach in Psychology Would Profit From a Functional Approach and Vice Versa. *Perspect Psychol Sci* **6**, 202-209 (2011).
10. Luce, R.D. & Suppes, P. Representational measurement theory. *Stevens' handbook of experimental psychology* (2002).
11. Michell, J. The psychometricians' fallacy: Too clever by half? *Br. J. Math. Stat. Psychol.* **62**, 41-55 (2009).
12. Estler, W.T. Measurement as inference: Fundamental ideas. *Cirp Annals 1999: Manufacturing Technology, Vol 48 No 2 1999*, 611-632 (1999).
13. Phillips, S.D., Estler, W.T., Doiron, T., Eberhardt, K.R. & Levenson, M.S. A Careful Consideration of the Calibration Concept. *J Res Natl Inst Stand Technol* **106**, 371-379 (2001).
14. BIPM, I., IFCC, ILAC, IUPAC, IUPAP, ISO, OIML. The international vocabulary of metrology—basic and general concepts and associated terms (VIM). *JCGM* **200**, 2012 (2012).

15. Shadish, W.R., Cook, T.D. & Campbell, D.T. *Experimental and quasi-experimental designs for generalized causal inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell* (Boston: Houghton Mifflin, 2002).
16. Cronbach, L.J. & Meehl, P.E. Construct validity in psychological tests. *Psychol Bull* **52**, 281-302 (1955).
17. Cronbach, L.J. Five perspectives on validity argument. *Test validity*, 3-17 (1988).
18. van der Maas, H.L., Molenaar, D., Maris, G., Kievit, R.A. & Borsboom, D. Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological review* **118**, 339-356 (2011).
19. Bach, D.R. & Friston, K.J. Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology* **50**, 15-22 (2013).
20. Bach, D.R.*, et al.* Psychophysiological modeling: Current state and future directions. *Psychophysiology*, e13214 (2018).
21. Bach, D.R. & Melinscak, F. Psychophysiological modelling and the measurement of fear conditioning. *Behav Res Ther* **127**, 103576 (2020).
22. Bach, D.R., Tzovara, A. & Vunder, J. Blocking human fear memory with the matrix metalloproteinase inhibitor doxycycline. *Molecular psychiatry* **23**, 1584-1589 (2018).
23. Novick, M.R. The axioms and principal results of classical test theory. *Journal of mathematical psychology* **3**, 1-18 (1966).
24. Lord, F.M. A strong true-score theory, with applications. *Psychometrika* **30**, 239-270 (1965).
25. Metzner, C., Mäki-Marttunen, T., Zurowski, B. & Steuber, V. Modules for Automated Validation and Comparison of Models of Neurophysiological and Neurocognitive Biomarkers of Psychiatric Disorders: ASSRUnit—A Case Study. *Computational Psychiatry* **2**, 74-91 (2018).
26. Rigdon, E.E., Sarstedt, M. & Becker, J.M. Quantify uncertainty in behavioral research. *Nat Hum Behav* **4**, 329-331 (2020).
27. Button, K.S.*, et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14**, 365-376 (2013).
28. Smaldino, P.E. & McElreath, R. The natural selection of bad science. *R Soc Open Sci* **3**, 160384 (2016).
29. Khemka, S., Tzovara, A., Gerster, S., Quednow, B.B. & Bach, D.R. Modeling startle eyeblink electromyogram to assess fear learning. *Psychophysiology* **54**, 204-214 (2017).
30. Bang, D. & Fleming, S.M. Distinct encoding of decision confidence in human medial prefrontal cortex. *Proc Natl Acad Sci U S A* **115**, 6082-6087 (2018).
31. Wager, T.D.*, et al.* An fMRI-based neurologic signature of physical pain. *N Engl J Med* **368**, 1388-1397 (2013).
32. Munafò, M.R.*, et al.* A manifesto for reproducible science. *Nature human behaviour* **1**, 0021 (2017).
33. Nosek, B.A., Ebersole, C.R., DeHaven, A.C. & Mellor, D.T. The preregistration revolution. *Proc Natl Acad Sci U S A* **115**, 2600-2606 (2018).

Competing interests

The authors declare no competing interests.

Acknowledgements

DRB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. ERC-2018 CoG-816564 ActionContraThreat). SMF is supported by a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (206648/Z/17/Z). The Wellcome Centre for Human Neuroimaging is funded by core funding from the Wellcome (203147/Z/16/Z).

Author contributions

All authors contributed to conception of the work. All authors wrote and revised the paper.

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

August 8, 2020

Contents

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 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:

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

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 $t_{ij} - \overline{t}_i$, 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:

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

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

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

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

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

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

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

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

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

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

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.

3 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.

4 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

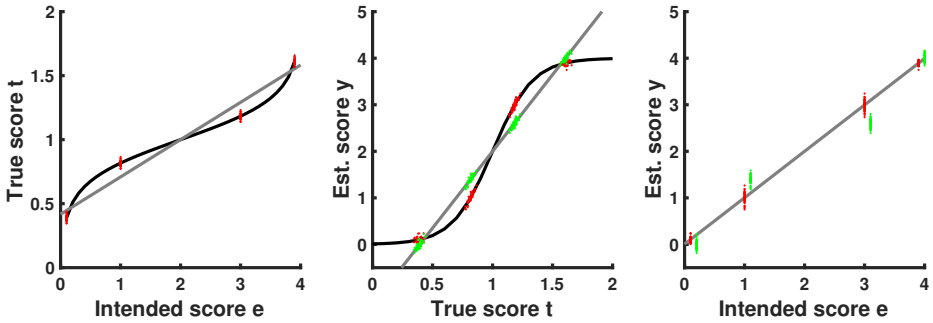

Figure 1: Scenario 4, validating a measure on itself. The systematic aberration in the experimental model is exactly the inverse of systematic error in the measurement model. The resulting estimated scores (red dots) are linearly related to intended scores. In comparison, a measurement model without 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.

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), 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.

5 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 infer-

ring 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.

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.