nature human behaviour

PERSPECTIVE

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

Calibrating the experimental measurement of psychological attributes

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

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.

When planning behavioural experiments, researchers must decide which observables to collect (observation) and how to preprocess 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 (for example, refs. [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 with 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 that 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 the reduction of measurement error 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.

[1]Wellcome Centre for Human Neuroimaging, University College London, London, UK. [2]Max Planck UCL Centre for Computational Psychiatry and Aging Research, University College London, London, UK. [3]Computational Psychiatry Research, Department of Psychiatry, Psychotherapy, and Psychosomatics, Psychiatric Hospital, University of Zurich, Zurich, Switzerland. [4]Department of Experimental Psychology, University College London, London, UK. [5]Psychological Research Methods, Humboldt University, Berlin, Germany. [6]Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany. ✉e-mail: d.bach@ucl.ac.uk

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 the 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 the 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 inter-individual 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 (Fig. 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 nonexhaustive list of further example applications across different subfields of psychology.

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 (preprocessing) 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 the 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: defining the measurand, identifying validity conditions and reporting their relationship.

Defining the measurand. The procedure that is used to create fear memory for calibration includes specifying the conditioned stimulus (CS; for example, triangles with specific size and colour), the unconditioned stimulus (US; for example, 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.

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 a 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 the 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.

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

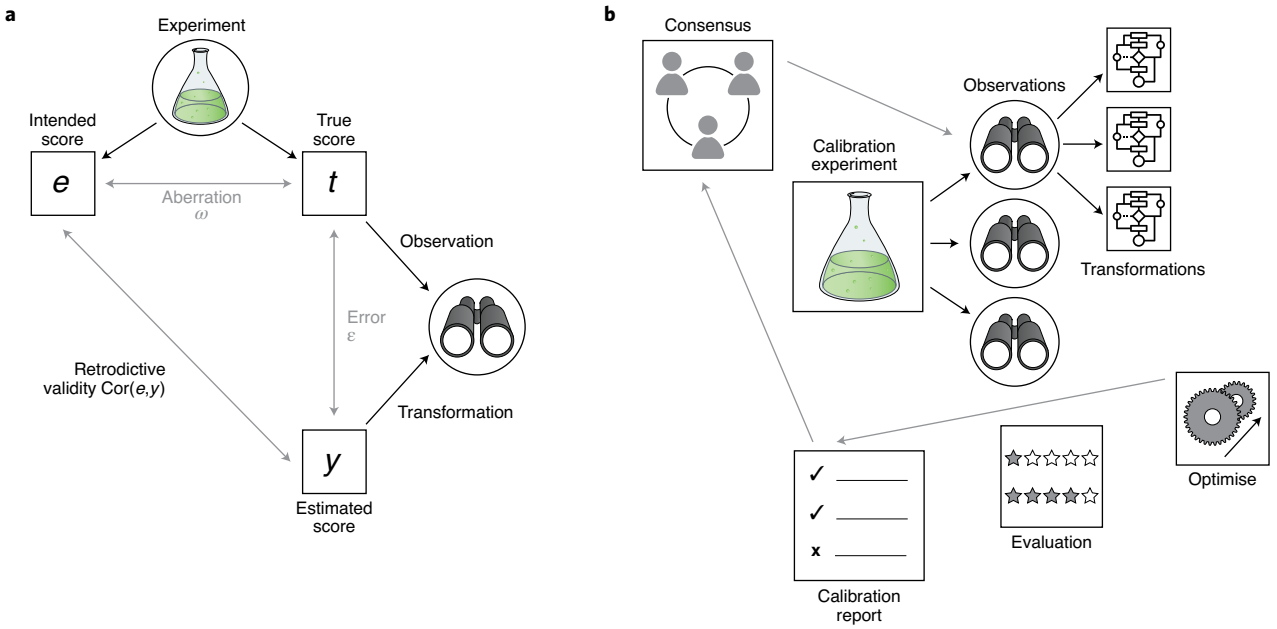

Fig. 1 | Retrodiction and calibration. a, A standardised experiment with intended attribute scores *e* generates true scores *t*. The measured attribute, *y*, is generated by transforming some observed data. Retrodictive validity denotes the observable correlation between *e* and *y* and is influenced 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 sub-fields of psychology for which calibration experiments appear feasible

Sub-field	Latent attribute	Possible calibration experiments	Specification of intended values per theory	Application outside calibration
Perception	Perceived stimulus property (for example, 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 three or more levels (associative learning theory)	Evaluation of learning interventions
Memory	Declarative memory	Number of repetitions in word lists	Interval scale with three 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

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 (Fig. 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

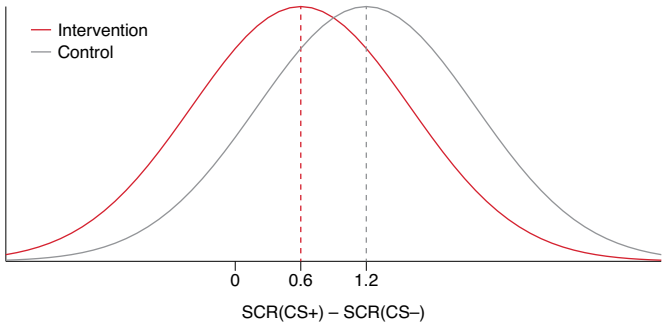

Fig. 2 | Power analysis. Measured scores *y* in a calibration experiment are affected by measurement error and experimental aberration. In this example, *y* is the difference between two skin conductance response (SCR) measurements and follows a standard normal distribution. The proposed experimental treatment is composed of the same manipulation as in the calibration experiment and either an additional intervention (red lines) or no intervention (control, grey 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 $P < 0.05$ in a one-tailed *t*-test.

within-participant CS+ vs 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 $P < 0.05$. 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.

Retrodictive validity and measurement accuracy

For a formal treatment, we now define key terms (see Fig. 1a for illustration and Supplementary Information for mathematical detail). As in classical test theory and other true score theories[23,24], we assume 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-participants) 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 nonlinearity in the mapping between these variables constitutes a misspecification of the intended values 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 (for example, 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 nonlinearity, a systematic (i.e., across participants) misspecification of *e*, which reduces the trueness of the experimental model. This is illustrated in Fig. 3a, where the black line denotes the actual nonlinear 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* (Fig. 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 Fig. 3a, which depict the true score differences under a constant value of *e*.

Next, we consider the mapping from *t* to *y* (Fig. 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 Information, we mathematically derive the conditions under which maximising the (observable) correlation between intended and estimated scores, $\mathrm{Cor}(e,y)$, 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 $\mathrm{Cor}(e,y)$ is guaranteed to increase measurement accuracy. Additionally, for any fixed measurement method, $\mathrm{Cor}(e,y)$ 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 Information, increasing $\mathrm{Cor}(e,y)$ may still increase measurement accuracy, but this is 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 it 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 these values 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

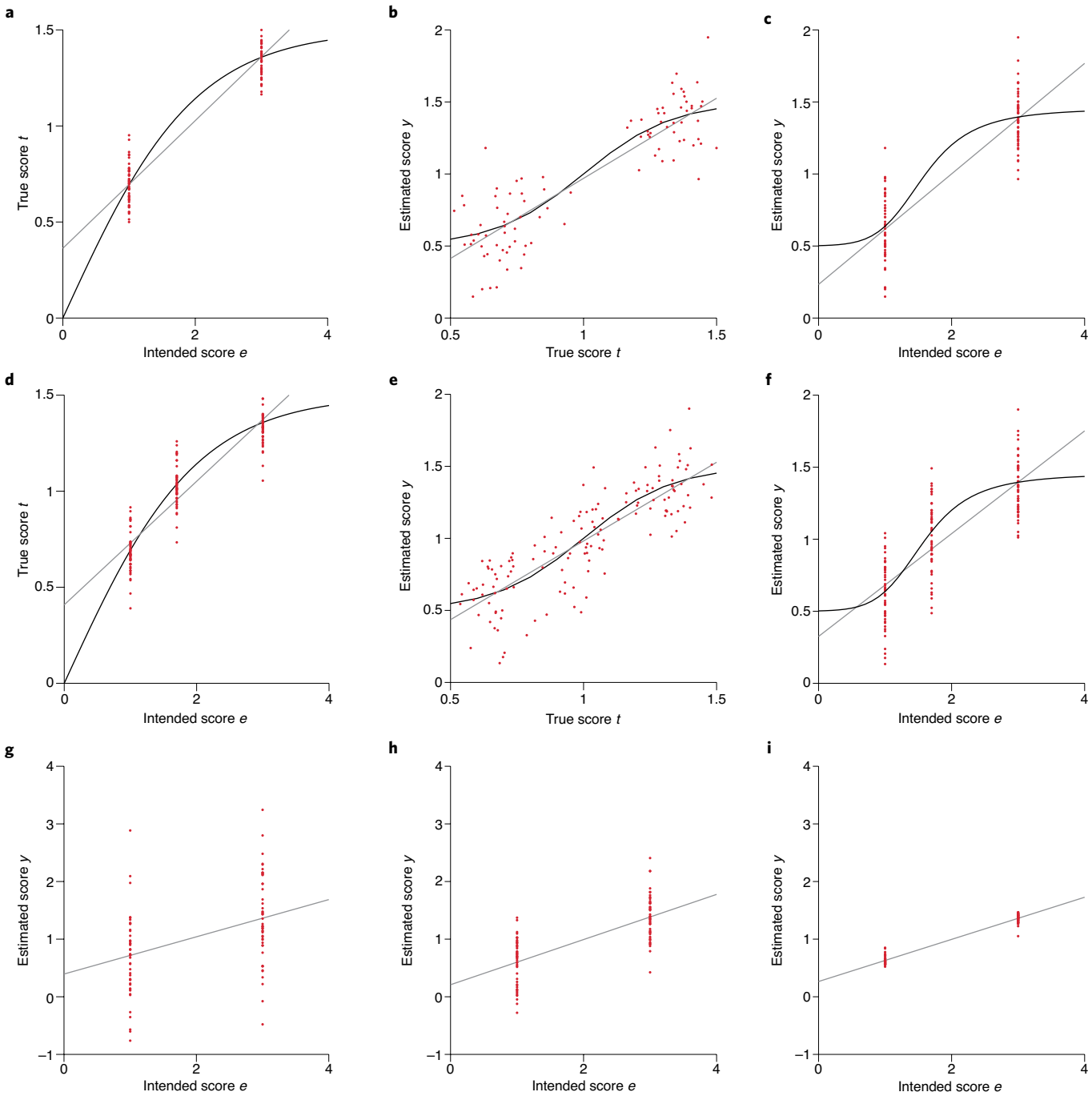

Fig. 3 | The retrodiction approach. a, The ideal relation between intended and true scores is a linear mapping with arbitrary coefficients (grey line), but the true relation is possibly nonlinear (systematic aberration, black line) and imprecise (distribution of red dots). Because there are only two 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 nonlinear (black line, composition of the two nonlinear functions in **a** and **b**), and imprecise (distribution of data points). **d**–**f**, Same model as in **a** and **b** 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**, $\varepsilon = 0$, but experimental aberration renders the resulting error non-zero.

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 the measurement method. Conditions known to impact 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 datasets. Often, the goal is to compare different measurement models (or transformation methods), which can be applied retrospectively to previously acquired datasets. To facilitate this in an iterative process (Fig. 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 dataset is published.

Further applications

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 (Fig. 2 and see section "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 a 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]. Finally, they can 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 it 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 (for example, 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 to 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 preprocessing 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 preprocessing before a study is conducted[32,33]. This leaves research practitioners in the uncomfortable situation of having to choose 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.

Received: 28 January 2020; Accepted: 21 September 2020;
Published online: 16 November 2020

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. *Adv. Methods Pract. Psychol. Sci.* **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**, e52465 (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. *Neurosci. Biobehav. Rev.* **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. *Psychol. Sci.* **22**, 1359–1366 (2011).
8. Lonsdorf, T. B., Merz, C. J. & Fullana, M. A. Fear extinction retention: is it what we think it is? *Biol. 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. in *Stevens' Handbook of Experimental Psychology* (ed. Pashler, H.) https://doi.org/10.1002/0471214426.pas0401 (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* **48**, 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. International Bureau of Weights and Measures (BIPM). The international vocabulary of metrology—basic and general concepts and associated terms (VIM). https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf (JCGM, 2012).
15. Shadish, W.R., Cook, T.D. & Campbell, D.T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (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. in *Test Validity* (eds. Wainer, H. & Braun, H. I.) 3–17 (Routledge, 1988).
18. van der Maas, H. L. J., 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. *Psychol. Rev.* **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* **55**, 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. *Mol. Psychiatry* **23**, 1584–1589 (2018).
23. Novick, M. R. The axioms and principal results of classical test theory. *J. Math. Psychol.* **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. *Comput. 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. USA* **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. *Nat. Hum. Behav.* **1**, 0021 (2017).
33. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci. USA* **115**, 2600–2606 (2018).

Acknowledgements

D.R.B. 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). S.M.F. 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 Trust (203147/Z/16/Z). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

D.R.B., F.M., S.M.F. and M.C.V. contributed to conception of the work. D.R.B. wrote and F.M. and M.C.V. contributed to the mathematical derivation. D.R.B., F.M., S.M.F. and M.C.V. wrote and revised the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-00976-8.

Correspondence should be addressed to D.R.B.

Peer review information Primary handling editor: Jamie Horder