An Open Publishing Platform of
the Organization for Human Brain Mapping
APERTURE NEURO

ORIGINAL RESEARCH ARTICLE

Putting behaviour back into brain–behaviour correlation analyses

Jeggan Tiego and Alex Fornito

Turner Institute for Brain and Mental Health, Monash Biomedical Imaging, and School of Psychological Sciences, Monash University, Clayton, Victoria, Australia

A fundamental challenge for human neuroscience is to relate imprecise measures of the brain with imprecise measures of behaviour. The recent study by Marek et al. (1) has demonstrated the perils of naively relating the two by showing that correlations between such measures are generally small and unstable under resampling unless very large samples (i.e., $N > 2,000$) are investigated. The importance of this seminal contribution cannot be overstated, as it comprehensively shows the limits of currently widespread approaches for brain-wide association studies (BWASs). Should we simply accept that small correlations are the norm, or can we improve our measurements in the hope of augmenting our effect sizes?

Correlations between two variables depend on the reliability and validity of the variables (2, 3). This basic fact suggests that effect sizes can be increased by improving the fidelity of our neuroimaging and/or behavioural measures. It is common knowledge that neuroimaging measures are noisy and indirect probes of the phenotypes of interest, being affected by multiple instrument, physiological, and person-specific artefacts (4–6). The development of improved data acquisition and processing strategies is a topic of intense investigation and will undoubtedly continue to yield increasingly precise measurements of brain structure and function in the future. Here, we draw attention to the fidelity of behavioural measurements, which has received comparatively little attention in the BWAS literature. We highlight four key psychometric considerations that suggest considerable gains can be made by adopting more refined methods for quantifying behaviour.

CONSIDERATION 1: GRANULARITY MISMATCH

Most experiments measure human behaviour using one or more (semi-) quantitative scales, such as cognitive test scores or psychological rating scales. The resulting scores are not a direct measurement of an objective property, such as height or weight, but instead reflect indirect estimates of a behavioural construct, such as intelligence or externalizing traits. The constructs indexed by commonly used scales are often defined without direct reference to brain function or neurobiological models; instead, they represent statistical abstractions from measures designed to achieve a specific purpose (e.g., identifying people with special educational needs or those with high levels of psychopathology). This can result in a granularity mismatch (7), such that the level of behavioural abstraction indexed by the scales does not correspond to the same level of abstraction in our neuroimaging measures. In more concrete terms, we should not expect a high correlation between a psychometric scale indexing, for example, total levels of psychopathology and the structure or function of any individual brain region (or pair of regions), since it is highly unlikely that any individual region strongly relates to such a complex summary of behaviour (an extreme form of biological reductionism) (8). Marek et al.'s (1) analysis indicates that the same rule holds for more specific, lower-level psychometric constructs as well (their Extended Data Fig. 1; although see Consideration 3, below).

It could be argued that granularity mismatch is a problem for mass univariate BWAS, and that multivariate analyses may offer greater power for linking brain and behaviour. If complex behaviour is a multifactorial, emergent property of brain–environment interactions and not the product of any individual brain region, surely it is related to combinations of different brain measures? (7, 8). Marek et al. (1) showed that multivariate analyses are still unstable for typical sample sizes of less than 100, but that reasonable out-of-sample associations ranging between 0.2 and 0.3 can be obtained between functional connectivity and cognition with canonical correlation analysis (CCA) for sample sizes between 300 and 1,000 (their Fig. S15). However, the considerable within-sample effect size inflation they observe (their Fig. S16) underscores the importance of out-of-sample cross-validation. The low correlations they observed in some cases between feature weights across model instances raise further concerns about model stability (their Fig. S13), as identified by others (9). These considerations notwithstanding, while multivariate models may improve effect sizes, we should be humble about the magnitude of brain–behaviour correlations that we can obtain if we measure behaviour using psychometric scales of complex constructs that are developed without reference to putative neural mechanisms.

CONSIDERATION 2: PHENOTYPIC RESOLUTION

For a measurement scale to accurately index an underlying construct, the scale must have adequate reliability across all levels of the latent dimension. In other words, the scale should measure both high and low levels of the trait with high reliability. If so, the scale is said to have a high phenotypic resolution (10). Phenotypic resolution can be examined using item response theory (IRT), a sophisticated approach to modelling item level data with respect to underlying constructs (11). In this context, reliability is not represented simply as a scalar value assigned to a given scale but represents a continuous function measured at each point across the latent trait continuum indexed by the scale.

To take a concrete example, Marek et al. (1) reported that the marginal reliability for one of the exemplar behavioural phenotypes from the child behaviour checklist (CBCL) was $r_{xx} = 0.94$ (p. 2). Supplementary analyses further suggested that, since the reliability of this measure is already near ceiling, any further improvements would not have an appreciable impact on brain–behaviour correlation magnitudes. However, for psychopathology traits, it is typical for measurement reliability to reach unacceptably low levels below the mean, where a large proportion of the sample score. This is because psychopathology constructs are unipolar, in that they have no adaptive end of the continuum or, at least, the scales designed to measure them do not have item content covering the adaptive end of the continuum (12).

We used IRT to calculate the reliability of each of the CBCL scales examined by Marek et al. (1) across the latent trait continuum (i.e., $M \pm 3SD$) in the 2-year follow-up data collection wave of the ABCD cohort (N = 5,820) (Figure 1A). We found that 10 of 11 scales (all scales except Total Problems) had reliability estimates below the minimum required for reliable research (i.e., $r_{xx} >.6$) (13) at $-1SD$ below the mean, with reliability reaching as low as 0.010 at $-3SD$ for the Withdrawn/Depressed and Rule-Breaking Behaviour scales. The proportion of the sample with unacceptable reliability (i.e., $r_{xx} <0.6$) based on latent trait continuum scores is, on average, 37.2% (5.9% Total Problems – 57.3% Rule-Breaking Behaviour). This result suggests that a large proportion of the data used for BWAS with these scales was noise, thereby attenuating correlations with neuroimaging phenotypes despite high-reliability estimates based on Cronbach's alpha (α = .68 to .95). This is a particularly pernicious problem when using scales developed in clinical contexts with normative samples.

CONSIDERATION 3: PHENOTYPIC COMPLEXITY

Phenotypic complexity is the extent to which a measured phenotype is multidimensional and reflects multiple sources of variance from one or more hierarchical levels of behaviour, extending from high generality (e.g., a general 'g' or 'p' factor) to high specificity (e.g., a specific cognitive or psychopathological construct). Failing to differentiate these sources of variance from each other can attenuate and otherwise confound relationships with neuroimaging results (3). We examined the phenotypic complexity of the CBCL data in the 2-year follow-up data collection wave of the ABCD cohort (N = 5,820) using a bifactor model, which parses variance into common (i.e., shared across subscales) and specific (i.e., unique to each subscale) sources (14). Almost half (48.8%) of the variance in the eight empirical syndrome scales was common and attributable to an overarching p-factor (15), suggesting a substantial contamination of variance that would obscure any specific sources of covariance with

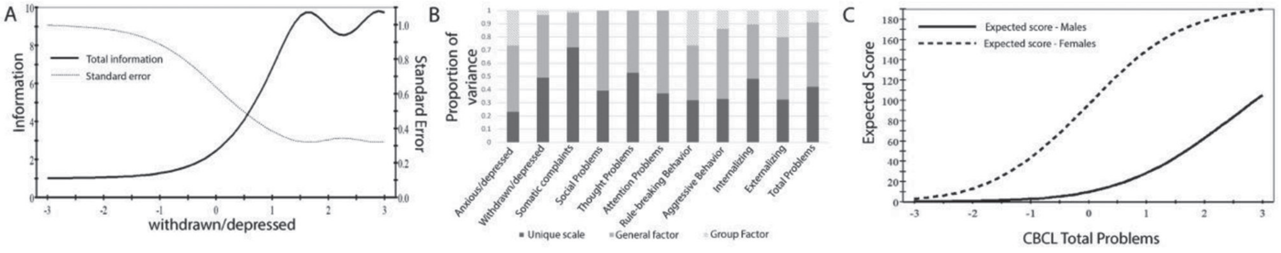

Fig. 1. **A:** Total information function for the child behaviour checklist (CBCL) withdrawn/depressed syndrome scale. Scale internal consistency reliability, r_{xx}, is related to the total information, I, as $r_{xx} = 1 - (1/I)$. **B:** Proportions of variance in CBCL scales attributable to a general p-factor, unique scale-specific variance, and each of the classical higher-order CBCL group factors (internalizing or externalizing). **C:** Test characteristic curves showing how the expected raw score (y axis) varies as a function of a participants' standing on the CBCL Total Problems latent trait continuum (x axis) for males (solid black line, n = 3,025) and females (dashed black line, n = 2,795). All analyses were performed on the 2-year follow-up data collection wave of the ABCD study cohort (N = 5,820).

neuroimaging measures (Figure 1B). Conversely, variance unique to each of the scales ranged from as little as 23.2% for Withdrawn/Depressed to 72.1% for Somatic Complaints (an average of 42.3% was observed across the eight scales; Figure 1B). If we consider the extraneous sources of variance from non-target phenotypes as noise, this phenotypic complexity results in substantial attenuation bias (16).

CONSIDERATION 4: MEASUREMENT NON-INVARIANCE

Measurement non-invariance refers to situations in which the measurement properties of a psychological or cognitive assessment instrument are not equivalent across subgroups within a sample (17, 18). This means that results are not directly comparable between groups because raw scores on the psychological or cognitive instrument have different substantive interpretations across groups. As an illustration, we examined measurement invariance between males and females for the CBCL Total Problems scale (the most reliable scale in the CBCL) with IRT using differential item function analyses, which is a powerful approach to detecting group differences in item measurement properties (19). The analysis yields test characteristic curves, which quantify the expected raw scores for each group as a function of their position on the underlying latent trait continuum, which represents a common metric for males (*n* = 3,025) and females (*n* = 2,795). These curves were not coincident at any point along the latent trait continuum, meaning that, for example, a raw score of 10 in males (equivalent to the mean of the latent trait) does not index the same level of severity in the underlying latent trait as it does in females (roughly equivalent to 2 standard deviations below the mean of the latent trait). These differences will confound any analysis that pools scores for males and females. Additional subgroups within the data (e.g., based on ethnicity, sociodemographic status, and the like) will compound this heterogeneity problem, which is particularly salient for large, heterogeneous cohorts (20).

WHERE TO FROM HERE?

These psychometric considerations are by no means an exhaustive list, and further issues are identified and discussed elsewhere (3, 16, 21–29). However, this brief discussion illustrates the complexity of measuring behaviour and of linking such measures to MRI phenotypes. It also suggests that substantial gains are possible by analysing behavioural phenotypes in a more refined way. In particular, latent variable modelling can be used in conjunction with IRT to derive more precise estimates of behavioural phenotypes and potentially augment effect sizes in BWAS. The degree to which such approaches can increase BWAS effect sizes is an empirical question, but one that we contend requires further investigation.

REFERENCES

1. Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S. et al. Reproducible brain-wide association studies require thousands of individuals. Nature 2022;603(7902):654–660.
2. Spearman, C. Correlation calculated from faulty data. Br J Psychol 1910:1904–1920.
3. Clark, L.A., Watson, D. Constructing validity: New developments in creating objective measuring instruments. Psychol Assess 2019;31:1412–1427.
4. Parkes, L., Fulcher, B., Yücel, M., Fornito, A. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. NeuroImage 2018;171:415–436.
5. Power, J.D., Plitt, M., Laumann, T.O., Martin, A. Sources and implications of whole-brain fMRI signals in humans. NeuroImage 2017;146:609–625.
6. Glasser, M.F., Coalson, T.S., Bijsterbosch, J.D., Harrison, S.J., Harms, M.P., Anticevic, A. et al. Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. NeuroImage 2018;181:692–717.
7. Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., Poeppel, D. Neuroscience needs behavior: Correcting a reductionist bias. Neuron 2017;93:480–490.
8. Borsboom, D., Cramer, A.O.J., Kalis, A. Brain disorders? Not really: Why network structures block reductionism in psychopathology research. Behav Brain Sci 2018;42:e2.
9. Tian, Y., Zalesky, A. Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? NeuroImage 2021;245:118648.
10. van der Sluis, S., Verhage, M., Posthuma, D., Dolan, C.V. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. PLOS ONE 2010;5:e13929.
11. Reise, S.P., Ainsworth, A.T., Haviland, M.G. Item response theory: Fundamentals, applications, and promise in psychological research. Cur Direct Psychol Sci 2005;14:95–101.
12. Reise, S.P., Waller, N.G. Item response theory and clinical measurement. Annu Rev Clin Psychol 2009;5:27–48.
13. Streiner, D.L. Starting at the beginning: An introduction to coefficient alpha and internal consistency. J Personality Assess 2003;80:99–103.
14. Reise, S.P. The rediscovery of bifactor measurement models. Multivariate Behavior Res 2021;47:667–696.
15. Caspi, A., Moffitt, T.E. All for one and one for all: Mental disorders in one dimension. Am J Psychiatr 2018;175(9):831–844.
16. Saccenti, E., Hendriks, M.H.W.B., Smilde, A.K. Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. Sci Rep 2020;10:438.
17. Vandenberg, R.J., Lance, C.E. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organ Res Methods 2000;3:4–70.
18. Meredith, W. Measurement invariance, factor-analysis and factorial invariance. Psychometrika 1993;58:525–543.
19. Stark, S., Chernyshenko, O.S., Drasgow, F. Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. J Appl Psychol 2006;91:1292–1306.
20. Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A.M., Nigg, J.T., Fair, D.A. The heterogeneity problem: Approaches to identify psychiatric subtypes. Trends Cognitive Sci 2019;23:584–601.
21. Patrick, C.J. Venables, N.C., Yancey, J.R., Hicks, B.M., Nelson, L.D., Kramer, M.D. A construct-network approach to bridging diagnostic and physiological domains: Application to assessment of externalizing psychopathology. J Abnormal Psychol 2013;122:902–916.
22. De Los Reyes, A., Thomas, S.A., Goodman, K.L., Kundey, S.M.A. Principles underlying the use of multiple informants' reports. Ann Rev Clin Psychol 2013;9:123–149.
23. Eid, M., Geiser, C., Koch, T. Measuring method effects: From traditional to design-oriented approaches. Cur Direct Psychol Sci 2016;25:275–280.
24. Podsakoff, P.M., MacKenzie, S.B., Podsakoff, N.P. Sources of method bias in social science research and recommendations on how to control it. Ann Rev Psychol 2012;63:539–569.

25. Feczko, E., Fair, D.A. Methods and challenges for assessing heterogeneity. Biol Psychiatr 2020;88(1):9–17.
26. Sanchez-Roige, S., Palmer, A.A. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. Nat Neurosci 2020;23:475–480.
27. Fried, E.I. Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. Front Psychol 2015;6:309.
28. Stanton, K., McDonnell, C.G., Hayden, E.P., Watson, D. Transdiagnostic approaches to psychopathology measurement: Recommendations for measure selection, data analysis, and participant recruitment. J Abnormal Psychol 2020;129:21–28.
29. Strauss, M.E., Smith, G.T. Construct validity: Advances in theory and methodology. Ann Rev Clin Psychol 2009;5:1–25.