Neuroscience and Biobehavioral Reviews 123 (2021) 72–82

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Back to Basics: The Importance of Measurement Properties in Biological Psychiatry

Daniel P. Moriarity *, Lauren B. Alloy

Temple University, United States

ARTICLE INFO

Keywords:
Biological psychiatry
measurement
methods
reliability
internal consistency
dimensionality

ABSTRACT

Biological psychiatry is a major funding priority for organizations that fund mental health research (e.g., National Institutes of Health). Despite this, some have argued that the field has fallen short of its considerable promise to meaningfully impact the classification, diagnosis, and treatment of psychopathology. This may be attributable in part to a paucity of research about key measurement properties ("physiometrics") of biological variables as they are commonly used in biological psychiatry research. Specifically, study designs informed by physiometrics are more likely to be replicable, avoid poor measurement that results in misestimation, and maximize efficiency in terms of time, money, and the number of analyses conducted. This review describes five key physiometric principles (internal consistency, dimensionality, method-specific variance, temporal stability, and temporal specificity), illustrates how lack of understanding about these characteristics imposes meaningful limitations on research, and reviews examples of physiometric studies featuring a variety of popular biological variables to illustrate how this research can be done and substantive conclusions drawn about the variables of interest.

1. Introduction

The integration of biological and psychopathological research into the field of biological psychiatry is prioritized highly at the National Institutes of Health. Whereas there is substantial discussion and standard reporting of certain types of measurement properties (e.g., dimensionality, retest reliability) for self-report questionnaires, less work has been done to investigate these measurement features for many relevant biological constructs and they are less frequently reported (Hajcak and Patrick, 2015). This is not to say that there has not been important investigation and regular reporting of measurement properties specific to biological variables (e.g., intra-assay coefficients of variation). Rather, several metrics key to common methodological and statistical practices in psychiatry research have not received comparable attention for biological variables. This may be due to greater confidence in the measurement of that which is directly observable (e.g., concentrations of analytes in blood). However, the ease with which a construct is operationally defined and measured does not directly translate to measurement qualities suitable for common statistical approaches.

It is important to remember Cronbach and Meehl's (1955) admonition, "One does not validate a test, but only a principle for making inferences" (p. 297). Confidence that a test can measure a variable accurately is not sufficient to know that the test facilitates the inferences tested in statistical models. For that, there is need for a thorough analysis of measurement properties germane to the intended data collection and statistical procedures. Armed with information about key measurement properties (henceforth referred to as "physiometrics"; Segerstrom & Smith, 2012), researchers can design more cost-effective and well-powered studies that are better indicators of the true associations between variables of interest.

2. The Perils of a Paucity of Physiometric Research

Variables with poor or unknown physiometrics impose multiple limitations to meaningful research. Thus, to ensure that biological psychiatry research reaches its maximum potential utility, it is important to evaluate measurement qualities key to typical methods used in biological psychiatry research to determine what study designs and analytic techniques are best suited to various biomarkers. In this section, we outline some of the risks and constraints imposed by research using variables with poor or unknown measurement properties.

* Corresponding author.
E-mail address: Daniel.moriarity@temple.edu (D.P. Moriarity).

https://doi.org/10.1016/j.neubiorev.2021.01.008
Received 27 July 2020; Received in revised form 12 January 2021; Accepted 16 January 2021
Available online 23 January 2021
0149-7634/© 2021 Elsevier Ltd. All rights reserved.

2.1. Internal Consistency

Many theories in biological psychiatry are about multifaceted biological constructs (e.g., reward processing, inflammation, etc.); however, studies commonly test multiple individual indices of these larger constructs (Segerstrom and Smith, 2012). Given concerns about the reliability of single-item measures and issues with multiple statistical comparisons, increased use of composite biological variables might benefit replicability in biological psychiatry. When used thoughtfully, composite measures also have the benefit of accentuating variance shared between components and reducing the impact of measurement error. When using composite measures, it is important to report internal consistency, which indicates the level of shared variance between component variables ("true score") relative to unshared ("error") variance (Cortina, 1993). Typically, researchers have hypotheses about the relationship between two constructs (e.g., inflammation and depression); consequently, it is beneficial to maximize the "true score" of their constructs of interest. Although reporting internal consistency for self-report questionnaires is standard practice, it is infrequently reported for applicable biological variables. For example, internal consistency is reported inconsistently for measures involving the creation of a single score from several trials of a task (e.g., error related negativity (ERN)), despite providing insight regarding consistent performance across the task and having implications for effect size (Hajcak et al., 2017). *Thus, whenever aggregate variables are used, it is important to report a measure of internal consistency (e.g., Cronbach's α, coefficient Ω).*

2.2. Dimensionality

Another important consideration when working with aggregate measures is the concept of dimensionality. Dimensionality refers to the degree to which a set of variables indicates the presence of one or more higher-order constructs. For example, under traditional conceptualizations of psychopathology, all behaviors on a depression questionnaire are associated with the construct of depression. Similarly, an assortment of biological variables (e.g., different proinflammatory proteins) could serve as markers of a higher-order construct (e.g., inflammation). It also is important to consider potential construct heterogeneity, the possibility that several lower-order constructs (e.g., pro- and anti-inflammatory processes) might comprise a larger construct of interest (e.g., inflammation).

Empirical evaluation of dimensionality is possible with dimension reduction techniques such as exploratory factor analysis (EFA) and principal components analysis (PCA). Both approaches investigate the structure of data with the logic that if all component variables are indicators of the same process, they should be strongly associated with one another (i.e., have high internal consistency, Clark & Watson, 1995, 2019; Loevinger, 1957). As such, dimension reduction approaches can help identify whether sets of variables are unidimensional or multidimensional in nature as well as components that might not load onto any of these processes (Tabachnick and Fidell, 2013). The primary theoretical distinction between the two is that the dimensions found in EFA are theorized to cause the variables, whereas the dimensions found in PCA are simply aggregates of observed variables. Statistically, only shared variance is analyzed in an EFA, but all variance is analyzed in a PCA.

Modeling decisions uninformed by dimensionality can have negative implications. Aggregating unrelated components into a single dimension or indicator reduces internal consistency and, consequently, the maximum observable true effect size (Hajcak et al., 2017). Relatedly, if only some dimensions/indicators are related to a criterion of interest, aggregating them with unrelated variables might wash out true effects. Alternatively, falsely assuming multidimensionality reduces power via failure to aggregate shared variance of interest. Further, it introduces issues with multiple comparisons.

However, these techniques are not appropriate for all datasets. It is important to consider that the maximum number of dimensions is constricted by the number of indicator variables tested. In other words, there needs to be enough variables per dimension to statistically anchor each dimension. Further, datasets with lower numbers of variables, higher dimensionality, and weaker associations between the variables and the dimensions require higher sample sizes to produce stable results (Guadagnoli and Velicer, 1988). Additionally, it is ill-advised to draw conclusions about dimensionality without thoughtful consideration of biological plausibility. *Consequently, it is important to consider dimensionality when multiple indicators of a broader construct of interest are collected before proceeding with hypothesis testing involving that construct. However, modeling decisions should be informed both by empirical investigation (if appropriate in the context of the dataset used) and biological plausibility.*

2.3. Method-specific Variance

Although not a "metric" in the sense of something explicitly testable and reportable like the other characteristics reviewed here, a critical measurement issue for biological psychiatry is method-specific variance. In addition to the "random" variance that contributes to measurement error, there is variability associated with the specific method of measurement (e.g., self-report, behavioral, psychophysiological) that is unrelated to the true construct of interest (Patrick et al., 2013). Consequently, two measures of the same construct using different methods will have smaller associations compared to two measures using similar modalities (e.g., self-report correlated with biological vs. self-report correlated with self-report). Given that biological psychiatry is, by definition, a multimodal field, this is a pervasive issue that needs to be considered when designing studies and interpreting results. *Thus, method-specific variance should be considered for all studies including multiple measurement modalities. This issue should inform power analyses, measurement error-adjusted analytic techniques, and consideration of aggregating multimethod assessments of the same construct. For a more detailed review of this issue and strategies to address it, see* Patrick et al. (2019).

2.4. Temporal Stability

Whereas a measure given to multiple people at a single time point has two sources of variance (between-person differences and measurement error), a measure given multiple times introduces a third source of variability: within-person variance. Measures with low within-person variability (small changes over time) have high temporal stability. Temporal stability is most frequently quantified using retest Pearson correlations (correlating scores on a measure at two different time points) and intraclass correlation coefficients (ICCs, which quantify the proportion of stable between-person differences across multiple time points). It is standard practice to report (or at least cite other work about) the temporal stability of self-report measures, but it is reported less consistently for biological variables (e.g., Moriarity et al., 2020b). This is concerning, given that information about temporal stability is necessary to interpret the probability with which a score at baseline will be similar to the score at follow-up. It is important to note that highly stable measures are not always the goal; many biological constructs would be expected to have both trait (relatively stable) and state (varying across time and situational factors) components. Target temporal stability should be informed by the conceptual stability of the construct in question (e.g., few would expect mood to be 100% stable in a community sample over the course of a year). *Temporal stability should be reported for all longitudinal studies. It should be calculated in the sample when repeated measures are available, or estimates reported from existing studies when calculation within the sample is impossible.*

2.5. Temporal Specificity

Somewhat related is the concept of temporal specificity.

Longitudinal data are necessary to establish directionality of associations; however, time between data points is an important methodological consideration. For example, the relationship between eating a hot pepper and experiencing pain after a couple minutes would not be as strong days after the meal. Thus, exploratory analyses are necessary to evaluate how the relationships between variables might fluctuate as a function of time (including potential developmental considerations). Temporally-informed study designs could improve replicability, provide information about when changes in biological risk factors manifest behaviorally (and vice-versa), and inform treatment studies given expected delays between interventions and symptom reduction (e.g., anti-inflammatory treatments for depression). *Thus, the field would benefit from more exploratory studies investigating the temporal specificity of associations of interest to identify optimal time lags between measurements.*

2.6. Effect Size and Power

The practical implications of many biological psychiatry studies are often questioned because they frequently have small effect sizes, which could be directly impacted by the use of measures uninformed by their physiometrics (such as those reviewed above). To illustrate, consider the formula for the maximum observable true correlation between two variables as a function of their reliability: $r_{xy}(\max) = \sqrt{(r_{xx}r_{yy})}$ where r_{xy} represents the maximum observable true correlation between variables x and y, r_{xx} represents the reliability of variable x, and r_{yy} represents the reliability of variable y (Davidshofer and Murphy, 2005). Only if two measures are perfectly reliable (both r_{xx} and $r_{yy} = 1$) can the maximum correlation = 1. As reliability decreases, so does the maximum observable true correlation. Consider two research teams testing the same hypothesis and using the same measure for variable x ($r_{xx} = .70$), but different measures for variable y ($r_{yy} = .70$ for Team A but $r_{yy} = .30$ for Team B). The maximum observable true correlation is .70 for Team A, but only .46 for Team B. Similar results have been found concerning the relationship between internal consistency and effect sizes (Hajcak et al., 2017).

This penalty is magnified in more complex designs. For example, many variables in biological psychiatry (e.g., inflammation) are theorized to be mediators between stress and psychopathology (e.g., Moriarity et al., 2018; Slavich and Irwin, 2014). Mediation analyses involve calculating the product of the association between i) the focal predictor and the mediator (a' pathway) and ii) the mediator and the outcome variable (b' pathway). Thus, unreliability of the mediator will result in misestimation of both estimates. Consequently, the bias introduced by poor reliability is effectively squared when calculating their product.

This bias also exists for group comparisons, which often occur in biological psychiatry in the form of case-control studies (e.g., Ng et al., 2019). The test statistics for these analyses (independent samples *t*-tests and between-subjects ANOVAs) are a ratio of the magnitude of the group difference divided by a variance component. Poor reliability inflates variability, decreasing the maximum observable true effect. For example, consider a researcher using an independent samples *t*-test to compare levels of interleukin (IL)-6 between participants with Major Depressive Disorder (MDD) and non-depressed controls. The formula for an independent samples t-test is $t = \frac{M1-M2}{SE}$. Suppose the true difference in IL-6 for individuals with MDD vs. non-depressed controls (M_1-M_2) is .30. In scenario A, the standard error of this difference (*SE*) is .15, and the *t*-score will = 2. The critical value that the *t*-score must be above to be significant at $p < .05$ is 1.96, so the researchers have a significant result. Now imagine scenario B, in which the group difference is the same, but the *SE* of this difference increases to .2 because of less reliable IL-6 measurement. Now the *t*-score is 1.5, which is not significant, despite having the same observed difference between the groups. The same logic applies for standardized (but not unstandardized) measures of effect size (e.g., Cohen's $d = \frac{M1-M2}{SD_{pooled}}$). Given the same difference between two means, as the standard deviation increases, *d* decreases. However, this does not mean that measurement error always results in attenuated effect sizes. Although it is true that the median standardized effect size will be lower when estimated with vs. without error, random error variance also can result in over-estimates (Segerstrom and Boggero, 2020), leading to false positives that could inspire misguided studies and intervention efforts. *Thus, inflated variability caused by unreliable measures can cause true effects to be overlooked both in terms of probability under null-hypothesis testing as well as their substantive implications via standardized effect sizes. Unreliable measures can also result in false positives and artificially inflated effect sizes.* Given the importance of individual differences research in the Research Domain Criteria (RDoC; Cuthbert & Kozak, 2013) initiative, this is a key (and addressable) source of bias in popular analytic strategies for NIH-funded research.

3. Examples of Physiometric Research in Biological Psychiatry

Below, several examples of physiometric research investigating a variety of biological variables are reviewed to illustrate the techniques used and conclusions about the variables of interest.

3.1. Internal Consistency

As previously discussed, strong internal consistency is evidence that various components of a measure are responded to similarly. To illustrate the importance of investigating internal consistency for neural measures, Hajcak et al. (2017) evaluated error-related negativity (ERN) averaged across multiple trials as a function of the number of trials completed by participants in two groups (with and without generalized anxiety disorder). The study reported two measures of internal consistency: Cronbach's α (how representative one trial was of all trials) and split-half reliability (correlating the average scores from the odd and even trials). They found that α increased sharply between four and eight trials, and modestly until approximately fourteen trials, after which α only increased subtly. Cronbach's α reached a maximum of .75 - .85, which was comparable to the Spearman-Brown corrected split-half reliability (r_{sb} = .71-.75). The lack of reliability when fewer trials were included is an expected feature of Cronbach's α, and dovetails with concerns about the reliability of single-item/few-item indicators. Further, the diminishing returns of increased trials reflects that more trials only decreases random error, not systematic error (e.g., error introduced by data collection techniques). In fact, there is a mathematically quadratic relationship between the number of indicators in a composite and the Spearman-Brown reliability such that, with enough indicators, nearly perfect reliability is achievable regardless of the true, systematic error. These results can help researchers plan the ideal number of trials to minimize participant burden without resulting in data with subpar measurement qualities and, consequently, limited utility. Additionally, they highlight one way of comparing different methods of data collection. For example, comparing the trajectories and plateaus of internal consistency as number of trials increases could provide insight on ratios of random vs. systematic error for two different ERN measures.

Kaye, Bradford, and Curtin (2016) present a thorough investigation of several measurement qualities (internal consistency, temporal stability, and effect size stability, the latter two will be discussed later) of acoustic startle (defensive reflex in response to brief, startling noise probes) and corrugator responses (reaction of the corrugator muscle associated with frowning) during a no-shock/predictable shock/unpredictable shock (NPU) task, an affective picture viewing task, and resting state task over two study visits (approximately one-week apart). Specifically, they evaluated Spearman-Brown corrected split-half reliability between odd and even trials as a measure of internal consistency. Further, the authors compared performance of within-person standardized ((raw score for a trial minus the participant's mean across all trials)/participant's standard deviation across all trials) vs. unstandardized scores for startle potentiation and the time domain and

frequency domain for corrugator potentiation. For the sake of brevity, this review will focus on startle potentiation. For the NPU task, the internal consistency for raw scores was higher than standardized scores for both predictable and unpredictable startle responses, with scores ranging from good to adequate (r_{sb} = .81, .64, .57, .52, respectively). For the affective picture viewing task, internal consistency for startle modulation was poor for all scores, but standardized scores were better for pleasant, and raw scores were better for unpleasant, startle modulation (raw pleasant $r_{sb} < .00$, standardized pleasant $r_{sb} = .16$, raw unpleasant $r_{sb} = .14$, standardized unpleasant $r_{sb} < .00$). Because within-subject standardized scores would have no utility for the resting state task, only internal consistency was reported for raw scores ($r_{sb} = .95$). Recalling the sources of variance (between-person, within-person, and error), it is unsurprising that raw scores typically had higher internal consistency than within-person standardized scores because true between-person variance was removed from the latter. In addition to their descriptive value, comparison of different types of responses and the influence of within-person standardization across several tasks is informative for the establishment of best-practices for these behavioral tasks.

Given the rise in popularity and high cost of functional magnetic resonance imaging (fMRI) in biological psychiatry, investigation of measurement properties of these methods is crucial. Luking et al. (2017) evaluated the split-half internal consistency for ERPs and blood oxygen level-dependent (BOLD) responses to monetary gain and loss feedback (an fMRI measure) within the ventral striatum and medial and/or lateral prefrontal cortex using Spearman-Brown corrected split-half reliability (comparing odd/even trials). Similar to Kaye et al. (2016), they compared several scoring methods: raw scores, difference scores (gain – loss), and residual scores (gain controlling for loss). Raw BOLD responses across all regions and ERPs to both gain and loss feedback demonstrated high internal consistency ($.66 \geq r_{sb} \geq .86$). Raw scores had consistently higher internal consistency than residual scores ($.26 \geq r_{sb} \geq .50$), which had uniformly higher internal consistency than difference scores ($.02 \geq r_{sb} \geq .36$). Thus, although residual scores may not have ideal internal consistency, they might be preferable over subtraction-based difference scores for studying between-person differences in within-person processes with these measures.

Instead of concluding that difference scores (common in many areas of biological psychiatry) are universally unreliable, it is important to consider *why* reliability was lowest for the difference scores, and under what context difference scores have utility. First, when variance associated with one variable is removed from another (either via subtraction or creating a residual term), the variance removed will be from the reliable variance because it is impossible for two variables to share *random* error. This reduction in reliability is greater when the two raw variables are highly correlated (Thomas and Zumbo, 2012). However, as emphasized in the discussion of temporal stability above, reliability needs to be considered in light of the expected true reliability. For reasons beyond the technical scope of this review (see Rogosa and Willett, 1983), when the individual differences in the difference score are not at least moderate, the reliability of the difference score will be more similar to the reliability of the raw scores. There also is evidence that BOLD difference scores that contrast win and loss conditions vs. neutral, instead of comparing win to loss conditions, can result in more reliable estimates (Holiga et al., 2018; Plichta et al., 2012), but the appropriateness of this approach depends on the research question at hand. Alternatively, many have argued that polynomial regression is a preferable technique to using difference scores altogether (Edwards, 2001).

It is important to note that residual/difference scores also hold the potential to isolate theoretically relevant variance in certain designs. For example, consider a study that compared P3 amplitudes (an event related potential) to aversive vs. neutral stimuli (used to index general reactivity) as predictors of threat sensitivity, finding the split-half reliability excellent for both conditions (r_{sb} = .92 and .90, respectively; Perkins et al., 2017). Split-half reliability for the difference between the two conditions (aversive-neutral) was poor (r_{sb} = .29). Recalling that variance removed when creating a difference score always comes from true variability, never random error, this decrease in reliability is not a surprise. Both the absolute value of the correlation between the difference score and threat sensitivity (r = -.12) and the correlation between general reactivity and threat sensitivity (r = .16) were small. However, a larger proportion of the systematic variance (true score) in the difference score was associated with threat sensitivity (i.e., $(-.12^2/.29) * 100 = 5.00\%$) compared to general reactivity (i.e., $(.16^2/.92) * 100 = 2.78\%$). This approach was particularly important when considering that the association between general reactivity and threat sensitivity was positive, but that the association between the variance unique to the aversive condition and threat sensitivity was negative. Thus, the variance from general reactivity could washout the association unique to the aversive condition if it were not removed from the variable. Consequently, it is important to consider how variables with modest reliability, but that include substantial amounts of criterion-related variance, can be informative.

3.2. Dimensionality

Recall the example of inflammation as a complex construct often indexed by several indicators (Segerstrom and Smith, 2012). One study of atherosclerosis (Egnot et al., 2018) assessed the dimensionality of seven inflammatory proteins and coagulation biomarkers (specifically, CRP, IL-6, fibrinogen, Lp(a), slCAM-1, PTX-3, and D-dimer) in a sample of 1103 adults. Thus, the sample was well-powered and there were enough indicators to find a one- or two-dimensional structure. The results of the EFA found a two-factor solution: Factor 1 consisted of CRP, IL-6, and fibrinogen; Factor 2 consisted of D-dimer and PTX-3, whereas slCAM-1 and Lp(a) did not load on either factor. Factor 1 was interpreted to represent a non-specific inflammatory process, whereas Factor 2 was interpreted to indicate coagulation burden. The authors then tested the factors as predictors of several outcomes, finding some associations unique to only one of the two factors. For example, although both factors were positively associated with risk for low ankle brachial index, higher levels of coagulation burden (Factor 2), but not inflammation (Factor 1), were associated with elevated common femoral artery intima-media thickness, suggesting that coagulation burden might be a better indicator of subclinical peripheral artery disease than inflammation.

Independent component analysis (ICA) is a technique for investigating dimensionality primarily used with neuroimaging and EEG data. Kakeda et al. (2020) used ICA as a data-driven approach to identify brain regions that might differ in grey matter volume between individuals with depression (n = 45) and controls (n = 38), and whether the volume in these regions correlated with serum TNFα. Specifically, they used source-based morphometry (which applies an ICA to a segmented image) to arrange the voxels into common morphological features of grey matter concentration among participants. Results indicated fourteen independent structural components; however, based on previous work (Williams, 2016), Kakeda et al. excluded four primarily cerebellar networks. Of the ten remaining components, two (a prefrontal network and an insula-temporal network) had less grey matter volume in a group of participants with depression compared to controls. Of these two, serum TNFα was significantly negatively correlated with the prefrontal network, but was not significantly correlated with the insula-temporal network. It is important to note (as the authors themselves do) that this study was limited by a small sample size, which constrains the number of components ICA can extract (Li et al., 2007), similar to how the number of indicators limits how many factors can be found using EFA.

3.3. Method-specific Variance

As described earlier, a major obstacle for biological psychiatry research is domain-specific method variance, the systematic tendency

for two measures of the same construct using different modalities (e.g., self-report vs. biological vs. behavioral) to have smaller associations than two measures using the same modality. Ostensibly, one reason for this is that measures from disparate modalities each contribute unique method-specific error (variance related to the measurement method and unrelated to the construct of interest; Patrick et al., 2013). This suggests that the integration of indices of a construct across multiple methods of measurement into single variables, described as the "cross-domain approach" (Patrick et al., 2013; Venables et al., 2018), might accentuate the shared variance related to the construct of interest, improving utility and construct validity.

To illustrate this, Nelson et al. (2011) measured three event-related potential (ERP) measures (ERN and P3 response to target stimuli from a flanker task and P3 response to feedback stimuli from a gambling feedback task) and investigated a) whether these measures represent overlapping indicators of externalizing proneness, and b) whether they index a shared neural process that accounts for their individual associations with externalizing proneness. Results of an EFA suggested that a single factor accounted for the covariance among all three variables, and that all three variables contributed similarly to this shared factor. To evaluate whether this factor represented brain processes associated with externalizing proneness, Nelson et al. (2011) ran another EFA including the three ERP measures as well as a self-report measure of externalizing proneness, again finding a single factor. Results of analyses using the aggregated ERP factor found that the aggregate measure had stronger correlations with the majority of physiological and psychometric externalizing proneness criterion variables tested than did the individual ERP measures. In fact, the composite factor out-performed comparison ERP measures (not included in the composite) in predicting externalizing proneness, likely due to the composite variable accentuating the shared externalizing proneness-related variance in the individual ERP variables. However, as described above (and discussed by the authors), a factor analysis on three ERP components and a self-report measure is not enough to provide a convincing evaluation of the true structure of these measures or provide enough options to support alternative models. In other words, there were not enough components to anchor more than one factor, so the factor analytic solution could, at most, feature one aggregate measure and/or unrelated variables. Still, this study serves as an example of how variable aggregation can result in variables with stronger predictive validity than the component parts.

To extend this work, Venables et al. (2018) first ran EFAs on several indices of inhibition-disinhibition within specific measurement domains (self-report, behavioral performance, brain response). Consistent with the ERP study above, indices within discrete measurement domains revealed single factor solutions. All possible pairwise correlations between these three domain factors were significantly positively correlated. Next, two confirmatory factor analyses (CFA) were estimated: the first specifying all indices across the three measurement domains loading onto a single factor, and the second specifying three lower order factors corresponding with each measurement method that, in turn, load onto a higher order *cross-domain* factor. The former demonstrated poor model fit, but the cross-domain factor model fit the data well. Further, comparative fit indices found significant differences in model fit, suggesting that inhibition-disinhibition is best represented by a cross-measurement domain, hierarchical factor structure. Additionally, the cross-domain factor frequently demonstrated significant correlations with the vast majority of criterion variables tested, whereas measurement-domain specific scores were less likely to be correlated with criterion variables from other measurement domains. Thus, these results demonstrate how thoughtful investigation of dimensionality in biological psychiatry can improve the construct validity of variables by the creation of cross-measurement domain composites that ameliorate concerns about a) the reliability of single-item measures (which are common in biological psychiatry) and b) downward-biased estimates due to measurement domain-specific variability.

3.4. Temporal Stability

Out of all the physiometric characteristics described above, biological psychiatry probably has done the best with assessing and reporting temporal stability (the reliability of a measure between different time points). However, there are many constructs of interest for which there is a paucity of research on this topic, especially when considering the wide breadth of study durations seen in behavioral health research. Before reviewing some examples of temporal stability research in biological psychiatry, it is important to emphasize that temporal stability estimates are only informative for the duration in which they are studied. Unfortunately, across all disciplines of behavioral health research, it is commonplace for previous work to be cited as evidence that a measure has sound temporal stability with no reference to the duration for which the measure's stability originally was assessed. Further, it also is essential to reiterate that having low temporal stability is not always indicative of a poor measure. The temporal stability of a measure is dependent on, and constrained by, stability of the construct under question. If one evaluated the 6-month temporal stability of depressed mood and height in a sample of adults, one would expect height to be more stable. Other contextual concerns, such as age, also are important to consider. For example, one would expect relatively lower 6-month temporal stability of height in a sample of 10-year-olds than a sample of adults. Finally, temporal stability, like many of the other measurement properties described in this review, can be misestimated due to unreliable measures.

The most straightforward metric of temporal stability is retest reliability using Pearson's *r*, the correlation between a measure at two different time points. In addition to internal consistency metrics, Kaye et al. (2016) (described above) also investigated one-week temporal stability of startle and corrugator responses to three tasks (NPU, affective picture viewing, and resting state) comparing raw vs. within-person standardized scores (Bradford et al., 2015) as well as differences in the effect size of task manipulations (predictable and unpredictable potentiation for the NPU task and pleasant and unpleasant modulation for the affective picture viewing task) between the two sessions. Similar to above, this review only will cover startle responses for the sake of brevity.

Temporal stability was higher for raw scores for both predictable and unpredictable startle potentiation during the NPU task (both $r = .71$) compared to within-person standardized scores ($r = .58$ and .49, respectively). When comparing the effect size of NPU manipulations between study visits, no significant differences were observed for raw or standardized predictable startle potentiation and raw unpredictable startle potentiation (all $\acute{\eta}_p^2 = .001$-.033, $p > .05$), but the standardized startle potentiation was smaller at the second visit ($\acute{\eta}_p^2 = .04$, $p = .03$), suggesting that the manipulation lost potency over time. Regarding the affective picture viewing task, one-week temporal stability was poor for both raw and standardized scores for pleasant startle modulation ($r <$ .00 and $= .08$, respectively), but was higher for the unpleasant startle modulation ($r = .50$ for raw, $r = .40$ for standardized). The effect sizes for the raw pleasant and unpleasant startle modulations were not significantly different after one week ($\acute{\eta}_p^2 = .02$, $p = .10$; $\acute{\eta}_p^2 = 03$; $p = .09$, respectively). It is interesting to note that the effect sizes for the standardized pleasant and unpleasant startle modulations differed between testing sessions ($\acute{\eta}_p^2 = .05$, $p = .02$; $\acute{\eta}_p^2 = .10$, $p < .001$, respectively), but in opposite directions (Visit 2 was smaller for pleasant startle modulation, but larger for unpleasant). As mentioned above, standardized scores for the resting state task have no utility, but the raw scores had high one-week temporal stability ($r = .89$) and scores were smaller at Visit 2 ($\acute{\eta}_p^2 = .21$, $p < .001$, respectively). There was no manipulation during (and consequently, no effect size for) the resting state task. In sum, these results demonstrate how different analytic approaches (i.e., raw vs. within-person standardized scores) can influence important temporal dynamics of behavioral tasks such as stability and the potency of the manipulation, which have important implications for designing

and interpreting research using repeated measures of these tasks.

Temporal stability also can be influenced by how extreme values are handled, as evidenced by Landau et al. (2019), a study investigating salivary CRP. Immunoassays use standard concentrations of an analyte to generate a standard curve, on which sample values are interpolated. Many samples have values that are flagged by the procedure as too high or low to fit onto the standard curve. In "strict" standard curve datasets, these extreme values are excluded; in "relaxed" standard curve datasets, they are extrapolated outside the standard curve range. There are several techniques currently used to handle these values: list-wise deletion, pairwise deletion, multiple imputation (extreme values replaced with multiply imputed values), and winsorization (extreme values replaced with the most extreme value on the standard curve). Landau et al. (2019) applied each of these four techniques to a strict and a relaxed dataset, resulting in eight total datasets. Additionally, they compared the reliability of samples taken in the morning compared to the evening, given evidence of diurnal variation in CRP (Out et al., 2012). The average two-day Pearson *r* was .49 for morning samples and .60 for evening samples, suggesting that evening samples might be more stable. Winsorization of extreme values resulted in the highest temporal stability, regardless of time of day (mean winsorized morning $r = .61$, mean winsorized evening $r = .77$, mean nonwinsorized morning $r = .45$, mean nonwinsorized evening $r = .54$) or whether the dataset was strict or relaxed (mean winsorized strict $r = .70$, mean winsorized relaxed $r = .68$, mean nonwinsorized strict $r = .47$, mean nonwinsorized relaxed $r = .52$). Relaxed data sets had an average stability of $r = .56$ compared to an average stability or $r = .52$ for strict datasets. However, it is important to always consider data management techniques in the context of one's specific dataset. For example, winsorization might be less appropriate when there are many extreme cases in a dataset. Further, the decision to modify observed values should always involve contemplation about how "extreme" values are defined, the likelihood that they are valid (not the result of measurement error), and the influence "extreme" values would have on planned analyses (e.g., assumptions of normality, sensitivity to outliers).

It will come as no surprise that, in addition to statistical procedure, measurement procedure can influence temporal stability as well. In addition to the actual method of data collection (e.g., specific self-report measure, particular imaging scanner model), some biological variables can be measured from different sources. For example, inflammatory proteins most frequently are measured via assaying blood samples (e.g., Moriarity et al., 2020a; Muscatell et al., 2016), but salivary measures have been increasing in popularity because they are less expensive and invasive than blood-based methods. However, the utility and comparability of these methods has been questioned as salivary markers of inflammation might reflect local, rather than systemic, immune function (Riis et al., 2015). Out et al. (2012) made an important contribution to this discussion by comparing the one- and two-year retest reliabilities of both plasma and salivary measures of CRP in a sample of adult women. Plasma CRP had higher one-year retest reliability than saliva CRP between years 2 and 3 ($r = .70$ vs. .57), but lower reliability between years 1 and 2 ($r = .53$ vs. .61). Plasma CRP also had higher two-year reliability ($r = .58$ vs. .46). Thus, results indicate comparable, but not identical, one and two-year retest stabilities when using these two methods to measure CRP.

Another important factor to consider when assessing temporal stability is the role of human development. Particularly for youth undergoing drastic growth and developmental changes, it is plausible that temporal stabilities of many biological variables will differ compared to adults. Riis et al. (2014) extended the previous study to a sample of adolescent girls using a similar design (i.e., 3 yearly measurements of plasma and saliva inflammatory analytes). This study assessed nine cytokines, but did not measure CRP, so results cannot be directly compared. Controlling for age, the average year 1 to year 2, year 2 to year 3, and year 1 to year 3 reliabilities were higher for serum compared to saliva (average *r*s = .61 vs .30, .33 vs. .25, and .40 vs. .34, respectively). However, when comparing the stability of individual proteins, a more complex picture emerged. One-year retest reliability was uniformly higher for plasma between years 1 and 2 (*r*s = .39 - .75 vs. .21 - .38). However, this discrepancy was less consistent between years 2 and 3 in which plasma reliability was higher for only four of the seven analytes (plasma *r*s = .10 - .54; saliva *r*s = .09 - .36) and for two-year reliability, for which saliva reliability was higher for four of the analytes (plasma *r*s = .16 - .57; saliva *r*s = .19 - .46). Thus, although these two studies suggest that serum measures of inflammation might be more stable than salivary measures, there might be important protein-level differences in ideal measurement methods. Also, the mouth is home to a complex microbiome that might introduce more confounding factors compared to circulating blood (Giannobile et al., 2009). Thus, future research establishing best practices for salivary methods of collection might find different estimates of temporal stability.

Another popular way to quantify temporal stability is intra-class correlation coefficients (ICCs), which assess the proportion of total variance (between-person + within-person) that is attributable to between-person differences. Thus, higher ICCs indicate less relative within-person variability and greater temporal stability. Conventionally, ICCs less than .40 are considered poor, between .40 and .59 are considered fair, between .60 and .74 are considered good, and above .75 are considered excellent indicators of temporal stability (Cicchetti, 1993), but desired ICCs should be considered in the context of the construct being studied. An important distinction between ICCs and retest reliability indexed by Pearson's *r* is that correlations primarily reflect rank-order stability (i.e., an individual will have the same relative ranking in a sample at Time 1 and Time 2), whereas ICCs reflect rank-order stability *and* mean-level changes between time points. Thus, ICCs are a preferable measure when evaluating how stable a given score is over time.

Continuing the discussion of inflammation, Shields et al. (2019) reported ICCs (in their supplemental material) for seven different salivary inflammatory proteins (CRP, IL-6, IL-8, IL-18, IL-1β, TNFα, MCP-1). They report stability estimates for two different durations: 120 minutes apart during the same testing session ("short-term reliability") and an 18-month follow-up ("long-term stability"). Importantly, testing stability of salivary analytes within the same testing session can help identify how many measurements of these proteins would be necessary to achieve a specific level of reliability. Short-term reliability ICCs ranged from .37 (for IL-8) to .80 (for CRP). To reach a goal short-term reliability of $r = .80$ using the Spearman-Brown prophecy formula, between one (CRP) and four measurements (IL-8 and IL-18) were needed. The number of measurements needed to reach a goal short-term reliability indexed by ICCs was not reported. ICCs were low for all 7 proteins at the 18-month follow-up (all ICCs $< .28$), suggesting lower temporal stability of salivary inflammatory proteins using ICCs compared to Pearson's *r*. Conceptually, this indicates that salivary inflammatory proteins might be more stable in terms of their person-level rank-order than their actual value.

Given the relative expense of much biological psychiatry research (e.g., neuroimaging), many studies are cross-sectional and prospective studies typically have small sample sizes. Thus, meta-analyses pooling the results of multiple studies together have the potential to be very useful in investigating the temporal stability of various measures. Elliott et al. (2020) evaluated temporal stability of task-related fMRI measures in regions of interest (ROIs) using a meta-analysis of 90 substudies (N = 1,008 and 1,146 ICC estimates). When selecting articles, the authors noticed that several of the studies reported thresholded ICCs (i.e., only reported ICCs above a threshold, comparable to only reporting effect sizes for results with $p < .05$). Due to concerns this might inflate estimates of reliability, meta-analyses were conducted separately for studies reporting unthresholded vs. thresholded ICCs. These concerns were supported by results showing that the average ICC for unthresholded results (77 substudies) was poor (mean ICC = .397; 95% CI, .330 - .460), whereas the average stability for tasks in thresholded substudies (13

substudies) was moderate (mean ICC = .705; 95% CI, .628 - .768). Further, a moderation analysis including all substudies confirmed that the decision to report thresholded ICCs was associated with significantly higher ICCs. Importantly, test-retest interval (the duration between the two points of measurement) was not found to be a significant moderator of temporal stability, although the authors do not provide information on the average test-retest interval or variability in the intervals between studies. The authors highlight several methodological limitations of their meta-analysis (e.g., different, potentially outdated scanners, different pre-processing and analysis pipelines).

These results suggest lower than ideal temporal stability for the study of individual differences. Importantly, the authors highlight that these tasks were created to robustly result in group-level changes, not to assess between-person differences in these changes. Therefore, the problem is not necessarily in the measures, but how researchers have extended their use to research questions they were not built to address. It also is important to highlight that this study only investigated ROIs. Similar analyses examining whole brain patterns might be more temporally stable. Additionally, some common ROIs not included in this paper (e.g., left nucleus accumbens and right anterior insula activity) have better temporal stability (e.g., ICC > .5) at large intervals (> 2.5 years) during the monetary incentive delay task included in Elliott et al. (2020) (Wu et al., 2014). In response to Elliott et al. (2020), Kragel et al. (2020, note this is a pre-print that has not undergone peer review) describe nine recent studies demonstrating strong short-term stability (i.e., less than five weeks) for task-based fMRI measures. They conclude that studies aggregating information across multiple brain regions (rather than ROIs) and/or aggregation across similar tasks, with larger samples, more data per participant (i.e., more time in the scanner), and shorter retest intervals paint a more promising picture of temporal stability for fMRI task measures than Elliott et al. (2020). It is worth note that many of these conditions involve using additional data (i.e., larger samples, more data per participant, aggregation across brain regions and similar tasks), underscoring that aggregating more data (e.g., across studies, see Segerstrom and Boggero (2020) below) will average out misestimations resulting from unreliable measures. Thus, further work is needed to identify best practices for individual differences research using various fMRI measures.

Recall that measures taken across multiple time points for multiple people have three sources of variability: between-person, within-person, and measurement error. Generalizability theory (Shavelson and Webb, 1991) is an extension of these principles that estimates what proportion of a single assessment is generalizable to other time points by separating variance due to stable individual differences, measurement occasions, and the interaction between the two. Results of generalizability analyses then can be used to inform the design of later studies with the goal of achieving a desired reliability. Segerstrom et al. (2014) applied this theory to investigate how many days of sampling would be needed to reliably characterize between-person differences and within-person changes in three cortisol metrics: diurnal mean, diurnal slope, and area under the curve (AUC) in two separate samples. Sample 1 consisted of young adults who provided five cortisol samples per day, for three consecutive days, across five separate occasions (mean time after previous occasion; Time 2: 44 days, Time 3: 57 days, Time 4: 36 days, Time 5: 29 days). Results indicated that three days were necessary for adequate reliability to facilitate individual differences research (defined as $r = .60$ in this study) for the diurnal mean, four days for the AUC, and 11 days for diurnal slope. Further, reliable measurement of within-person changes would require three days of data for the mean, four for AUC, and eight for slope. Correlations comparing slopes calculated with 2, 3, and 4 time points per day suggested that collecting two samples per day (taken during the morning and evening) were excellent at reproducing slope estimates using four samples ($r = .97$), suggesting that collecting more than two samples per day does not substantively improve measurement. To evaluate whether these results replicate in a demographically different sample, a second study was conducted in older adults that resulted in comparable estimates. These results suggest that collecting two samples per day for several days will provide more reliable estimates than collecting more samples, but across fewer days.

3.5. Temporal Specificity

In addition to temporal stability, temporal specificity of effects is integral to advance longitudinal research. To illustrate this, consider the following studies of inflammation as a risk factor for depression. Miller and Cole (2012) reported that CRP predicted depression symptoms at a six-month follow-up, but only in female adolescents exposed to childhood adversity. Gimeno et al. (2009) found that CRP and IL-6 predicted depression symptoms 12 years in the future. However, neither van den Biggelaar et al. (2007; five years of annual follow-ups) nor Stewart et al. (2009; six-year follow-up) found significant associations between IL-6 and future depression symptoms, but van der Biggelaar et al. found that CRP predicted future depression. Further, Copeland et al. (2012) did not find that CRP predicted future depression in a sample of adolescents with up to nine assessments over a 12-year period. Although there might be (and likely are) many moderators influencing this heterogeneity in results, time to follow-up is a plausible candidate that could inform design of future, and interpretation of past, studies.

Moriarity et al. (2019) explored this possibility in a sample of 201 adolescents with a baseline blood draw and a total of 582 assessments of depression symptoms (time to follow-up ranged from .07 – 30.49 months). Using hierarchical linear models, they tested main effects models of five inflammatory proteins on change in depression symptoms as well as five exploratory models testing interactions between the five biomarkers, sex, and time to follow-up. The only protein with a significant unconditional main effect was CRP; however, three of the four remaining proteins demonstrated significant three-way interactions. Specifically, both IL-6 and TNFα had stronger, more positive associations with change in depression symptoms as time to follow-up increased, but only for females (e.g., Fig. 1). Conversely, IL-8 had a stronger association with change in depression symptoms for males as time to follow-up increased, but the association was negative. These results highlight how associations might not replicate between samples with different demographic characteristics (e.g., sex) or different intervals between assessments. This line of inquiry might be particularly important during adolescence, which is both a time of elevated risk for first onset of many psychopathologies (e.g., depression; Cummings et al., 2014) as well as a time of rapid social, biological, and psychological development. Although testing individual proteins maximized this

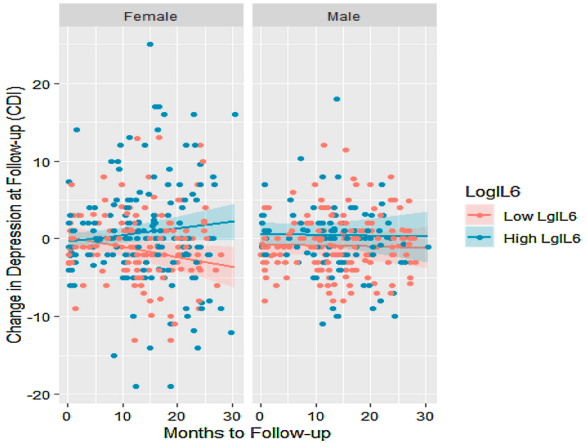

Fig. 1. Temporal specificity of Log IL-6 predicting change in depression symptoms by sex. This figure was first presented in Moriarity et al. (2019). Note: IL = interleukin, CDI = Children's Depression Inventory. Shaded regions indicate 95% confidence intervals.

study's relevance (as this is the most common approach in immunopsychiatry) it is worth considering how results might have changed if an aggregate variable of "inflammation" was also tested. Possibly, by aggregating shared variance and increasing power, an empirically-supported aggregate variable may have predicted change in depression at a wider range of follow-up intervals or had larger effect sizes.

The rise in popularity of intensive longitudinal designs allows for a wealth of new opportunities to investigate temporal specificity on a smaller time scale. For example, Graham-Engeland et al. (2018) measured serum levels of seven inflammatory proteins (combined into an inflammatory composite) and CRP (analyzed individually) after a 14-day ecological momentary assessment (EMA) protocol. Before starting the EMA protocol, participants completed questions about recalled positive and negative affect "over the past month". Then, participants completed questions about experienced positive and negative affect five times per day for 14 days leading up to the blood draw. Neither the inflammatory composite nor CRP were significantly predicted by positive or negative affect "over the past month" or aggregated positive or negative affect over the 14-day EMA protocol. However, when the affect variables were separated by week, Week 2 (closest to the blood draw), but not Week 1, negative affect significantly predicted the inflammatory composite variable. Exploratory analyses found that the association between negative affect and inflammation consistently increased in strength as the lag between measurements shortened. Thus, these two studies illustrate how it is possible to leverage longitudinal studies of different time scales to identify whether risk factors for psychopathology operate on a proximal or distal time scale, providing important insight to study design and intervention efforts.

3.6. Effect Size and Power

As reviewed in the conceptual portion of this paper, all of the physiometric examples reviewed thus far have implications for model performance; however, some researchers have empirically tested the relationship between physiometrics and effect size/power in biological psychiatry. For example, Hajcak et al.s' (2017) paper on how internal consistency of ERN changes as a function of trials completed in two groups of participants with, and without, generalized anxiety disorder (reviewed above) also tested how between-group effect sizes were related to internal consistency. Cohen's *d* increased almost parallel to increases in internal consistency as the number of trials increased ($r = .94$). Given that two primary goals of biological psychiatry are understanding i) group differences between those with and without mental illness, and ii) the between-person variability in within-person effects contributing to psychiatric risk, resilience, and treatment, this is noteworthy.

Simulation studies present a powerful option to evaluate the state of current measurement practices. Segerstrom and Boggero (2020) used 212 study designs included as part of a meta-analysis (Boggero et al., 2017) on the relationship between various psychosocial correlates and cortisol awakening response to investigate the probability of misestimates using these data. 100,000 data sets were simulated for each study design with sample sizes and reliability estimates extracted from the original studies. Boggero et al. (2020) found a meta-analytic effect size of less than $r = 0.10$, which was used as the "true" effect size for the purposes of the simulation study. Two types of misestimates were assessed: 1) sign errors (i.e., when the association was negative, instead of positive like the "true" effect); and 2) magnitude errors (i.e., when the estimate was more than .10 away from the "true" effect size). Consistent with literature reviewed above, more days of sampling in cortisol studies are associated with higher reliability. More days of sampling (and, by extension, reliability) was, in turn, consistently negatively correlated with both sign and magnitude errors in the simulations. Given that results found that around 20% of all simulations resulted in sign errors, and nearly 40% in magnitude errors, this study highlights increased cortisol sampling as a way to increase reliability and overall study quality.

4. The Promise of Biological Psychiatry

Biological psychiatry has the potential to enhance both physical and mental health through the investigation of the reciprocal associations between the body and mind. However, this potential only can be realized with carefully crafted theory and rigorous methodology. Many have argued that the field has fallen short of its promise to meaningfully impact psychiatric classification, diagnosis, prevention, and treatment so far (Kapur et al., 2012; Miller, 2010; Venkatasubramanian and Keshavan, 2016). One important reason for this may be that a lack of sufficient attention to key measurement properties of biological variables has constrained the utility of these data in statistical modeling, and thus, inference generation, despite rapid technological advances allowing for more precise data acquisition in many biological subfields.

Although the physiometric characteristics covered in this review are far from exhaustive, we would like to reiterate five steps that would improve biological psychiatry research: 1) thoughtful investigation of the dimensionality of complex biological constructs in datasets including multiple indicators of these constructs; 2) standardized reporting of internal consistency when using aggregate measures; 3) careful consideration of the implications of method-specific variance; 4) standardized reporting of temporal stability, preferably calculated with the sample being analyzed or at least a reference to previous research with a similar time frame; and 5) increased exploration into the temporal specificity of associations between biological and behavioral phenomena. Further, it is imperative to keep in mind how the results of these investigations might be contingent on other analytic choices (e.g., handling of extreme values; Landau et al., 2019) and sample characteristics (e.g., sex; Moriarity et al., 2019).

A physiometric awakening in biological psychiatry would promote a wide array of benefits to the field and those whom this work is intended to benefit. Projects uninformed by basic measurement principles germane to their study methods risk inflating the noise-to-signal ratio in statistical models. As a result, there is an increased risk for false-negatives and false-positives, hindering the actual progress of the field as well as belief in its utility relative to the associated costs. Further, many standardized effect sizes between biological and psychological variables likely are biased downward due to less than ideal matching of measures to procedures and method specific variance, weakening the appearance of their practical implications. Thoughtful application of measurement principles can reduce error-related variability in future studies via improvement of both study design and statistical modeling, resulting in improved replicability of findings and less biased effect sizes.

Moreover, physiometric studies can provide guidance about which variables have the most utility, under what research designs they operate well, and how to optimally model constructs of interest. To illustrate this, consider designing a study of experienced negative affect as a predictor of inflammatory and coagulatory markers in adolescents. Having read Nelson et al. (2011), you know that aggregating variables containing overlapping variance can accentuate the shared variance related to other variables, increasing power. You originally considered the same panel of biomarkers as Egnot et al. (2018), but you decided not to assay and analyze slCAM-1 and Lp(a) because neither loaded onto either of the two factors in their study. This decision saves you money, enabling recruitment of more participants, hiring additional staff, or purchasing other supplies. Additionally, because Engeland et al. (2018) found that the association between negative affect and inflammation was stronger at shorter intervals, you might plan a one-week EMA protocol rather than a two-week protocol, saving money, time, and participant burden. However, instead of testing separate regressions for each day of negative affect, you could improve statistical rigor of this comparison by testing for moderations by time interval using multilevel

models like Moriarity et al. (2019).

In addition to improving study design, thoughtful application of various statistical approaches holds the potential to ameliorate physiometric issues in biological psychiatry. One example is structural equation modeling (SEM), a powerful tool for reducing the impact of poor reliability on statistical models. SEM allows the estimation of latent factors from the shared variance between items, removing measurement error associated with individual observed variables and accentuating shared variance between biomarkers of interest. However, SEM models require larger samples than traditional models. Thus, multi-study collaborations might be necessary to permit model testing for more expensive measures.

As described in Perkins et al. (2017), many physiological variables of interest are associated with many different psychological constructs. Thus, when possible, researchers should carefully consider whether building statistical models that can isolate portions of variance relevant to one trait vs. another would be beneficial. However, we would like to underscore that the suitability of various variance isolation techniques is context dependent. As described above, variance removed from a variable always comes from the "true" and reliable variance, never from error variance. Thus, difference scores or predictors with variance partialled out for covariates are almost always less reliable and have a lower signal-to-error ratio (Lynam et al., 2006). This is amplified when the predictors are highly correlated (Thomas and Zumbo, 2012). Finally, it also is critical to remember that difference scores (or predictors with variance partialled out in multiple regression) are conceptually different than the raw variables. These interpretive concerns are more extreme with more heterogenous (lower internal consistency) measures, because it is more likely that the variance removed might only be associated with a subset of the components of the original variable.

Additionally, most of this article has discussed physiometric work anchored in classical test theory. Future work could utilize generalizability theory, an extension of classical test theory described above in the review of Segerstrom et al. (2014). Alternatively, item response theory (IRT) estimates reliability for varying levels of a continuum rather than the entire range of a measure. Typically, IRT requires binary or polytomous indicators, but continuous response models (CRM) are an extension of IRT models that allow for continuous variables (Samejima, 1973). Physiometric research utilizing these approaches might lead to useful insight for how to best collect and model biological data.

Increasing the efficiency of study design and statistical modeling will improve the ability to accurately detect associations and their effect sizes. These advancements have the potential to smooth the transition from basic research to the improvement of interventions and policy via increasing confidence in results and the ability to gauge their utility. Importantly, with lower rates of false positives, there is a reduced chance that ineffective biological interventions may be explored that have little to no real-world utility.

Fortunately, as reviewed above, some researchers are working to arm the rest of the field with this crucial information. As more physiometric work is published, the value of comprehensive reviews of this literature increases. Recently, Segerstrom (2020) and Gloger et al. (2020) published reviews of salivary and serum biomarker physiometrics, respectively, but many more topics would benefit from a focused physiometric review (e.g., neuroimaging, ERP, heart rate variability).

However, it is critical to admonish the dangers of treating particular levels of physiometric characteristics as benchmarks to hit, without careful consideration of what they mean in relation to the constructs being studied. Several methodologists have warned that primarily focusing on creating measures with high internal consistency can result in the removal of items/components that contribute to lower internal consistency, but would help capture the true breadth of the construct of interest (Clark and Watson, 2019; Cronbach and Meehl, 1955). This sacrifices construct validity for higher internal consistency and faux-unidimensionality. However, it is important to note that this concern is only applicable to the creation of measures using different biomarkers (e.g., different inflammatory cytokines), not repeated measures of the same variable. Further, internal consistency increases as a function of the number of components included in its calculation, potentially resulting in larger, but not better, measures. Additionally, although there are many contexts in which high temporal stability can be beneficial, it is critical to avoid overvaluing components of larger constructs (e.g., brain regions for neuroimaging studies) with higher reliability. Rather, there should be reciprocal interplay between methodology and theory.

Creating a solid physiometric foundation for biological psychiatry is not without obstacles. First and foremost, biological variables often are more expensive to measure than psychological variables, some of which can be measured via self-report questionnaires administered online from the comfort of participants' homes. Measurement research and construct validation are, by their nature, iterative processes, amplifying the associated cost of this work. However, it is crucial to appreciate that good physiometric research is an investment; it will result in increased statistical power and better study design in the future, saving money and time. This requires investment both on the part of researchers as well as funding agencies. Fortunately, there is a lot of important work that can be done with existing data sets. Any study with repeated measures of a variable can estimate its temporal stability. Any study using an aggregate measure can assess the internal consistency of its components. In fact, there are many publicly available data sets that offer great opportunities for physiometric research (e.g., the Human Connectome Project; Van Essen et al., 2013).

Finally, this work can, at times, be statistically intensive and conceptually abstract. One of the strengths of biological psychiatry is that, by nature, it is an interdisciplinary pursuit with experts along the biology—psychology spectrum. Collaboration with statisticians and measurement specialists can serve as a catalyst for the efficient, high-quality research that is needed for biological psychiatry to reach its full academic, clinical, and policy-informing potential.

5. Conclusion

It is important to end on a clarification that the issues highlighted in this article should not be received with apprehension or pessimism. Rather, it is an invitation to ask new questions of the data collected to help the field of biological psychiatry realize its potential. Biological psychiatry has been criticized for falling short of its considerable promise in advancing knowledge about the interplay between biology and behavior in ways that will translate to substantive impact on clinical outcomes (Kapur et al., 2012; Miller, 2010; Venkatasubramanian and Keshavan, 2016). One addressable barrier to meaningfully advancing biological psychiatry is an understanding and appreciation of measurement properties for biological variables. By leveraging existing data sets and prioritizing funding for physiometric research, it is possible to advance current methods to allow for more informative and replicable studies that will provide greater clarity into what areas of research offer the greatest promise to make meaningful impacts on mental health, and how best to integrate them into intervention efforts.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

Thank you to Drs. Michelle Bryne, Thomas Olino, Lauren Ellman, David Smith, Suzanne Segerstrom, and Robin Nusslock for providing feedback on drafts of this article.

References

Boggero, I.A., Hostinar, C.E., Haak, E.A., Murphy, M.L.M., Segerstrom, S.C., 2017. Psychosocial functioning and the cortisol awakening response: Meta-analysis, P-curve analysis, and evaluation of the evidential value in existing studies. Biol. Psychol. 129, 207–230. https://doi.org/10.1016/j.biopsycho.2017.08.058.

Bradford, D.E., Starr, M.J., Shackman, A.J., Curtin, J.J., 2015. Empirically based comparisons of the reliability and validity of common quantification approaches for eyeblink startle potentiation in humans. Psychophysiology 52, 1669–1681. https://doi.org/10.1111/psyp.12545.

Cicchetti, D.V., 1993. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Clark, L.A., Watson, D., 2019. Constructing validity: New developments in creating objective measuring instruments. Psychol. Assess. 31, 1412–1427. https://doi.org/10.1037/pas0000626.

Clark, L.A., Watson, D., 1995. Constructing validity: Basic issues in objective scale development. Psychol. Assess. 7, 309–319.

Copeland, W.E., Shanahan, L., Worthman, C., Angold, A., Costello, E.J., 2012. Cumulative depression episodes predict later C-reactive protein levels: A prospective analysis. Biol. Psychiatry 71, 15–21. https://doi.org/10.1016/j.biopsych.2011.09.023.

Cortina, J.M., 1993. What is coefficient alpha? An examination of theory and applications. J. Appl. Psychol. 78, 98–104. https://doi.org/10.1037//0021-9010.78.1.98.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychol. Bull. 52, 281–302.

Cummings, C., Caporino, N., Kendall, P.C., 2014. Comorbidity of anxiety and depression in children and adolescents: 20 Years After. Psychol. Bull. 140, 816–845. https://doi.org/10.1037/a0034733.

Cuthbert, B.N., Kozak, M.J., 2013. Constructing constructs for psychopathology: The NIMH research domain criteria. J. Abnorm. Psychol. 122, 928–937. https://doi.org/10.1037/a0034028.

Davidshofer, K.R., Murphy, C.O., 2005. Psychological testing: principles and applications.

Edwards, J.R., 2001. Ten difference score myths. Organ. Res. Methods 4, 265–287. https://doi.org/10.1177/109442810143005.

Egnot, N.S., Barinas-Mitchell, E., Criqui, M.H., Allison, M.A., Ix, J.H., Jenny, N.S., Wassel, C.L., 2018. An exploratory factor analysis of inflammatory and coagulation markers associated with femoral artery atherosclerosis in the San Diego Population Study. Thromb. Res. 164, 9–14. https://doi.org/10.1016/j.thromres.2018.02.003.

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T.E., Caspi, A., Hariri, A.R., 2020. What is the test-retest reliability of common task-fMRI measures? New empirical evidence and a meta-analysis. Psychol. Sci. 681–700. https://doi.org/10.1101/681700.

Giannobile, W.V., Beikler, T., Kinney, J.S., Ramseier, C.A., Wong, D.T., 2009. Saliva as a diagnostic tool for periodontal disease: current state and future directions. Periodontol 2000, 52–64. https://doi.org/10.1111/j.1600-0757.2008.00288.x. Saliva.

Gloger, E.M., Smith, G.T., Segerstrom, S.C., 2020. Stress physiology and physiometrics. Handb. Res. Methods Heal. Psychol. 127–140.

Graham-Engeland, J.E., Sin, N.L., Smyth, J.M., Jones, D.R., Knight, E.L., Sliwinski, M.J., Almeida, D.M., Katz, M.J., Lipton, R.B., Engeland, C.G., 2018. Negative and positive affect as predictors of inflammation: Timing matters. Brain. Behav. Immun. 74, 222–230. https://doi.org/10.1016/j.bbi.2018.09.011.

Guadagnoli, E., Velicer, W.F., 1988. Relation of Sample Size to the Stability of Component Patterns. Psychol. Bull. 103, 265–275. https://doi.org/10.1037/0033-2909.103.2.265.

Hajcak, G., Meyer, A., Kotov, R., 2017. Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. J. Abnorm. Psychol. 126, 823–834. https://doi.org/10.1037/abn0000274.

Hajcak, G., Patrick, C.J., 2015. Situating psychophysiological science within the Research Domain Criteria (RDoC) framework. Int. J. Psychophysiol. 98, 223–226. https://doi.org/10.1016/j.ijpsycho.2015.11.001.

Holiga, Š., Sambataro, F., Luzy, C., Greig, G., Sarkar, N., Renken, R.J., Marsman, J.B.C., Schobel, S.A., Bertolino, A., Dukart, J., 2018. Test-retest reliability of task-based and resting-state blood oxygen level dependence and cerebral blood flow measures. PLoS One 13, 1–16. https://doi.org/10.1371/journal.pone.0206583.

Kakeda, S., Watanabe, K., Nguyen, H., Katsuki, A., Sugimoto, K., Igata, N., Abe, O., Yoshimura, R., Korogi, Y., 2020. An independent component analysis reveals brain structural networks related to TNF-α in drug-naïve, first-episode major depressive disorder: a source-based morphometric study. Transl. Psychiatry 10. https://doi.org/10.1038/s41398-020-00873-8.

Kapur, S., Phillips, A.G., Insel, T.R., 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it. Mol. Psychiatry 17, 1174–1179. https://doi.org/10.1038/mp.2012.105.

Kaye, J.T., Bradford, D.E., Curtin, J.J., 2016. Psychometric properties of startle and corrugator response in NPU, affective picture viewing, and resting state tasks. Psychophysiology 53, 1241–1255. https://doi.org/10.1111/psyp.12663.

Kragel, P.A., Han, X., Kraynak, T.E., Gianaros, P.J., Wagner, T.D., 2020. fMRI can be highly reliable, but it depends on what you measure. PsyArXiv.

Landau, E.R., Trinder, J., Simmons, J.G., Raniti, M., Blake, M., Waloszek, J.M., Blake, L., Schwartz, O., Murray, G., Allen, N.B., Byrne, M.L., 2019. Salivary C-reactive protein among at-risk adolescents: A methods investigation of out of range immunoassay data. Psychoneuroendocrinology 99, 104–111. https://doi.org/10.1016/j.psyneuen.2018.08.035.

Li, Y.O., Adali, T., Calhoun, V.D., 2007. Estimating the number of independent components for functional magnetic resonance imaging data. Hum. Brain Mapp. 28, 1251–1266. https://doi.org/10.1002/hbm.20359.

Loevinger, J., 1957. Objective tests as instruments of psychological theory. Psychol. Rep. 3, 635–694.

Luking, K.R., Nelson, B.D., Infantolino, Z.P., Sauder, C.L., Hajcak, G., 2017. Internal consistency of functional magnetic resonance imaging and electroencephalography measures of reward in late childhood and early adolescence. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 2, 289–297. https://doi.org/10.1016/j.bpsc.2016.12.004.

Lynam, D.R., Hoyle, R.H., Newman, J.P., 2006. The Perils of Partialling Cautionary Tales From Aggression and Psychopathy. Assessment 13, 328–341. https://doi.org/10.1177/1073191106290562.

Miller, G.A., 2010. Mistreating psychology in the decades of the brain. Perspect. Psychol. Sci. 5, 716–743. https://doi.org/10.1038/jid.2014.371.

Miller, G.E., Cole, S.W., 2012. Clustering of depression and inflammation in adolescents previously exposed to childhood adversity. Biol. Psychiatry 72, 34–40. https://doi.org/10.1016/j.biopsych.2012.02.034.

Moriarity, D.P., Mac Giollabhui, N., Ellman, L.M., Klugman, J., Coe, C.L., Abramson, L.Y., Alloy, L.B., 2019. Inflammatory proteins predict change in depressive symptoms in male and female adolescents. Clin. Psychol. Sci. 7, 754–767. https://doi.org/10.1177/2167702619826586.

Moriarity, D.P., McArthur, B.A., Ellman, L.M., Coe, C.L., Abramson, L.Y., Alloy, L.B., 2018. Immunocognitive model of depression secondary to anxiety in adolescents. J. Youth Adolesc. 47, 2625–2636. https://doi.org/10.1007/s10964-018-0905-7.

Moriarity, D.P., Ng, T., Curley, E., Mcarthur, B.A., Ellman, L.M., Coe, C.L., Abramson, L. Y., Alloy, L.B., 2020a. Reward sensitivity, cognitive response style, and inflammatory response to an acute stressor in adolescents. J. Youth Adolesc. 49, 2149–2159.

Moriarity, D.P., Ng, T., Titone, M.K., Chat, I.K., Nusslock, R., Miller, G.E., Alloy, L.B., 2020b. Reward sensitivity and ruminative response styles for positive and negative affect interact to predict inflammation and mood symptomatology. Behav. Ther. 51, 829–842. https://doi.org/10.1016/j.beth.2019.11.007.

Muscatell, K.A., Moieni, M., Inagaki, T.K., Dutcher, J.M., Jevtic, I., Breen, E.C., Irwin, M. R., Eisenberger, N.I., 2016. Exposure to an inflammatory challenge enhances neural sensitivity to negative and positive social feedback. Brain. Behav. Immun. 57, 21–29. https://doi.org/10.1016/j.bbi.2016.03.022.

Nelson, L.D., Patrick, C.J., Bernat, E.M., 2011. Operationalizing proneness to externalizing psychopathology as a multivariate psychophysiological phenotype. Psychophysiology 48, 64–72. https://doi.org/10.1111/j.1469-8986.2010.01047.x.

Ng, T.H., Alloy, L.B., Smith, D.V., 2019. Meta-analysis of reward processing in Major Depressive Disorder: Distinct abnormalities within the reward circuit? Transl. Psychiatry 9, 2–10.

Out, D., Hall, R.J., Granger, D.A., Page, G.G., Woods, S.J., 2012. Assessing salivary C-reactive protein: Longitudinal associations with systemic inflammation and cardiovascular disease risk in women exposed to intimate partner violence. Brain. Behav. Immun. 26, 543–551. https://doi.org/10.1016/j.bbi.2012.01.019.

Patrick, C.J., Iacono, W.G., Venables, N.C., 2019. Incorporating neurophysiological measures into clinical assessments: Fundamental challenges and a strategy for addressing them. Psychol. Assess. 31, 1512–1529. https://doi.org/10.1037/pas0000713.

Patrick, C.J., Venables, N.C., Yancey, J.R., Hicks, B.M., Nelson, L.D., Kramer, M.D., 2013. A construct-network approach to bridging diagnostic and physiological domains: Application to assessment of externalizing psychopathology. J. Abnorm. Psychol. 122, 902–916. https://doi.org/10.1037/a0032807.

Perkins, E.R., Yancey, J.R., Drislane, L.E., Venables, N.C., Balsis, S., Patrick, C.J., 2017. Methodological issues in the use of individual brain measures to index trait liabilities: The example of noise-probe P3. Int. J. Psychophysiol. 111, 145–155. https://doi.org/10.1016/j.ijpsycho.2016.11.012.

Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A.B. M., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. Neuroimage 60, 1746–1758. https://doi.org/10.1016/j.neuroimage.2012.01.129.

Riis, J.L., Granger, D.A., Dipietro, J.A., Bandeen-Roche, K., Johnson, S.B., 2015. Salivary cytokines as a minimally-invasive measure of immune functioning in young children: Correlates of individual differences and sensitivity to laboratory stress. Dev. Psychobiol. 57, 153–167. https://doi.org/10.1002/dev.21271.

Riis, J.L., Out, D., Dorn, L.D., Beal, S.J., Denson, L.A., Pabst, S., Jaedicke, K., Granger, D. A., 2014. Salivary cytokines in healthy adolescent girls: Intercorrelations, stability, and associations with serum cytokines, age, and pubertal stage. Dev. Psychobiol. 56, 797–811. https://doi.org/10.1002/dev.21149.

Rogosa, D.R., Willett, J.B., 1983. Demonstrating the reliability of the difference score in the measurement of change. J. Educ. Meas. 20, 335–343.

Samejima, F., 1973. Homogenous case of the continuous response model. Psychometrika 38, 203–2019.

Segerstrom, S.C., 2020. Physiometrics in Salivary Bioscience. Int. J. Behav. Med. 27, 262–266.

Segerstrom, S.C., Boggero, I.A., 2020. Expected Estimation Errors in Studies of the Cortisol Awakening Response: A Simulation. Psychosom. Med. 82, 751–756. https://doi.org/10.1097/PSY.0000000000000850.

Segerstrom, S.C., Boggero, I.A., Smith, G.T., Sephton, S.E., 2014. Variability and reliability of diurnal cortisol in younger and older adults: Implications for design decisions. Psychoneuroendocrinology 49, 299–309. https://doi.org/10.1016/j.psyneuen.2014.07.022.

Segerstrom, S.C., Smith, G.T., 2012. Methods, variance, and error in psychoneuroimmunology research: The good, the bad, and the ugly. In:

Segerstrom, S.C. (Ed.), Oxford Handbook of Psychoneuroimmunology. Oxford U Press, New York, NY, pp. 421–432.
Shavelson, R.J., Webb, N.M., 1991. Generalizability Theory: A Primer. Sage.
Shields, G.S., Slavich, G.M., Perlman, G., Klein, D.N., Kotov, R., 2019. The short-term reliability and long-term stability of salivary immune markers. Brain. Behav. Immun. 81, 650–654. https://doi.org/10.1016/j.bbi.2019.06.007.
Slavich, G.M., Irwin, M.R., 2014. From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. Psychol. Bull. 140, 774–815. https://doi.org/10.1037/a0035302.
Stewart, J.C., Rand, K.L., Muldoon, M.F., Kamarck, T.W., 2009. A prospective evaluation of the directionality of the depression-inflammation relationship. Brain. Behav. Immun. 23, 936–944. https://doi.org/10.1016/j.bbi.2009.04.011.
Tabachnick, B.G., Fidell, L.S., 2013. Using multivariate statistics, Sixth. ed. Pearson, Boston, MA.
Thomas, D.R., Zumbo, B.D., 2012. Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. Educ. Psychol. Meas. 72, 37–43. https://doi.org/10.1177/0013164411409929.
van den Biggelaar, A.H.J., Gussekloo, J., de Craen, A.J.M., Frölich, M., Stek, M.L., van der Mast, R.C., Westendorp, R.G.J., 2007. Inflammation and interleukin-1 signaling network contribute to depressive symptoms but not cognitive decline in old age. Exp. Gerontol. 42, 693–701. https://doi.org/10.1016/j.exger.2007.01.011.
Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., 2013. The WU-Minn Human Connectome Project: An overview. Neuroimage 80, 62–79. https://doi.org/10.1038/jid.2014.371.
Venables, N.C., Foell, J., Yancey, J.R., Kane, M.J., Engle, R.W., Patrick, C.J., 2018. Quantifying inhibitory control as externalizing proneness: A cross-domain model. Clin. Psychol. Sci. 6, 561–580. https://doi.org/10.1177/2167702618757690.
Venkatasubramanian, G., Keshavan, M.S., 2016. Biomarkers in psychiatry – A critique. Ann. Neurosci. 23, 3–5. https://doi.org/10.1159/000443549.
Williams, L.M., 2016. Precision psychiatry: A neural circuit taxonomy for depression and anxiety. The Lancet Psychiatry 3, 472–480. https://doi.org/10.1016/S2215-0366(15)00579-9.Precision.
Wu, C.C., Samanez-Larkin, G.R., Katovich, K., Knutson, B., 2014. Affective traits link to reliable neural markers of incentive anticipation. Neuroimage2014 84, 279–289. https://doi.org/10.1007/978-3-319-55511-9_5.

Daniel P. Moriarity was supported by National Research Service Award F31MH122116.

Lauren B. Alloy was supported by National Institute of Mental Health grants R01MH077908 and R01MH101168.