THE PHYSIOMETRICS OF INFLAMMATION AND IMPLICATIONS FOR MEDICAL AND PSYCHIATRIC RESEARCH: TOWARD EMPIRICALLY-INFORMED INFLAMMATORY COMPOSITES

A Dissertation

Submitted to

the Temple University Graduate Board

In Partial Fulfillment

of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

---

by

Daniel P. Moriarity, M.A.

August 2022

Examining Committee Members:

Lauren B. Alloy, Ph.D., Advisory Chair, Department of Psychology

Thomas Olino, Ph.D., Core Committee Member, Department of Psychology

Lauren M. Ellman, Ph.D., Core Committee Member, Department of Psychology

Michael McCloskey, Ph.D., Committee Member, Department of Psychology

David Smith, Ph.D., Committee Member, Department of Psychology

Debra Bangasser, Ph.D., External Member, Department of Psychology

ABSTRACT

Most psychoneuroimmunology research examines individual proteins; however, some studies have used summed score composites of all available inflammatory markers without evaluating the appropriateness of this decision. Using three different samples (MIDUS-2: N = 1,255 adults, MIDUS-R: N =863 adults, and ACE: N = 315 adolescents), this study investigates the dimensionality of eight inflammatory proteins (C-reactive protein (CRP), interleukin (IL)-6, IL-8, IL-10, tumor necrosis factor-α (TNF-α), fibrinogen, E-selectin, and intercellular adhesion molecule (ICAM)-1) and compares the resulting factor structure to a) an "a priori" factor structure in which all inflammatory proteins equally load onto a single dimension (a technique that has been used previously) and b) proteins modeled individually (i.e., no latent variable) in terms of model fit, replicability, reliability, temporal stability, and their associations with medical history and depression symptoms. A hierarchical factor structure with two first-order factors (Factor 1A: CRP, IL-6, fibrinogen; Factor 2A: TNF-α, IL-8, IL-10, ICAM-1, IL-6) and a second-order general inflammation factor was identified in MIDUS-2 and replicated in MIDUS-R and partially replicated in ACE (which unfortunately only had CRP, IL-6, IL-8, IL-10, and TNF-α but, unlike the other two, has longitudinal data). Both the empirically-identified structure and modeling proteins individually fit the data better compared to the one-dimensional "a priori" structure. Results did not clearly indicate whether the empirically-identified factor structure or the individual proteins modeled without a latent variable had superior model fit. Modeling the empirically-identified factors and individual proteins (without a latent factor) as outcomes of medical diagnoses resulted in comparable conclusions, but modeling empirically-identified factors resulted in fewer results "lost" to correction for multiple comparisons. Importantly, when the factor scores were recreated in a longitudinal dataset, none of the individual proteins, the "a priori" factor, or the empirically-identified general inflammation factor significantly predicted

concurrent depression symptoms in multilevel models. However, both empirically-identified first-order factors were significantly associated with depression, in opposite directions. Measurement properties are reported for the different aggregates and individual proteins as appropriate, which can be used in the design and interpretation of future studies. These results indicate that modeling inflammation as a unidimensional construct equally associated with all available proteins does not fit the data well. Instead, empirically-supported aggregates of inflammation, or individual inflammatory markers, should be used in accordance with theory. Further, the aggregation of shared variance achieved by constructing empirically-supported aggregates might increase predictive validity compared to other modeling choices, maximizing statistical power.

ACKNOWLEDGEMENTS

From the bottom of my heart, I thank my advisor, Dr. Lauren Alloy, for giving me the opportunity to realize a professional dream and pursue a career that amplifies my most deeply seeded interests and passions. You have demonstrated patience, enthusiasm, and a sincere, personal care for my well-being as you mentored me toward a career trajectory that, six years ago, seemed a pipedream. In retrospect, I consider not being accepted into any graduate programs the first year I applied one of the greatest blessings of my life because it led me to working with you in this lab. I look forward to a lifetime of continued collaboration and commit to championing my future mentees with the same passion and patience you have mentored me for over half a decade.

Second, I would like to thank Dr. Lauren Ellman. Until graduate school I had no training in immunology and only basic training in biology, and despite a rapidly growing lab and family, you always found the time to give me mentorship and support in the complexities of psychoneuroimmunology. You simultaneously affirm my growth and push me further and have always been a constant catalyst for my development. I commit to similarly challenging my future mentees to reach their full potential.

Third, I would like to thank Dr. Tom Olino for being a model of how to weave methodological rigor and implications into substantive work. My first academic interests in psychology pertained to issues of measurement, work I thought I had to set aside to focus on a more fundable research narrative. Your mentorship has facilitated the fusion of my substantive and methodological interests, a combination that has led to some of my favorite projects to date. I promise to similarly foster a passion for research methodology in my mentees and my subfield to the best of my ability.

I also would like to thank the Mood and Cognition Lab and my cohort for providing support, intellectual engagement, and sources of levity that breathed life into the more difficult parts of graduate school. Thank you also to Drs. Michael McCloskey, David Smith, and Debra Bangasser, for devoting the time and energy to serve as members of my dissertation committee.

Finally, I also would like to thank God, my family (Mom, Dad and Barb, Connor, Grandma Betty, Grandpa Bruce, Aunt Sis and Uncle Drew, Alex, Jean, Maddy), and my fiancé, Uma. This journey wouldn't be worth it or possible without the support you all have shown me through the years.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

MANUSCRIPT IN JOURNAL ARTICLE FORMAT

Introduction

On average, measurement error attenuates effect sizes, meaning that increased consideration of measurement issues in immunopsychiatry might translate to larger observed effect sizes and more replicable studies. Larger observable effect sizes also translate to increased statistical power. More power reduces necessary sample sizes/numbers of repeated measures to detect effects, thus increasing the cost effectiveness of research. Consequently, a full characterization of the measurement properties of biological variables (henceforth referred to as "physiometrics" (Moriarity & Alloy, 2021; Segerstrom & Miller, 2004)) will help improve the replicability of findings, decrease the cost of immunopsychiatry research, and improve the research to practice timeline. In this paper, several different approaches for modeling inflammatory proteins are compared to illustrate the importance of using standard measure-building procedures and to inform the use of inflammatory proteins in future research.

An important first step to this characterization is determining how to create a variable that quantifies a construct of interest. It is commonplace in immunopsychiatry to measure several inflammatory proteins and test them all as independent or dependent variables, which results in several complications. First, many immunopsychiatry theories and hypotheses are described in terms of "inflammation" (e.g., Dooley et al., 2018; Miller et al., 2009; Moriarity et al., 2018; Slavich, 2020), not individual proteins, underscoring a disconnect between theory and analysis. This approach also invites concerns with multiple comparisons that could be addressed using composite variables. To ameliorate these concerns, some studies (e.g., Moriarity, Ng, Titone, et al., 2020) have used "a priori" composite variables consisting of the sum or average of all the z-

standardized inflammatory proteins in a dataset to create a measure of general inflammation. However, this decision risks aggregating proteins that do not load strongly onto the same dimension of inflammatory physiology (e.g., pro- vs. anti-inflammatory processes) and makes the unlikely assumption that all inflammatory proteins are identically associated with a higher order inflammation construct. Assuming unidimensionality is particularly worrisome with multifaceted constructs like inflammation. If different dimensions have different associations with outcomes of interest, aggregating this variance could result in inconsistency of the size and direction of effects.

There is theoretical rationale for believing that different inflammatory proteins might be influenced by shared underlying processes. For example, they are all known to be broadly involved in the initiation and maintenance of, or recovery from, inflammatory activity (hence their classification as "inflammatory" proteins) in what is sometimes referred to as the inflammatory cascade (Cavaillon & Adib-Conquy, 2002). However, given the many component processes of inflammation (e.g., the acute phase response and upregulation of pro-inflammatory proteins, activation of the vascular and endocrine systems, neutrophil migration to the site of injury, fibrinolysis, apoptosis, coagulation, and the induction of regulatory anti-inflammatory processes to return the system to homeostasis (Gruys et al., 2005)), it is plausible that a unidimensional model might not best represent the complexity of this system. It is also important to note that many inflammatory proteins are pleotropic and can contribute to different, even opposing (e.g., pro- and anti-inflammatory), processes under certain conditions. For example, interleukin (IL)-6 typically is described in terms of proinflammatory functions (mediated by classic signaling), but also has some anti-inflammatory functions (mediated by trans-signaling) (Scheller et al., 2011). Further, it is also possible that the resting concentrations of various

proteins are partially determined by shared cellular sources (i.e., proteins produced by myokines might have higher intercorrelations than proteins produced by different types of cells).

To our knowledge, only one other study has tested the dimensionality of inflammatory proteins in a non-medical sample. Egnot and colleagues (2018) tested the dimensionality of several inflammatory and coagulatory proteins (C-reactive protein (CRP), IL-6, fibrinogen, intercellular adhesion molecule (ICAM)-1, D-dimer, Lipoprotein (Lp)-a, and pentraxin (PTX)-3) in a community-living sample. It was concluded that CRP, IL-6, and fibrinogen loaded onto an inflammatory factor and D-dimer and PTX-3 loaded onto a factor indicative of a thrombogenic process and the concomitant vascular perturbation. ICAM-1 and Lp-a did not load onto any observed factors.

A few other studies have conducted factor analyses of inflammatory proteins in medical samples, which may or may not generalize to community samples. In a sample of acute coronary syndrome patients, Tziakas et al. (2007) tested the dimensionality of CRP, fibrinogen, HDL cholesterol, IL-10, IL-18, and ICAM-1 and found three inflammatory factors: a "systemic inflammation" factor consisting of CRP and fibrinogen, a "local inflammation—endothelial dysfunction" factor consisting of IL-18 and ICAM-1, and an "anti-inflammatory factor" comprised of IL-10 and HDL cholesterol. Koukkunen et al. (2001) conducted an exploratory factor analysis (EFA) on CRP, fibrinogen, IL-6, tumor necrosis factor-α (TNF-α), troponin T, and creatine kinase MB mass in a sample of adults with unstable angina pectoris. This study described two factors: an "inflammation factor" including CRP, fibrinogen, and IL-6, and an "injury" factor including TNF-α, troponin T, and creatine kinase MB mass. Finally, Sakkinen and colleagues (2000) conducted an EFA on 21 different biological characteristics (including several inflammatory proteins) in participants with insulin resistance syndrome. Six biomarkers

(fibrinogen, CRP, plasmin-α_2-antiplasmin, Factor VIIIc, Factor IXc, and fibrin fragment D-dimer) loaded onto what was interpreted as an inflammation factor. HDL cholesterol was tested, but was not found to load onto a factor with other inflammatory markers, as was found by Tziakas et al. (2007).

Some notable similarities between these studies emerge; specifically, 1) CRP, fibrinogen, and IL-6 frequently load onto the same factor (potentially due to their interactive role in the acute phase reaction) when these proteins are included in the dataset and 2) TNF-α and ICAM-1 never loaded onto the same factor as CRP, fibrinogen, and IL-6. Thus, these studies do not support the use of "a priori" unidimensional composites created by summing standardized values of all inflammatory proteins in a dataset (consequently, giving all of the proteins equal weight in the composite) without investigating the empirical structure of the data first. McNeish and Wolf (2020) provide a thorough review of concerns about using sum scores in this manner when the structure of the data is more complex. Additional work must be done to test and replicate the structure of inflammation in populations of interest (e.g., community samples, cancer patients, individuals with HIV/AIDs) to determine the best way to aggregate different inflammatory proteins (and whether this is appropriate). It is also worth noting that two of these four studies found multiple inflammation factors, suggesting that inflammation might be best represented as a multidimensional construct. Additionally, none of these studies tested the same panel of inflammatory proteins. Thus, the direct replicability of the structure of an array of inflammatory proteins has never been tested.

The Present Study

This study sought to investigate the dimensionality of several inflammatory proteins and compare the empirically-identified factor structure to an "a priori" factor structure computed by

summing the z-standardized values of all available proteins in the dataset (a technique used in previously published studies (e.g., Moriarity, Ng, Titone, et al., 2020)). First, the structure of eight inflammatory proteins was investigated in a sample of adults with a mean age of 55.42 years. Replicability of this structure was tested in a second sample of similarly aged adults and model fit was compared to the "a priori" data structure. Structural equation modeling also was used in this replication dataset to compare 1) the empirically-identified structure, 2) the "a priori" structure, and 3) modeling the inflammatory proteins individually (without a latent variable) as outcomes of several different medical criterion variables (i.e., heart disease, diabetes, asthma, tuberculosis, thyroid disease, peptic ulcer disease, and arthritis). Internal consistencies for the empirically-identified factors and the "a priori" factor were evaluated in this dataset. Because these datasets were cross-sectional in nature, a longitudinal sample of adolescents (mean age of 16.44 years) was used for prospective analyses. Specifically, the model fit of a short-form (this dataset only included five of the proteins available in the adult datasets) version of the empirically-identified factors and the "a priori" composite were estimated and compared to establish developmental generalizability. Then, 1) the empirically-identified factors, 2) the "a priori" composite, and 3) individual proteins were tested as predictors of depression (the primary outcome of the adolescent sample's study and a psychiatric outcome associated with inflammation (Moriarity, Kautz, Mac Giollabhui, et al., 2020)). Temporal stability estimates for these inflammatory modeling options are reported.

Methods

Participants and Procedure

Participants were selected from three pre-existing datasets: Midlife in the United States (MIDUS)-2, MIDUS-R, and Project Adolescent Cognition and Emotion (ACE). MIDUS-2 (Ryff

et al., 2017) consisted of 1,255 (Mage = 55.42 years, 50% female, 78% White) participants between the ages of 25 and 75 who were fluent in English and volunteered to participate in a biomarker collection that included a sera assessment of eight inflammatory proteins (C-reactive protein (CRP), interleukin (IL)-6, IL-8, IL-10, tumor necrosis factor-α (TNF-α), fibrinogen, E-selectin, and intercellular adhesion molecule (ICAM)-1). MIDUS-R (Weinstein et al., 2017) was designed to parallel the procedure of the MIDUS-2 study, and included 863 adults (Mage = 53.53 years, 50% female, 87% white).

Project ACE (Alloy et al., 2012) is a longitudinal study of adolescent depression. Inclusion criteria for the original study were that: (i) adolescents were aged 12-13 years, (ii) their mothers were willing to participate, and (iii) adolescents self-identified as Caucasian, African-American, or biracial (examining racial differences in depression was a goal of Project ACE). Exclusion criteria were if the adolescent or mother had: (i) insufficient English reading/speaking skills to complete the assessments, or (ii) a psychotic, developmental, medical, or learning disorder. Complete details on recruitment and sample characteristics have been published (Alloy et al., 2012). A subsample nested within ACE (n = 315, Mage at first blood draw = 16.44 years, 53% female, 42% White, 53% Black, 5% Biracial) agreed to participate in a supplementary sera assessment of inflammation (CRP, IL-6, IL-8, IL-10, TNF-α). Blood draws were offered as an optional part of the protocol annually. The Children's Depression Inventory (CDI: Kovacs, 1985) was collected at these sessions as well. The dataset included a total of 866 observations with complete inflammation and CDI data across seven waves (Ns = 313, 226, 165, 101, 46, 13, 2, chronologically).

Measures

MIDUS-2 and MIDUS-R

Inflammatory proteins. Fasting blood draws were collected between 6:00 and 8:30 am for both MIDUS-2 and MIDUS-R. Blood was centrifuged and stored in a -60 °C to -80 °C freezer. Samples were shipped to the MIDUS Biocore Lab on dry ice, where they were stored at -65 °C until they were assayed. C-reactive protein (CRP) originally was analyzed in plasma via BNII nephelometer (Dade Behring Inc.). Samples falling below the assay range for this method were re-assayed using immunoelectrochemiluminescence using a high-sensitivity array kit (Meso Scale Diagnostics (MSD)). Comparisons of these two methods showed results to be highly correlated. Beginning in 2016, all participants (150 from MIDUS-R) had CRP assayed using the MSD platform using serum. Corrections to account for these changes were applied before the data were made publicly available. Fibrinogen was measured using the same BNII nephelometer. E-Selectin and Intercellular Adhesion Molecule (ICAM)-1 both were measured using ELISA assays (R&D Systems, Minneapolis, MN). Lot-to-lot changes in both E-Selectin and ICAM-1 assays were made throughout the course of the study and adjusted for prior to the data being made available to the public. Cytokines (interleukin (IL)-6, IL-8, IL-10, and tumor necrosis factor alpha (TNF-α)) were quantified by V-plex Custom Human Cytokine Kit (MSD, Rockville, MD), MSD Sulfo-tag, and MSD Sector Imager. E-Selectin and ICAM-1 values outside of the detectable range (LLOD = <.1 ng/mL and <45 mg/L, respectively) were set at .09 ng/mL and 44 ng/mL, respectively. MIDUS documentation indicates that none of the other proteins had values outside of the detectable range. Assay ranges and variability for all proteins can be found in the MIDUS documentation available online. Bivariate correlations between the proteins in MIDUS-2 and MIDUS-R can be found in Supplemental Tables 1 and 2, respectively.

C-reactive protein. CRP is a pentameric protein, generalized marker of inflammation, and acute phase reactant upregulated by IL-6 (Davidson, 2013). CRP is primarily synthesized by the liver and activates the complement system, promoting phagocytosis.

Fibrinogen. Fibrinogen is a glycoprotein complex and acute phase protein made in the liver and is upregulated by IL-6. It is involved in creating blood clots, regulating thrombin, and influencing leukocyte migration (Amrani, 1990; Davidson, 2013). Additionally, it influences the induction of cytokine/chemokine expression (e.g., IL-6 and TNF-α) via MAC-1 signaling. Breakdown products of fibrinogen (e.g., D-dimers) stimulate the release of several inflammatory proteins including CRP and IL-6 (Davidson, 2013).

E-selectin. E-selectin is a selectin cell adhesion molecule expressed by endothelial cells and activated by cytokines. Local release of IL-1 and TNF-α induces over-expression of E-selectin, which then recruits leukocytes to the site of injury (Imhof & Dunon, 1995).

Intracellular Adhesion Molecule-1. ICAM-1 concentrations increase rapidly in response to TNF-α and IL-1 and influence neutrophil adhesion (Divietro et al., 2001) and the recruitment of macrophages. It is expressed by the vascular endothelium, macrophages, and lymphocytes. There is also evidence that ICAM-1 is involved in the secretion of TNF-α (Etienne-Manneville et al., 1999).

Interleukin-6. IL-6 is responsible for stimulating acute phase protein synthesis in the liver and production/trafficking of neutrophils (Davidson, 2013; Fielding et al., 2008). It is produced by a wide variety of cells including liver cells, macrophages, osteoblasts, and monocytes (Davidson, 2013). In addition to its pro-inflammatory roles, it also has anti-inflammatory properties and is involved in the regulation of TNF-a and IL-10 (Scheller et al., 2011).

Interleukin-8. IL-8 (also known as neutrophil chemotactic factor) is a chemokine produced by macrophages and other types of cells (e.g., epithelial cells, smooth muscle cells in the airway, and endothelial cells) (Hedges et al., 2000). It is one of the most important proteins in neutrophil adhesion (both in terms of adhesion promotion and inhibition) and migration toward injury sites via chemotaxis, and also stimulates phagocytosis (Divietro et al., 2001; Dixit & Simon, 2012; Luscinskas et al., 1992).

Interleukin-10. IL-10 is predominantly produced by monocytes and lymphocytes, and primarily, is an anti-inflammatory cytokine that regulates pro-inflammatory proteins (e.g., TNF-a, IL-8, IL-6) as well as enhancing B-cell survival and antibody production (Kessler et al., 2017; Sun et al., 2009).

Tumor necrosis factor alpha. TNF-α is a predominantly proinflammatory cytokine primarily released by macrophages (in addition to other cells such as lymphoid cells, adipose tissue, and mast cells) to recruit other cells to activate immune processes (Olszewski et al., 2007). Many of the proinflammatory functions of TNF-α are apoptotic (promotes programmed cell death) in nature (Gough & Myles, 2020). TNF-α also influences migration of neutrophil to injury sites (Smart & Casal, 1994) and induces ICAM-1 (Burke-Gaffney & Hellewell, 1996).

Medical Status. Participants' medical history was assessed on the day of the study visit via interview. This interview found that 11.5/9.3 % of the sample (MIDUS-2/MIDUS-R, respectively) reported a history of heart disease, 12.4/10.9% reported a history of diabetes, 12.6/18.1 % reported a history of asthma, 0.6/0.9% reported a history of tuberculosis, 12.4/12.1% reported a history of thyroid disease, 43.4/36.4% reported a history of arthritis, and 5.4/5.2% reported experiencing a peptic ulcer.

Project ACE

Inflammatory Proteins. Blood samples were obtained via antecubital venipuncture by a certified phlebotomist into a 10 mL vacutainer designed for freezing plasma separated from the cell fraction within the vial (BD Hemogard with K2 EDTA). Vacutainers were stored in an ultracold freezer at -80 ^{o}C, and later thawed on the day of assay. Four cytokines were quantified by multi-cytokine array (IL-6, IL-8, IL-10, and TNF-α), and high-sensitivity CRP was determined in a singleplex assay, using an electrochemiluminescence platform and a QuickPlex SQ 120 imager for analyte detection (MSD, Gaithersburg, MD). Each specimen was assayed in duplicate and the intra-assay coefficients of variation averaged 5.36 and 2.29 for the cytokines and CRP, respectively. Plasma was diluted 1:2 for the cytokine assay and 1:1000 for the CRP assay. CRP is present in blood at higher concentrations, and thus, plasma was diluted to correspond to the standard curve. Values were calculated with respect to a standard curve generated from 7 calibrators with known concentrations. The lower limit of detection (LLOD) for the cytokines was 0.1 pg/mL, with a large dynamic range up to 2000 pg/mL. The LLOD for CRP was 0.1 mg/L. Values below the LLOD were set at the LLOD. Values were converted to mg/L units to be consistent with the clinical literature (Breen et al., 2011; Dabitao et al., 2011). See above for a brief description of the inflammatory proteins. Bivariate correlations between these proteins can be found in Supplemental Table 3.

Depression Symptoms. Symptoms of depression were measured using the Children's Depression Inventory (CDI; Kovacs, 1985). It consists of 27 items reflecting affective, behavioral, and cognitive symptoms of depression. Items are rated on a 0 to 2 scale and total scores range from 0 to 54. The CDI has been demonstrated to be a reliable and valid measure of depressive symptoms in youth samples (Klein et al., 2005) and has strong measurement invariance over time (Stumper et al., 2019), making it a good option for longitudinal analyses

with adolescents. Internal consistency ranged from α = .84 - .92 across each time point. The CDI was administered at all annual assessments during the same visit as the blood draws. All items in the scale were summed (after reverse scoring several items, in accordance with the manual) to create the total CDI score.

Analyses

All analyses were conducted in R 3.6.2 (R Core Team, 2013). All proteins were z-standardized for the purposes of all analyses. Outliers were identified using the median absolute deviation (MAD)-median rule and winsorized. To avoid biasing estimates toward participants with more blood draws in ACE, only the first blood draw (which had the most participants) was used to identify cutoffs for outliers. No proteins were skewed following winsorization.

Exploratory Factor Analysis

"Bass-ackwards" exploratory factor analyses (Goldberg, 2006) were conducted with all eight proteins in MIDUS-2 using the Geomin rotation to allow for a hierarchical factor structure and correlated factors using *efaUtilities* (Zhang et al., 2020). Parallel analysis was used to determine the number of factors to retain using *EFA.dimensions* (Connor, 2020). Unlike other factor retention methods, parallel analysis allows for correction for the effects of sampling error. Eigenvalues were generated using permutations of the raw dataset to create eigenvalues that could be compared to the eigenvalues produced by the factor analysis on the dataset. When an eigenvalue generated from the factor analysis is higher than that generated from the parallel analysis, it can be assumed that the eigenvalue represents a real factor that accounts for more variance than a factor based on the permutated data (Horn, 1965). Thus, the number of factors retained was determined by the number of eigenvalues for which the eigenvalue for the factor analysis is higher than the eigenvalue of the parallel analysis using the permutations of the

dataset. Complex loadings (i.e., items loading onto more than one factor) were allowed due to the pleotropic nature of these biomarkers. Loadings at or above .30 were considered to load onto a factor; however, if a protein was close to this threshold (above .20), it was included in an alternative CFA (described below) and model fits were compared. Proteins that did not meet this threshold for any factor were removed and the EFA re-estimated. Factor scores were calculated using the regression method. Models were estimated using maximum likelihood.

Confirmatory Factor Analysis/Structural Equation Modeling

Confirmatory factor analyses (CFAs) were estimated in MIDUS-R for the factor structure found in the EFA from MIDUS-2. CFAs were conducted in *lavaan* (Rosseel, 2012). The variance of latent variables was set to 1 and all factor loadings, unless otherwise noted in the results, were freely estimated. Alternative models with proteins with factor loadings between .20 and .30 also were estimated and model fit was compared to select the final "empirically-identified" factor structure. Next, model fit of the empirically-identified factor structure was compared to the "a priori" unidimensional factor structure with the factor loadings of all proteins constrained to equality. In addition to interpretation of standard goodness-of-fit indices (i.e., good fit indicated by comparative fit index [CFI] ≥ .95, root mean square-error of approximation [RMSEA] ≤ .06, and standardized root-mean-square residual [SRMR] ≤ .08, and non-significant chi-square test; Hu et al., 2009), AIC and BIC will be compared between models. Unlike most other fit indices, AIC and BIC can be directly compared, with lower values indicating preferable model fit. Additionally, it is important to note that the chi-square test of significance is over-powered in large sample sizes (Bollen, 1989). SEM models also were conducted in MIDUS-R comparing the empirically-identified structure, "a priori" structure, and individual inflammatory proteins as outcomes of several medical criterion variables (heart disease, diabetes, asthma,

tuberculosis, thyroid disease, peptic ulcer disease, and arthritis) and model fit was compared again. Finally, CFAs of the empirically-identified and "a priori" models were estimated in ACE to test generalizability to a more racially-diverse, adolescent sample. Three of the proteins used in MIDUS were not measured in ACE, so this is analogous to testing a short form of a self-report measure. Missing data was handled with maximum likelihood estimation.

Predictive Validity—Depression Symptoms

Mixed linear models using *lme4* (Bates et al., 2015) tested the predictive validity of i) the empirically-identified weighted composites, ii) the "a priori" composite, and iii) the individual proteins for depression symptoms at the time of blood draw in random intercept, fixed slope models. Multilevel modeling was chosen over standard regression because the data were clustered within individuals, which would result in greater probability of Type I error and less efficient estimates of coefficients compared to multilevel modeling. Models were estimated using restricted maximum likelihood. These analyses used a total of 866 observations across seven waves (Ns = 313, 226, 165, 101, 46, 13, 2, chronologically).

Reliability and Temporal Stability

Factor reliability was quantified via coefficient ω (Raykov, 2001) using the R package *semTools* (Jorgensen et al., 2021) in MIDUS-2. Reliability for higher-order latent variables will be quantified using $\omega_{partial}$, which indicates the proportion of observed variance explained by a higher-order factor after partialing out the lower-order factors. Coefficient ω was selected over Cronbach's α because a) ω makes fewer and more realistic assumptions compared to α and b) problems with misestimation of reliability are far less likely. For a more in-depth argument for the adoption of ω over α as the new field standard refer to Dunn et al. (2014).

Temporal stability was calculated two ways: Pearson's *r* and intra-class correlation coefficients (ICCs). Importantly, temporal stability estimates are only informative for the time between data points they were estimated under. Time between blood draws in Project ACE were highly variable due to the optional nature of the blood draw (.13 – 70.70 months), so observations only were chosen that fell into "waves" of data collection at increments of 11.5 months (the modal duration between observations) +/- 3 months. This resulted in a number of observations being excluded due to not fitting in these waves. Pearson correlations were calculated between the first two waves (N = 97). ICCs were calculated using random intercept, fixed slope hierarchical linear models with weighted composite scores across seven waves (Ns = 187, 97, 96, 76, 48, 15, 6, chronologically). Pearson correlations reflect stability of ordinal rankings, whereas ICCs indicate the within-person variance relative to between-person variance.

Results

Exploratory Factor Analysis

The initial parallel analysis indicated two lower-order factors (i.e., two factors with eigenvalues greater than the corresponding factors in the random data) in MIDUS-2. The initial EFA found that E-selectin did not load onto any factors, and thus, it was removed (EFA results can be found in Table 1). The new parallel analysis still indicated two factors. CRP, IL-6, and fibrinogen loaded onto Factor 1A above the cutoff of .30. TNF-α, IL-8, IL-10, and ICAM-1 loaded onto Factor 2A above .30 (Table 1). IL-6 had a subthreshold (.22) cross-loading on Factor 2A. These factors accounted for 20% and 16% of the total variance, respectively. In comparison, a single factor model only accounted for 24% of the total variance. The two-factor model also fit the data better than a single factor model (AIC/BIC = 42027.253/42145.153 vs 42323.654/42431.302, respectively). Both first order factors strongly correlated with a higher

order factor (Factor 1B, loadings = .84 and .53, respectively), suggesting a multidimensional, hierarchical factor structure (Loehlin & Goldberg, 2014).

Table 1. Exploratory Factor Analyses (EFAs) in MIDUS-2

	Initial EFA		Without E-selectin	
	Factor 1A	Factor 2A	Factor 1A	Factor 2A
CRP	.78	.00	.78	.00
Interleukin-6	.58	.22	.58	.22
Tumor Necrosis Factor-α	.03	.75	.02	.78
Interleukin-8	.02	.34	-.02	.32
Interleukin-10	.05	.52	-.05	.50
Fibrinogen	.68	.06	.67	-.06
Intercellular Adhesion Molecule-1	.11	.30	.11	.30
E-selectin	.17	.11	x	x
Correlation Between Factors	.33		.33	
Proportion of Variance Explained	.18	.14	.20	.16
Proportion of Variance Explained by Single Factor	.22		.24	

Note: x = not included in EFA

Confirmatory Factor Analyses/Structural Equation Modeling

First, the factor structure obtained using a .30 loading threshold for the exploratory analyses was compared to an identical model except with IL-6 loading onto both first-order factors in MIDUS-R (Figures 1a and 1b, respectively). In all CFA models, factor loadings between the two first-order factors (Factors 1A and 2A) and the second-order factor (Factor 1B) were constrained to be equal because a latent variable needs 3 indicators to be properly identified and constraining the loadings to equality frees up degrees of freedom to identify the model. The model with IL-6 loading onto both first order factors had lower AIC/BIC (Table 2: 15909.725/16018.839 vs 15943.590/16047.960, respectively), suggesting better model fit. The chi-squared difference test that compared model fit found that this difference in model fit was significant ($\Delta\chi^2 = 35.865$, $p < .001$). This model demonstrated good model fit with a CFI = .970, RMSEA = .055 (90% CI: .037-.073), and SRMR = .036. The chi-square test was significant ($\chi^2(12) = 42.331$, $p < .001$), but this is uninformative with sample sizes this large. Additionally, the EFA was well-replicated at the factor loading level as well. All of the originally estimated factor loadings were between the 95% confidence intervals for the factor loadings in the replication models (Table 3).

Second, this empirically-identified factor structure was compared to the "a priori" factor structure with all proteins having an equal weight on a single factor (Figure 1). AIC/BIC are not comparable for models with different numbers of indicators; thus, E-selectin was not included in this model despite this decision being informed by the results of the original EFA. Model fit for the "a priori" model was poor: CFI = .601, RMSEA = .155 (90% CI = .142-.167), and SRMR = .119. The chi-square test was significant ($\chi^2(12) = 425.386$, $p < .001$). The empirically-identified

model also had a lower AIC/BIC (15909.73/16018.84 vs 16276.780/16347.940, respectively), indicating worse fit for the "a priori" model.

Third, several medical conditions associated with inflammation (heart disease, diabetes, asthma, tuberculosis, thyroid disease, peptic ulcer disease, and arthritis) were modeled as predictors of a) the empirically-identified factors (Figure 2), b) the "a priori" factor (Figure 3), and c) individual proteins (modeled without a latent variable; Figure 3) in separate models. Note that the first-order and second-order factors in the empirically-identified model could not be entered as outcomes in the same model because Factor 1B was entirely composed of variance from Factors 1A and 2A; thus, two models had to be estimated (the only difference being which level of the structure was predicted by the medical diagnoses). The first-order and second-order empirically-identified models (Table 2: AIC/BIC = 14617.221/14789.709; 14612.368/14752.223, respectively) and individual protein models (AIC/BIC = 14586.231/14977.827) all outperformed the "a priori" model (AIC/BIC = 14958.274/15060.834). Both empirically-identified models had better BIC, but worse AIC than the individual protein model, suggesting no clear answer as to which fit the data better. Because the individual protein model was just-identified, no other fit statistics were generated. Model fit for both the first-order and second-order empirically-identified models were generally good (RMSEA = .046 (90% CI = .036-.056), SRMR = .034; RMSEA = .044 (90% CI = .034-.053), SRMR = .036, respectively), except the CFI was slightly below the ideal cutoff of .95 (CFI = .926 and .924, respectively). As expected with this sample size, both chi-square tests were significant ($\chi^2(47) = 124.989$, $p < .001$; $\chi^2(54) = 131.136$, $p < .001$, respectively). Associations between the inflammation variables and medical conditions are described below in the section "Predictive Validity".

Finally, the empirically-identified and "a priori" factor structures were partially recreated (using just CRP, IL-6, IL-8, IL-10, and TNF-α) in the more racially-diverse, longitudinal dataset of ACE adolescents. This is analogous to validating a "short-report" version of a self-report questionnaire. Note that, because Factor 1A only had two indicators (CRP and IL-6), these loadings had to be constrained to equality to fit the model in this dataset. Similar to the MIDUS-R dataset, the empirically-identified model outperformed the "a priori" model (Table 2: AIC/BIC = 4290.608/4350.649 vs. 4405.040/4446.319). Model fit for the "a priori" model was poor: CFI = .395, RMSEA = .209 (90% CI = .178-.241), SRMR = .148, $\chi^2(9) = 132.672$, $p < .001$. The empirically-identified dataset fit the data well with a nonsignificant chi-square test ($\chi^2(4) = 8.239$, $p = .083$), CFI = .979, RMSEA = .058 (90% CI = .000-.115), and SRMR = .033. The only parameter suggesting potentially poor fit was the upper bound of the RMSEA 90% CI was slightly above the ideal cutoff of .100. However, despite global indicators of good fit, the original factor loading estimates did not consistently fall within the 95% CIs for the CFA. Given general support for the empirically-identified model found in MIDUS-2, the predictive validity and temporal stability of these factors were tested in ACE.

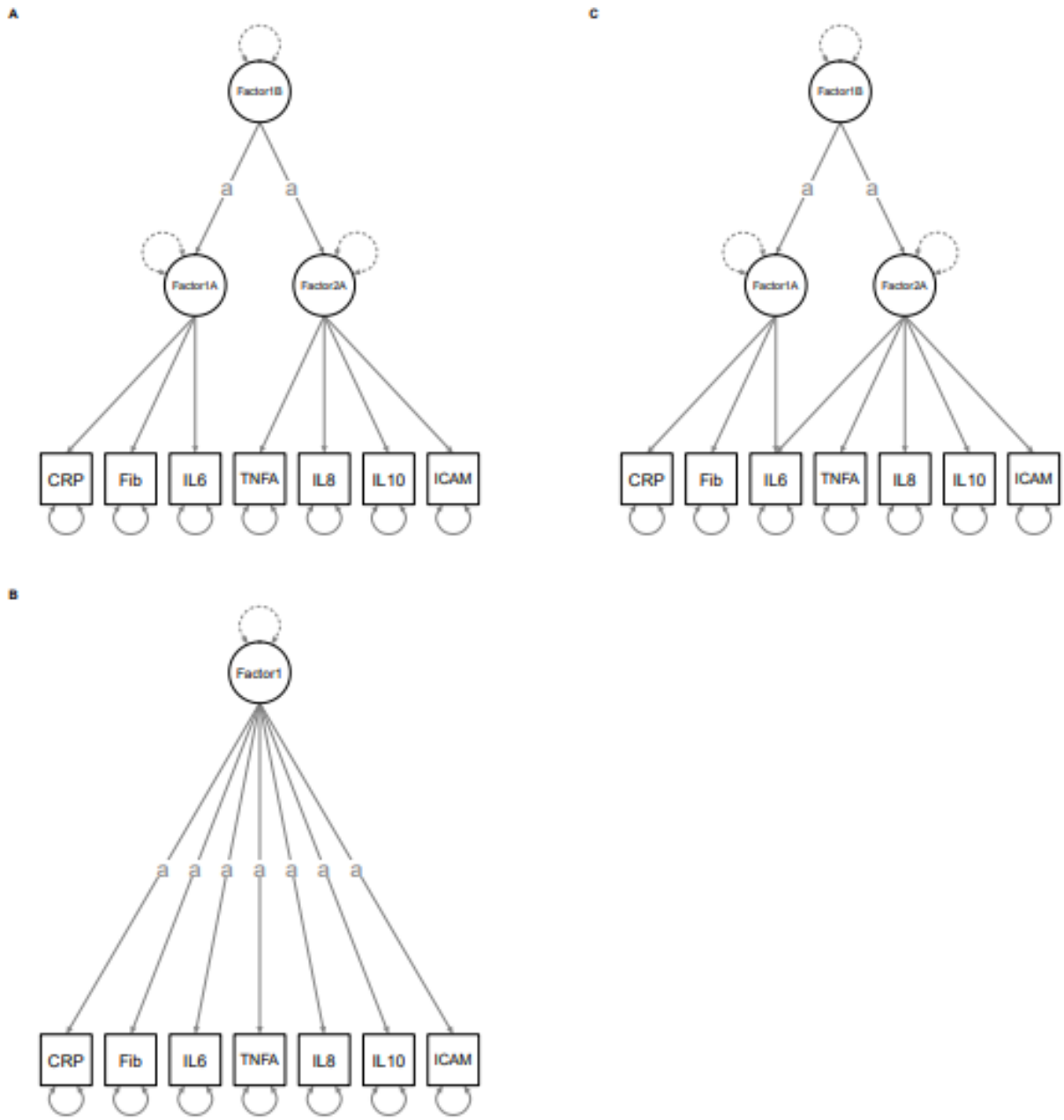

Figure 1. Structural Models of Inflammation. A: Empirically-identified structure without IL-6 cross-loading. B: Empirically-identified model with IL-6 cross-loading. C: A priori structure. "a" denotes loadings constrained to equality, z= z-standardized, CRP = C-reactive Protein, Fib=fibrinogen, IL = interleukin, TNFA= Tumor Necrosis Factor-α, ICAM = Intracellular Adhesion Molecule-1.

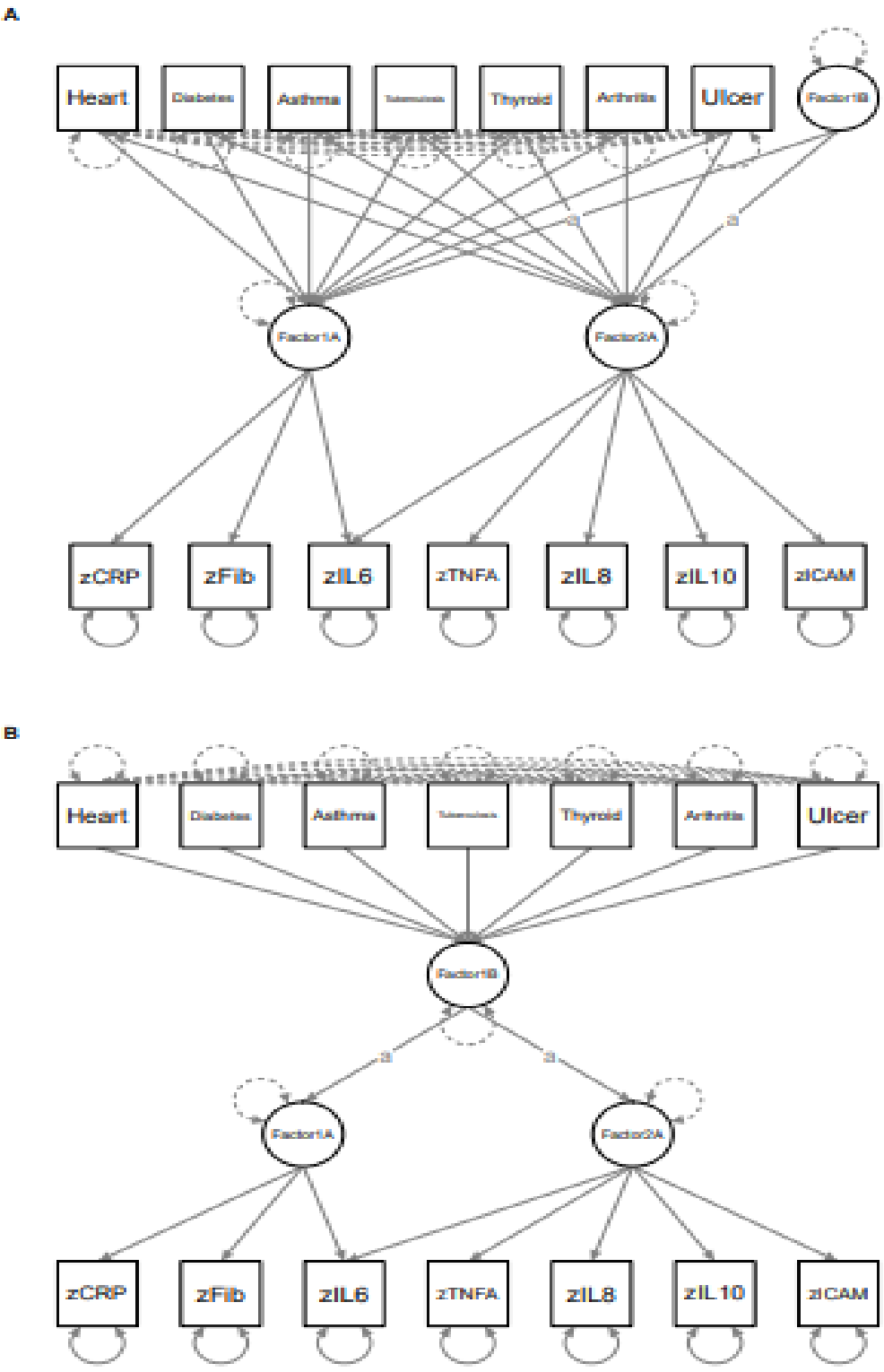

Figure 2. Empirically-identified Structural Equation Models with Medical Criterion. A: Empirically-identified structure: Medical predictors of 1st order factors. B: Empirically-identified structure: Medical predictors of 2nd order factors. "a" denotes loadings constrained to equality, z= z-standardized, CRP = C-reactive Protein, Fib=fibrinogen, IL = interleukin, TNFA= Tumor Necrosis Factor-α, ICAM = Intracellular Adhesion Molecule-1.

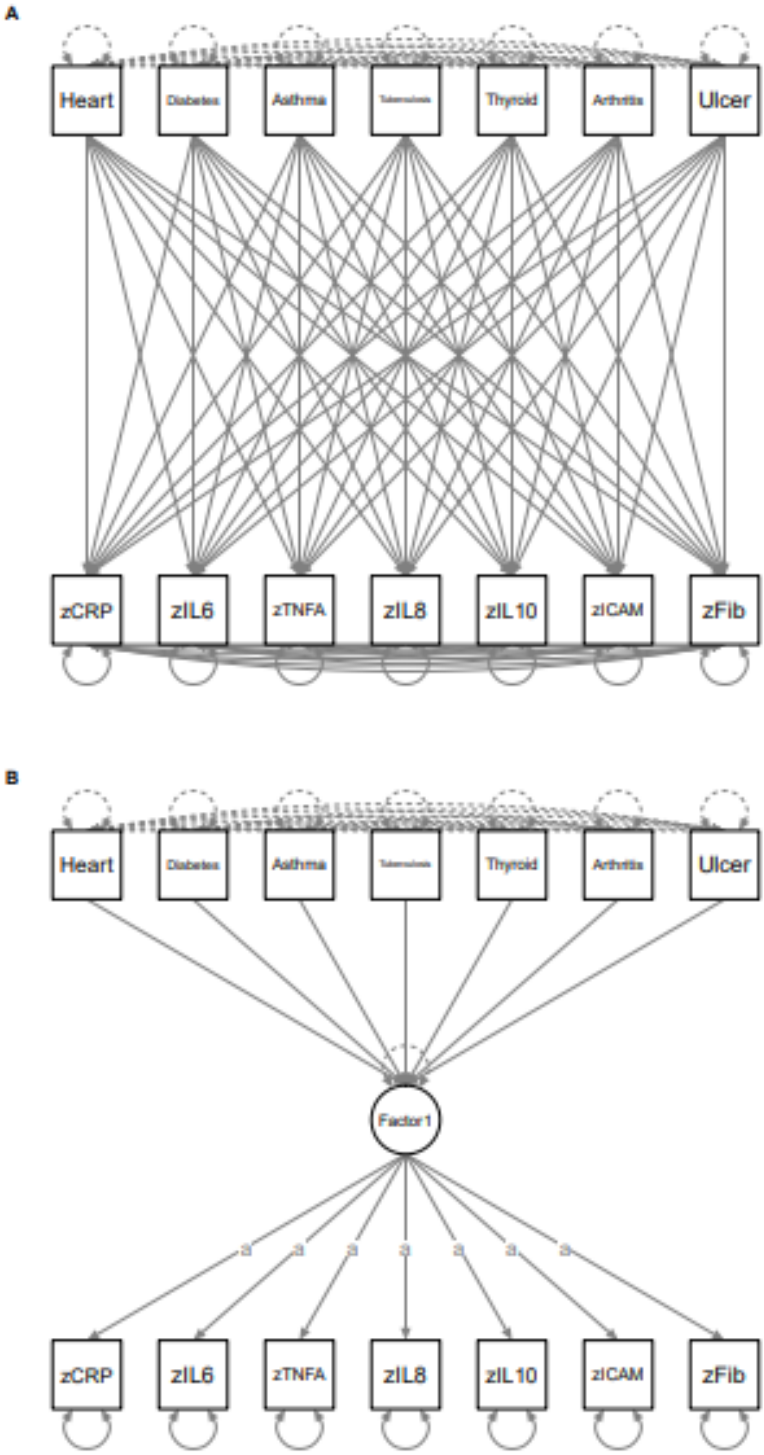

Figure 3. Other Structural Equation Models with Medical Criterion. A. Medical predictors of individual proteins. B. Medical predictors of a priori factor. "a" denotes loadings constrained to equality, z = z-standardized, CRP = C-reactive Protein, Fib=fibrinogen, IL = interleukin, TNFA= Tumor Necrosis Factor-α, ICAM = Intracellular Adhesion Molecule-1.

Table 2. Fit Statistics of Different Inflammatory Models

	χ^2	χ^2 df	p_{χ^2}	CFI	RMSEA	90% CI RMSEA	SRMR	AIC	BIC
MIDUS-R CFAs									
A priori	425.386	20	.001	.601	.155	.142-.167	.119	16276.780	16347.940
Empirically-identified (no IL-6 cross-loading)	78.196	13	.001	.936	.077	.061-.094	.042	15943.590	16047.960
Empirically-identified	42.331	12	.001	.970	.055	.037-.073	.036	15909.725	16018.839
MIDUS-R SEMs with Medical Criteria									
A priori	496.042	62	.001	.588	.095	.087-.102	.073	14958.274	15060.834
Empirically-identified: 1st order	124.989	47	.001	.926	.046	.036-.056	.034	14617.221	14789.709
Empirically-identified: 2nd order	131.136	54	.001	.924	.044	.034-.053	.036	14612.368	14752.223
Individual proteins	0.000	0	N/A	1.000	0.000	0.000	0.000	14586.231	14977.827
ACE CFAs									
A priori	136.672	9	.001	.395	.209	.178-.241	.148	4405.040	4446.319
Empirically-identified	8.239	4	.083	.979	.058	.000-.115	.033	4290.608	4350.649

Note: χ^2= Chi-squared, df = degrees of freedom, p = p-value, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation, CI = Confidence Interval, SRMS = Standardized Root Mean Square Residual, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion, MIDUS-R = Midlife in the United States-Replication, ACE = Adolescent Cognition and Emotion, CFA= Confirmatory Factor Analysis, SEM = Structural Equation Model, IL-6 = Interleukin-6

Table 3. Protein Loadings

	MIDUS-2 EFA		MIDUS-R CFA		ACE CFA	
	Factor 1A	Factor 2A	Factor 1A	Factor 2A	Factor 1A	Factor 2A
CRP	.78		**.78** **90% CI = .71-.85**		**.71*** **90% CI = .62-.80**	
Interleukin-6	.58	.22	**.59** **90% CI = .52-.67**	**.23** **90% CI = .16-.31**	.71* 90% CI = .62-.80	**.18** **90% CI = .05-.31**
Tumor Necrosis Factor-α		.78		**.80** **90% CI = .71-.90**		.59 90% CI = .41-.77
Interleukin-8		.32		**.34** **90% CI = .27-.42**		.04 90% CI = -.10-.18
Interleukin-10		.50		**.47** **90% CI = .40-.55**		.75 90% CI = .54-.96
Fibrinogen	.67		**.69** **90% CI = .62-.76**		x	x
Intercellular Adhesion Molecule-1		.30		**.31** **90% CI = .23-.39**	x	x

Note: * = constrained to equality, x = unavailable in dataset. Loadings < .20 not depicted. CFA confidence intervals that include the original EFA estimate are bolded. MIDUS = Midlife in the United States, EFA = Exploratory Factor Analysis, CFA = Confirmatory Factor Analysis, CI = Confidence Interval, ACE = Adolescent Cognition and Emotion

Predictive Validity

Because of poor model fit (Table 2), the associations between the medical conditions and the "a priori" inflammation variable are not reported. The associations between the medical conditions and a) the empirically-identified factors and b) the individual proteins are reported in Table 4. All three empirically-identified inflammatory factors were significantly associated with a history of diabetes (Factor 1A b =.936, SE = .166, p < .001; Factor 2A b =.596, SE = .174, p = .001; Factor 1B b =1.355, SE = .260, p < .001) and arthritis (Factor 1A b =.256, SE = .106, p = .016; Factor 2A b =.284, SE = .113, p = .012; Factor 1B b =.470, SE = .151, p = .002). Additionally, Factor 1A was associated with asthma (b =.272, SE = .129, p = .035) and both Factor 2A and 1B were associated with a history of thyroid disease (Factor 2A b =.333 SE = .161, p = .038; Factor 1B b =.417, SE = .207, p = .044); however, these three results (33% of significant results) were no longer significant after family-wise (grouped by independent variable) Benjamini-Hochberg false-discovery rate corrections (BH-FDR).

In the model with the medical conditions predicting individual proteins, diabetes was significantly associated with CRP (b =.556, SE = .114, p < .001), IL-6 (b =.566, SE = .113, p < .001), TNF-α (b =.335, SE = .115, p = .004), and IL-8 (b =.613, SE = .112, p < .001) and arthritis was associated with IL-6 (b =.235, SE = .075, p = .002), IL-8 (b =.340, SE = .074, p < .001), and fibrinogen (b =.208, SE = .075, p = .005). These seven results were robust to BH-FDR corrections. The following seven associations (50% of significant results), were no longer significant after BH-FDR corrections: heart disease was associated with IL-6 (b =.257, SE = .127, p = .042), both asthma and tuberculosis were associated with CRP (b =.183, SE = .092, p = .048 and b =.791, SE = .371, p = .033, respectively), thyroid disease was associated with IL-8 (b =.214, SE = .106, p = .043), arthritis was associated with TNF-α (b =.156, SE = .076, p = .041),

and ulcers were associated with IL-6 and ICAM-1 (b =.378, SE = .159, p = .018 and b =.340, SE = .164, p = .038, respectively).

Four hierarchical linear models were estimated with inflammation predicting contemporaneous depression in Project ACE. Following the same logic described above, separate "first-order" and "second-order" models were estimated for the empirically-identified structure. The model with the five available proteins modeled as separate predictors in a single model had no significant associations with depression (CRP b = .224, SE = .199, p = .261; IL-6 b = .186, SE = .211, p = .380; IL-8 b = -.212, SE = .191, p = .268; IL-10 b = -.417, SE = .218, p = .056; TNF-α b = -.101, SE = .219, p = .628). Neither the unidimensional "a priori" aggregate nor the second order (Factor 1B) aggregate were significantly associated with depression (a priori b = -.065, SE = .061, p = .289; Factor 1B b = -.023, SE = .134, p = .864). Both first-order factors were significantly associated with depression, in opposite directions (Factor 1A b = .355, SE = .172, p = .039; Factor 2B b = -.391, SE = .149, p = .009).

Table 4. Medical Disorders Predicting Inflammatory Outcomes

	Heart	Diabetes	Asthma	Tuberculosis	Thyroid	Arthritis	Ulcer
	Coefficient (SE) *p*	Coefficient (SE) *p*	Coefficient (SE) *p*	Coefficient (SE) *p*	Coefficient (SE) *p*	Coefficient (SE) *p*	Coefficient (SE) *p*
Empirically-identified Factors							
Factor 1A	.241 (.178) .177	**.936 (.166)** **<.001**	**.272 (.129)** **.035**	.593 (.519) .254	.156 (.150) .298	**.256 (.106)** **.016**	.057 (.224) .801
Factor 2A	.243 (.188) .197	**.596 (.174)** **.001**	.041 (.136) .764	-.760 (.550) .167	**.333 (.161)** **.038**	**.284 (.113)** **.012**	.255 (.239) .287
Factor 1B	.423 (.244) .083	**1.355 (.260)** **<.001**	.284 (.176) .105	-.078 (.697) .911	**.417 (.207)** **.044**	**.470 (.151)** **.002**	.260 (.304) .391
Proteins Modeled Individually							
zCRP	.087 (.128) .497	**.556 (.114)** **<.001**	**.183 (.092)** **.048**	**.791 (.371)** **.033**	.093 (.108) .389	.085 (.075) .258	-.04 (.161) .803
zIL-6	**.257 (.127)** **.042**	**.566 (.113)** **<.001**	.123 (.091) .179	-.051 (.368) .890	.153 (.107) .153	**.235 (.075)** **.002**	**.378 (.159)** **.018**
zTNF-α	.195 (.129) .130	**.335 (.115)** **.004**	.058 (.093) .535	-.287 (.374) .443	.158 (.109) .146	**.156 (.076)** **.041**	.063 (.162) .063
zIL-8	.216 (.125) .085	**.613 (.112)** **<.001**	-.042 (.091) .644	-.426 (.364) .242	**.214 (.106)** **.043**	**.340 (.074)** **<.001**	-.034 (.158) .828
zIL-10	-.062 (.131) .637	.143 (.117) .223	-.038 (.095) .685	-.665 (.381) .081	.177 (.111) .110	.051 (.077) .514	.043 (.165) .796
zICAM-1	-.168 (.130) .199	.081 (.116) .485	.045 (.094) .635	-.020 (.379) .959	.144 (.110) .189	-.065 (.077) .400	**.340 (.164)** **.038**
zFib	.149 (.126) .238	.599 (.113) <.001	.153 (.091) .093	-.099 (.367) .788	096 (.106) .368	**.208 (.075)** **.005**	-.129 (.159) .415

Note: **bolded** = significant at an alpha-level of .05 (uncorrected), slashed cell = Benjamini-Hochberg correction shifted significant result to non-significant, SE = standard error, *p* = *p*-value, z = z-standardized, CRP = C-reactive Protein, IL = interleukin, TNF-α= Tumor Necrosis Factor-α, ICAM-1 = Intracellular Adhesion Molecule-1, Fib=fibrinogen.

Reliability and Temporal Stability

According to standard interpretive thresholds, reliability was questionable for the "a priori" factor structure (ω = .690). It was acceptable for Factor 1A (ω = .750) and poor for both Factor 2A (ω = .579) and Factor 1B ($\omega_{partial}$ = .579). Pearson correlations found the 8.5-14.5-month rank order stability to be consistently higher for all three empirically-identified factors compared to the "a priori" factor (a priori *r* = .242, Factor 1B *r* = .314, Factor 1A *r* = .471, Factor 2A *r* = .286). Rank order stability was variable for the individual proteins (CRP *r* = .473; IL-6 *r* = .374; IL-8 *r* = .205; IL-10 *r* = .363; TNF-α *r* = .384). Longitudinal ICCs for Factor 1A = .534, Factor 2A = .235, Factor 1B = .405, the "a priori" factor = .317, CRP = .457, IL-6 = .450, IL-8 = .164, IL-10 = .343, and TNF-α = .222.

Discussion

Most studies investigating inflammation in relation to medical or psychiatric illnesses test a variety of individual proteins as predictors and/or outcomes. This invites problems with multiple comparisons and creates a disconnect between theories about generalized inflammation and the analyses conducted. Alternatively, some studies use composite variables created without first investigating the appropriateness of this decision in what we describe as "a priori" composites. Results suggested that seven of the eight proteins available in the MIDUS datasets were best organized into a two-level factor structure with a general "inflammation" factor and two lower-order subfactors. In all comparisons made, this structure fit the data better than an "a priori" structure in which all inflammatory proteins are equally associated with a single factor, an approach that has been used previously (e.g., Moriarity, Ng, Titone et al., 2020). Compared to the "a priori" composite, empirically-identified factors and modeling proteins individually

(without an aggregate variable) consistently fit the data better (in fact, model fit was not acceptable for the “a priori” composite in all models tested).

Interestingly, fit criteria diverged on whether the empirically-identified structure fit the model with external criteria better than modeling the proteins individually. Specifically, BIC supported the empirically-identified structures and AIC supported individual markers. Unfortunately, because all variables in the individual protein model were observed variables, it is impossible to evaluate other fit statistics. Given the large sample size and the fact that BIC, but not AIC, is a consistent metric (i.e., the probability of selecting the true model approaches 100% as the sample size increases to infinity), it might be that the empirically-identified factors are preferable (see Dziak et al., 2020 for a more thorough comparison of information criteria). However, these findings are far from conclusive and future research is necessary. Additionally, careful consideration should be given to whether variable aggregation is a theoretically-appropriate course of action (i.e., if specific biomarkers are theorized to drive effects rather than generalized inflammation). Further, this study adopted a latent modeling approach to form aggregates; it is possible that other types of causal modeling (e.g., causal indicator models, network models, multiple-indicator multiple-cause (MIMIC) models) might be more suitable for inflammatory proteins.

In the empirically-identified structure, the inflammatory factor including CRP, IL-6, and fibrinogen was interpreted to reflect proteins involved in acute phase reaction processes localized in the liver. Both CRP and fibrinogen are acute phase reactants synthesized in the liver and upregulated by IL-6 (some of which is also made by liver cells, although it is produced by a variety of cell types) in response to acute stress and inflammatory challenges (Amrani, 1990; Davidson, 2013). Further, breakdown products of fibrinogen (e.g., D-dimers) can upregulate

both IL-6 and CRP, creating a positive feedback loop (Davidson, 2013). Additionally, this factor might represent chronic inflammatory processes (the temporal stability of this factor was higher than any other inflammatory variable assessed). Interestingly, these three proteins also consistently loaded onto the same factor in prior research in a community sample of older adults (Egnot et al., 2018) and a sample of adults with unstable angina pectoris (Koukkunen et al., 2001). Additionally, although IL-6 was not measured, CRP and fibrinogen also loaded onto the same factor in a sample of patients with insulin resistance syndrome (Sakkinen et al., 2000) and acute coronary syndrome (Tziakas et al., 2007). Also consistent with this study, ICAM-1 was measured, but did not load onto this common factor in Egnot et al. (2018) or Tziakas et al. (2007). Similarly, TNF-α also was measured, but did not load onto this factor in Koukkunen et al. (2001). Consequently, there is evidence that this grouping is not an artifact of the proteins available in this study or other study-specific methods, demonstrating both internal replicability in this study and external replicability with previous studies.

The second inflammatory factor, consisting of IL-6, TNF-α, IL-8, IL-10 and ICAM-1 was interpreted to reflect a more general dimension of inflammation that might be related to more acute processes (both pro- and anti-inflammatory) particularly associated with neutrophil activity (Korthuis et al., 1994). IL-6 is a key regulator of neutrophil trafficking (Fielding et al., 2008), IL-8 and ICAM-1 can influence neutrophil adhesion (Divietro et al., 2001), and both TNF-α and IL-8 help neutrophils migrate to injury sites (Dixit & Simon, 2012; Smart & Casal, 1994). IL-6 also is involved in the regulation of TNF-α and IL-10, which might account for some of the shared variance between these indicators. Similarly, TNF-α often is co-released with IL-6. TNF-α also induces ICAM-1 (Burke-Gaffney & Hellewell, 1996), which, in turn, might increase TNF-α secretion (Etienne-Manneville et al., 1999) in a positive feedback loop. IL-10 serves as an

important mediator of neutrophil activity via regulation of proinflammatory cytokines (e.g., TNF-a and IL-6), and CXC keratinocyte-derived chemokine molecules (e.g., IL-8; Kessler et al., 2017; Sun et al., 2009). Further, given the different functions these factors are interpreted to reflect and differences in their temporal stability over time, it might be that this factor reflects more acute processes compared to the CRP, IL-6, and fibrinogen factor. Future research testing changes in these inflammatory factors in response to acute inflammatory challenges is necessary to investigate this possibility further.

It is critical to underscore the pleotropic nature of many inflammatory proteins and the complexity of inflammatory processes. For example, IL-6, commonly conceptualized as a "pro-inflammatory" protein, has several anti-inflammatory functions that are most likely dependent on classic-signaling via the membrane-bound non-signaling α-receptor IL-6R (Scheller et al., 2011). Given the contextual functioning of the immune system, shared variance analyses might be more informative in the context of acute inflammatory activity compared to the resting data used in this study. Truly, due to the complexity of this system and comparable model fit to the empirically-indicated factor structure, it might be more appropriate to analyze specific proteins and plan around the inherent complications of doing so. However, should the decision be made to use aggregate variables of inflammation, the results of this study consistently discourage the use of aggregates in which all proteins are equally weighted (throughout referred to as an "a priori" composite). Instead, standard aggregate measure-building procedures should be utilized (for more detailed concerns about the use of straight sum scoring to create aggregates when not all component variables are equally associated with the construct of interest, see McNeish & Wolf (2020)). Until a greater understanding of how to model inflammation is achieved,

immunopsychiatry might benefit from simultaneously testing multiple scales of inflammatory measurement (Moriarity & Alloy, 2020).

If a choice between modeling individual proteins and empirically-supported composites is made, this decision should be made prior to data collection to facilitate complementary study design (e.g., sample size, time between observations in longitudinal studies) with the physiometrics of the variables to be used in mind. Armed with this information, researchers can design studies that require fewer resources (money, participants, repeated measures), with greater odds of finding replicable effects, and the research to intervention pipeline can be optimized (Moriarity & Alloy, 2021).

These empirically-identified factors, individual proteins, and the "a priori" factor were also tested in models with medical and psychiatric outcomes of interest. The associations between medical conditions and the "a priori" factor were not interpreted because of poor model fit. After using BH-FDR corrections, the same two medical conditions (diabetes and arthritis) predicted inflammatory outcomes in both the 1) empirically-identified factors and 2) individual proteins models. However, the use of the empirically-identified factors resulted in fewer results shifting from significance to non-significance post-correction (33% vs. 50%). When predicting depression symptoms, none of the individual proteins, the "a priori" composite, or the 2nd order empirically-identified composite (Factor 1B) were significantly associated with depression symptoms. However, the two empirically-identified first-order factors were associated with depression symptoms in opposite directions. Thus, aggregating these sources of variance would wash each other out, increasing the risk for false negatives and missing nuances in the relationship between inflammation and depression. Specifically, the "short-form" Factor 1A (CRP and IL-6) was positively associated with concurrent depression and "short-form" Factor

2A (IL-8, IL-10, TNF-α, and IL-6) was negatively associated with depression. Should these factors represent more chronic vs. acute inflammatory processes, these opposing associations could be explained by the theory that inflammatory risk for depression symptoms is characterized by chronic low-grade inflammation (i.e., Factor 1A) as opposed to acute-reactions (which are a necessary part of biological reactivity to illness and injury). Further, all of the proteins in the "short-form" Factor 2A (i.e., IL-10, TNF-α, IL-6, IL-8) have functions that regulate pro-inflammatory processes (Kessler et al., 2017; Luscinskas et al., 1992; Scheller et al., 2011; Zakharova & Ziegler, 2005), which could explain an inverse association with depression symptoms.

This study has several important strengths. First, multiple sizeable datasets were used for replication featuring a combination of cross-sectional and longitudinal data allowing for a more thorough physiometric characterization of the variables under study. Second, the samples featured important differences in sample characteristics (i.e., age, race) supporting the generalizability of the empirically-identified factor structure to different demographic groups. Third, the fact that all three samples were community samples maximizes generalizability to other non-medical samples. Fourth, these three datasets allowed for both "complete" replication and a "conceptual replication" in which study methods differed. Indeed, MIDUS-R specifically was designed as a replication dataset for MIDUS-2, maximizing comparability of methods. This was complemented by the generally good replication in ACE, which did not use the same methods to measure biomarkers, suggesting that the results are not completely contingent on study methods.

However, this study must be interpreted in the context of several limitations. First, although eight proteins are more than most psychoneuroimmunology studies, there are many

inflammatory markers that were not included (e.g., IL-1β, T-cells, B-cells). Thus, the empirically-identified structure might look different when a broader array/different proteins are used. However, the finding that CRP, IL-6, and fibrinogen loaded onto one factor (which did not include ICAM-1 or TNF-α) is consistent with previous investigations of the dimensionality of inflammation in medical and community samples (e.g., Egnot et al. (2017)), supporting the generalizability of this factor. Second, the number of identifiable latent factors is constrained by the number of indicator variables. Consequently, a study with more than eight indicator variables might find a more multidimensional factor structure. Third, although a hierarchical structure of inflammation is theoretically plausible, it was impossible to statistically compare a hierarchical structure with two lower-order factors to a model with just two lower-order factors because they are equivalent models with alternative parameterizations. Fourth, the adolescent dataset used to test temporal stability and predictive validity only had five of the eight proteins included in the MIDUS datasets, forcing the investigation of a "short-form" aggregate. This concern is ameliorated somewhat by generally good model fit for the subset of inflammatory proteins, but future research should replicate these tests in a sample with all eight proteins. Additionally, despite good model fit, the original factor loadings (particularly IL-8 onto Factor 2A) in MIDUS-2 did not consistently fall in the confidence intervals of the CFA in ACE. It is important to note that this level of comparison might be more sensitive to missing two indicators (one from each first-order factor). These differences also could be due to differences in sample characteristics (e.g., developmental (Hager et al., 1994) or racial differences (Mayr et al., 2008)). Future research should conduct structural invariance tests to rule out these possibilities. Additionally, the medical criterion variables were measured via self-report of a *history* of a diagnosis. These results would be more informative if current medical conditions were known. Finally, many

inflammatory proteins have diurnal variations (Dominguez-Rodriguez et al., 2009) and Project ACE did not consistently restrict times for blood sampling (although efforts were made to ensure the majority of blood draws happened in the morning or early afternoon), unlike MIDUS-2 and MIDUS-R. However, patterns of model fit were comparable between the MIDUS projects and ACE, demonstrating that this methodological inconsistency did not render these samples incompatible for this analysis.

Conclusion

This study used standard aggregate measure building procedures to investigate the structure of eight inflammatory proteins. Results support the use of an empirically-identified hierarchical factor structure of inflammation or the modeling of individual proteins over an "a priori" aggregate in which all inflammatory proteins equally load onto a single dimension. Use of empirically-identified inflammatory factors as outcomes for medical history resulted in comparable conclusions to using individual proteins while losing fewer significant results to correction for multiple comparisons. When predictive validity to depression symptoms was compared, only the two first-order empirically-identified factors predicted depression, and the associations were in opposite directions, underscoring the potential benefits of this modeling approach and the increased risk of Type II error associated with falsely assuming unidimensionality. To facilitate other researchers in designing studies and analyzing data, reliability and temporal stability of the aggregates and proteins were described as appropriate. By building a strong physiometric foundation, immunopsychiatry can become a more replicable, cost-effective, and impactful field of research.

Supplemental Table 1. *Bivariate Correlations of Proteins in MIDUS-2*

Protein	1.	2.	3.	4.	5.	6.	7.	8.
1. CRP	——							
2. IL-6	.51	——						
3. TNF-α	.22	.33	——					
4. IL-8	.03	.15	.24	——				
5. IL-10	.10	.18	.37	.19	——			
6. Fibrinogen	.51	.42	.13	.07	.05	——		
7. E-selectin	.16	.17	.11	.11	.12	.13	——	
8. ICAM-1	.18	.17	.28	.07	.13	.12	.05	——

Note: CRP = C-reactive protein; IL- = Interleukin; TNF-α = tumor necrosis factor alpha; ICAM= intracellular adhesion molecule

Supplemental Table 2. *Bivariate Correlations of Proteins in MIDUS-R*

Protein	1.	2.	3.	4.	5.	6.	7.	8.
1. CRP	——							
2. IL-6	.52	——						
3. TNF-α	.21	.34	——					
4. IL-8	.06	.19	.28	——				
5. IL-10	.06	.17	.39	.04	——			
6. Fibrinogen	.53	.46	.18	.16	.03	——		
7. E-selectin	.22	.30	.29	.10	.19	.19	——	
8. ICAM-1	.19	.18	.23	.08	.18	.16	.39	——

Note: CRP = C-reactive protein; IL- = Interleukin; TNF-α = tumor necrosis factor alpha; ICAM= intracellular adhesion molecule

Supplemental Table 3. *Bivariate Correlations of Proteins in Project ACE*

Protein	1.	2.	3.	4.	5.
1. CRP	——				
2. IL-6	.54	——			
3. TNF-α	.14	.22	——		
4. IL-8	-.05	-.09	.10	——	
5. IL-10	.13	.28	.44	.01	——

Note: CRP = C-reactive protein; IL- = Interleukin; TNF-α = tumor necrosis factor alpha

References

Alloy, L. B., Black, S. K., Young, M. E., Goldstein, K. E., Shapero, B. G., Stange, J. P., Boccia, A. S., Matt, L. M., Boland, E. M., Moore, L. C., & Abramson, L. Y. (2012). Cognitive vulnerabilities and depression versus other psychopathology symptoms and diagnoses in early adolescence. *Journal of Clinical Child and Adolescent Psychology*, *41*(5), 539–560. https://doi.org/10.1080/15374416.2012.703123

Amrani, D. L. (1990). Regulation of fibrinogen biosynthesis: glucocorticoid and interleukin-6 control. *Blood Coagulation & Fibrinolysis*, *1*, 443–446. https://doi.org/10.1097/00001721-199010000-00013

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01>.

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. *Biological Psychology*, *129*(January), 207–230. https://doi.org/10.1016/j.biopsycho.2017.08.058

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

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*(12), 1669–1681. https://doi.org/10.1111/psyp.12545

Breen, E. C., Reynolds, S. M., Cox, C., Jacobson, L. P., Magpantay, L., Mulder, C. B., Dibben, O., Margolick, J. B., Bream, J. H., Sambrano, E., Martínez-Maza, O., Sinclair, E., Borrow, P., Landay, A. L., Rinaldo, C. R., & Norris, P. J. (2011). Multisite comparison of high-sensitivity multiplex cytokine assays. *Clinical and Vaccine Immunology*, *18*(8), 1229–1242. https://doi.org/10.1128/CVI.05032-11

Burke-Gaffney, A., & Hellewell, P. G. (1996). Tumour necrosis factor-ac-induced ICAM-1 expression in human vascular endothelial and lung epithelial cells : modulation by tyrosine kinase inhibitors. *British Journal of Pharmacology*, *119*, 1149–1158.

Cavaillon, J. M., & Adib-Conquy, M. (2002). The Pro-Inflammatory Cytokine Cascade. *Immune Response in the Critically Ill*, 37–66. https://doi.org/10.1007/978-3-642-57210-4_4

Cicchetti, D. V. (1993). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://search.proquest.com/docview/1756276764?accountid=9630%0Ahttp://pmt-eu.hosted.exlibrisgroup.com/openurl/44LSE/44LSE_services_page?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=unknown&sid=ProQ:ProQ%3Abusinesspremium&atitle=Stat

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Connor, B. P. (2020). *EFA.dimensions: Exploratory Factor Analysis Functions for Assessing Dimensionality. R package version 0.1.6.* https://cran.r-project.org/package=EFA.dimensions

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. *Biological Psychiatry, 71*(1), 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. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037//0021-9010.78.1.98

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Cummings, C., Caporino, N., & Kendall, P. C. (2014). Comorbidity of anxiety and depression in children and adolescents: 20 Years After. *Psychological Bulletin, 140*(3), 816–845. https://doi.org/10.1037/a0034733

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology, 122*(3), 928–937. https://doi.org/10.1037/a0034028

Dabitao, D., Margolick, J. B., Lopez, J., & Bream, J. H. (2011). Multiplex measurement of proinflammatory cytokines in human serum: comparison of the Meso Scale Discovery electrochemiluminescence assay and the Cytometric Bead Array. *Journal of Immunological Methods, 372*(1–2), 71–77. https://doi.org/10.1038/jid.2014.371

Davidshofer, K. R., & Murphy, C. O. (2005). *Psychological testing: principles and applications*.

Davidson, S. J. (2013). Inflammation and acute phase proteins in haemostasis. In *Acute Phase Proteins* (pp. 31–54).

Divietro, J. A., Smith, M. J., Smith, B. R. E., Petruzzelli, L., Larson, R. S., Lawrence, M. B., Larson, R. S., & Lawrence, M. B. (2001). Immobilized IL-8 Triggers Progressive Activation of Neutrophils Rolling In Vitro on P-Selectin and Intercellular Adhesion Molecule-1. *Journal of Immunology*, *167*, 4017–4025. https://doi.org/10.4049/jimmunol.167.7.4017

Dixit, N., & Simon, S. I. (2012). Chemokines, selectins and intracellular calcium flux: Temporal and spatial cues for leukocyte arrest. *Frontiers in Immunology*, *3*, 1–9. https://doi.org/10.3389/fimmu.2012.00188

Dominguez-Rodriguez, A., Abreu-Gonzalex, P., & Kaski, J. C. (2009). Inflammatory systemic biomarkers in setting acute coronary syndromes - effects of the diurnal variation. *Current Drug Targets*, *10*(10), 1001–1008.

Dooley, L. N., Kuhlman, K. R., Robles, T. F., Eisenberger, N. I., Craske, M. G., & Bower, J. E. (2018). The role of inflammation in core features of depression: Insights from paradigms using exogenously-induced inflammation. *Neuroscience and Biobehavioral Reviews*, *94*(March), 219–237. https://doi.org/10.1016/j.neubiorev.2018.09.006

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and

specificity of information criteria. *Briefings in Bioinformatics*, *21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*(3), 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. *Thrombosis Research*, *164*(October 2017), 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. *Psychological Science*, 681–700. https://doi.org/10.1101/681700

Etienne-Manneville, S., Chaverot, N., Strosberg, A. D., & Couraud, P. O. (1999). ICAM-1-coupled signaling pathways in astrocytes converge to cyclic AMP response element-binding protein phosphorylation and TNF-alpha secretion. *Journal of Immunology*, *163*(2), 668–674. http://www.ncbi.nlm.nih.gov/pubmed/10395656

Fielding, C. A., Mcloughlin, R. M., Mcleod, L., Colmont, C. S., Najdovska, M., Grail, D., Jones, S. A., Topley, N., Brendan, J., Stat, I., Fielding, C. A., Mcloughlin, R. M., Mcleod, L., Colmont, C. S., Najdovska, M., Grail, D., Ernst, M., Jones, S. A., Topley, N., & Jenkins, B. J. (2008). IL-6 regulates neutrophil trafficking during acute inflammation via STAT3. *Journal of Immunology*, *181*, 2189–2195. https://doi.org/10.4049/jimmunol.181.3.2189

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*, *50*, 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. In *Handbook of Research Methods in Health Psychology* (pp. 127–140). Routledge.

Goldberg, L. R. (2006). Doing it all Bass-Ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, *40*(4), 347–358. https://doi.org/10.1016/j.jrp.2006.01.001

Gough, P., & Myles, I. A. (2020). Tumor Necrosis Factor Receptors: Pleiotropic Signaling Complexes and Their Differential Effects. *Frontiers in Immunology*, *11*(November), 1–14. https://doi.org/10.3389/fimmu.2020.585880

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, Behavior, and Immunity*, *74*(August), 222–230. https://doi.org/10.1016/j.bbi.2018.09.011

Gruys, E., Toussaint, M. J. M., Niewold, T. A., & Koopmans, S. J. (2005). Acute phase reaction and acute phase proteins. *Journal of Zhejiang University: Science*, *6B*(11), 1045–1056. https://doi.org/10.1631/jzus.2005.B1045

Guadagnoli, E., & Velicer, W. F. (1988). Relation of Sample Size to the Stability of Component Patterns. *Psychological Bulletin*, *103*(2), 265–275. https://doi.org/10.1037/0033-2909.103.2.265

Hager, K., Machein, U., Krieger, S., Platt, D., Seefried, G., & Bauer, J. (1994). Interleukin-6 and Selected Plasma Proteins in Healthy Persons of Different Ages. *Neurobiology of Aging*, *15*(6), 771–772.

Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology*, *126*(6), 823–834. https://doi.org/10.1037/abn0000274

Hajcak, G., & Patrick, C. J. (2015). Situating psychophysiological science within the Research Domain Criteria (RDoC) framework. *International Journal of Psychophysiology*, *98*(2), 223–226. https://doi.org/10.1016/j.ijpsycho.2015.11.001

Hedges, J. C., Singer, C. A., & Gerthoffer, W. T. (2000). Mitogen-activated protein kinases regulate cytokine gene expression in human airway myocytes. *American Journal of Respiratory Cell and Molecular Biology*, *23*(1), 86–94. https://doi.org/10.1165/ajrcmb.23.1.4014

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*(11), 1–16. https://doi.org/10.1371/journal.pone.0206583

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L., Bentler, P. M., & Hu, L. (2009). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A*

Multidisciplinary Journal, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Imhof, B. A., & Dunon, D. (1995). Leukocyte migration and adhesion. *Advances in Immunology*, *58*, 345–416. https://doi.org/10.1016/s0065-2776(08)60623-9

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling. R package version 0.5-4* (p. 2021).

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. *Translational Psychiatry*, *10*(1). 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. *Molecular Psychiatry*, *17*(12), 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*(8), 1241–1255. https://doi.org/10.1111/psyp.12663

Kessler, B., Rinchai, D., Kewcharoenwong, C., Nithichanon, A., Biggart, R., Hawrylowicz, C. M., & Bancroft, G. J. (2017). Interleukin 10 inhibits pro- inflammatory cytokine responses and killing of Burkholderia pseudomallei. *Nature Publishing Group*, *February*, 1–11. https://doi.org/10.1038/srep42791

Klein, D., Dougherty, L. R., & Olino, T. M. (2005). Toward guidelines for evidence-based

assessment of depression in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *34*(3), 412, 432. https://doi.org/10.1207/s15374424jccp3403

Korthuis, J., Anderson, C., & Granger, D. N. (1994). Role of Neutrophil-Endothelial Cell Adhesion Inflammatory Disorders. *Journal of Critical Care*, *9*(1), 47–71.

Koukkunen, H., Penttilä, K., Kemppainen, A., Halinen, M., Penttilä, I., Rantanen, T., & Pyörälä, K. (2001). C-reactive protein, fibrinogen, interleukin-6 and tumour necrosis factor-α in the prognostic classification of unstable angina pectoris. *Annals of Medicine*, *33*(1), 37–47. https://doi.org/10.3109/07853890109002058

Kovacs, M. (1985). The Chidlren's Depression Inventory (CDI). *Psychopharmacology Bulletin*, *21*(4), 995–998.

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*(August 2018), 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. *Human Brain Mapping*, *28*(11), 1251–1266. https://doi.org/10.1002/hbm.20359

Loehlin, J. C., & Goldberg, L. R. (2014). Do personality traits conform to lists or hierarchies?

Personality and Individual Differences, *70*, 51–56. https://doi.org/10.1016/j.paid.2014.06.018

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(60), 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. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(3), 289–297. https://doi.org/10.1016/j.bpsc.2016.12.004

Luscinskas, F. W., Kiely, J. M., Ding, H., Obin, M. S., Hebert, C. A., Baker, J. B., & Gimbrone, M. A. (1992). In vitro inhibitory effect of IL-8 and other chemoattractants on neutrophil-endothelial adhesive interactions. *The Journal of Immunology*, *149*, 2163–2171.

Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The Perils of Partialling Cautionary Tales From Aggression and Psychopathy. *Assessment*, *13*(3), 328–341. https://doi.org/10.1177/1073191106290562

Mayr, F. B., Spiel, A. O., Leitner, J. M., Firbas, C., Kliegel, T., & & Jilma, B. (2008). Ethnic differences in plasma levels of interleukin-8 (IL-8) and granulocyte colony stimulating factor (G-CSF). *Translational Research*, *149*(1), 10–14.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Miller, A. H., Maletic, V., & Raison, C. L. (2009). Inflammation and its discontents: the role of

cytokines in the pathophysiology of major depression. *Biological Psychiatry*, *65*(9), 732–741. https://doi.org/10.1016/j.biopsych.2008.11.029

Miller, G. A. (2010). Mistreating psychology in the decades of the brain. *Perspectives on Psychological Science2*, *5*(6), 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. *Biological Psychiatry*, *72*(1), 34–40. https://doi.org/10.1016/j.biopsych.2012.02.034.

Moriarity, D. P., & Alloy, L. B. (2020). Beyond diagnoses and total symptom scores: Diversifying the level of analysis in psychoneuroimmunology research. *Brain, Behavior, and Immunity*, *89*, 1–2. https://doi.org/10.1016/j.bbi.2020.07.002

Moriarity, D. P., & Alloy, L. B. (2021). Back to basics: The importance of measurement properties in biological psychiatry. *Neuroscience and Biobehavioral Reviews*, *123*, 72–82. https://doi.org/10.1016/j.neubiorev.2021.01.008

Moriarity, D. P., Kautz, M. M., Mac Giollabhui, N., Klugman, J., Coe, C. L., Ellman, L. M., Abramson, L. Y., & Alloy, L. B. (2020). Bidirectional associations between inflammatory biomarkers and depressive symptoms in adolescents: Potential causal relationships. *Clinical Psychological Science*, *8*(4), 690–703. https://doi.org/10.1017/CBO9781107415324.004

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. *Clinical Psychological Science*, *7*(4), 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. *Journal of Youth and Adolescence*, *47*(12), 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. (2020). Reward sensitivity, cognitive response style, and inflammatory response to an acute stressor in adolescents. *Journal of Youth and Adolescence*, *49*, 2149–2159.

Moriarity, D. P., Ng, T., Titone, M. K., Chat, I. K., Nusslock, R., Miller, G. E., & Alloy, L. B. (2020). Reward sensitivity and ruminative response styles for positive and negative affect interact to predict inflammation and mood symptomatology. *Behavior Therapy*, *51*(5), 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, Behavior, and Immunity*, *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*(1), 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? *Translational Psychiatry*, *9*(293), 2–10.

Olszewski, M. B., Groot, A. J., Dastych, J., & Knol, E. F. (2007). TNF Trafficking to Human Mast Cell Granules: Mature Chain-Dependent Endocytosis. *The Journal of Immunology*, *178*(9), 5701–5709. https://doi.org/10.4049/jimmunol.178.9.5701

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, Behavior, and Immunity*, *26*(4), 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. *Psychological Assessment*, *31*(12), 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. *Journal of Abnormal Psychology*, *122*(3), 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. *International Journal of Psychophysiology*, *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*(3), 1746–1758.

https://doi.org/10.1016/j.neuroimage.2012.01.129

Raykov, T. (2001). Estimation of congeneric scale reliablity using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54*, 315–323.

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. *Developmental Psychobiology*, *57*(2), 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. *Developmental Psychobiology*, *56*(4), 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. *Journal of Educational Measurement*, *20*(4), 335–343.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Ryff, C. D., Seeman, T., & Weinstein, M. (2017). Midlife in the United States (MIDUS 2): Biomarker Project, 2004-2009. *Ann Arbor, MI: Inter-University Consortium for Political and Social Research [Distributor]*, 10.

Sakkinen, P. A., Wahl, P., Cushman, M., Lewis, M. R., & Tracy, R. P. (2000). Clustering of procoagulation, inflammation, and fibrinolysis variables with metabolic factors in insulin

resistance syndrome. *American Journal of Epidemiology*, *152*(10), 891–907.

Samejima, F. (1973). Homogenous case of the continuous response model. *Psychometrika*, *38*(2), 203–2019.

Scheller, J., Chalaris, A., Schmidt-Arras, D., & Rose-John, S. (2011). The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1813*(5), 878–888. https://doi.org/10.1016/j.bbamcr.2011.01.034

Segerstrom, S. C. (2020). Physiometrics in Salivary Bioscience. *International Journal of Behavioral Medicine*, *27*, 262–266.

Segerstrom, S. C., & Boggero, I. A. (2020). Expected Estimation Errors in Studies of the Cortisol Awakening Response: A Simulation. *Psychosomatic Medicine*, *82*(8), 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*(1), 299–309. https://doi.org/10.1016/j.psyneuen.2014.07.022

Segerstrom, S. C., & Miller, G. E. (2004). Psychological Stress and the Human Immune System : A Meta-Analytic Study of 30 Years of Inquiry. *Psychological Bulletin*, *130*(4), 601–630. https://doi.org/10.1037/0033-2909.130.4.601

Segerstrom, S. C., & Smith, G. T. (2012). Methods, variance, and error in psychoneuroimmunology research: The good, the bad, and the ugly. In S. C. Segerstrom (Ed.), *Oxford Handbook of Psychoneuroimmunology* (pp. 421–432). Oxford U Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*.

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, Behavior, and Immunity*, *81*(January 2020), 650–654. https://doi.org/10.1016/j.bbi.2019.06.007

Slavich, G. M. (2020). Social Safety Theory: A Biologically Based Evolutionary Perspective on Life Stress, Health, and Behavior. *Annual Review of Clinical Psychology*, *16*, 265–295. https://doi.org/10.1146/annurev-clinpsy-032816-045159

Slavich, G. M., & Irwin, M. R. (2014). From stress to inflammation and major depressive disorder: A social signal transduction theory of depression. *Psychological Bulletin*, *140*(3), 774–815. https://doi.org/10.1037/a0035302

Smart, S. J., & Casal, T. B. (1994). Pulmonary epithelial cells facilitate TNF-alpha-induced neutrophil chemotaxis. A role for cytokine networking. *Journal of Immunology*, *152*, 4087–4094.

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, Behavior, and Immunity*, *23*(7), 936–944. https://doi.org/10.1016/j.bbi.2009.04.011

Stumper, A., Olino, T. M., Abramson, L. Y., & Alloy, L. B. (2019). A factor analysis and test of longitudinal measurement invariance of the Children's Depression Inventory (CDI) across adolescence. *Journal of Psychopathology and Behavioral Assessment*, *41*, 692–698.

Sun, L., Guo, R., Newstead, M. W., Standiford, T. J., Macariola, D. R., & Shanley, T. P. (2009). *Effect of IL-10 on Neutrophil Recruitment and Survival after Pseudomonas aeruginosa*

Challenge. *11*. https://doi.org/10.1165/rcmb.2008-0202OC

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (Sixth). Pearson.

Team, R. C. (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.r-project.org

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. *Educational and Psychological Measurement*, *72*(1), 37–43. https://doi.org/10.1177/0013164411409929

Tziakas, D. N., Chalikias, G. K., Kaski, J. C., Kekes, A., Hatzinikolaou, E. I., Stakos, D. A., Tentes, I. K., Kortsaris, A. X., & Hatseras, D. I. (2007). Inflammatory and anti-inflammatory variable clusters and risk prediction in acute coronary syndrome patients: A factor analysis approach. *Atherosclerosis*, *193*(1), 196–203. https://doi.org/10.1016/j.atherosclerosis.2006.06.016

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. *Experimental Gerontology*, *42*(7), 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. *Clinical Psychological Science*, *6*(4), 561–580. https://doi.org/10.1177/2167702618757690

Venkatasubramanian, G., & Keshavan, M. S. (2016). Biomarkers in psychiatry – A critique. *Annals of Neurosciences*, *23*(1), 3–5. https://doi.org/10.1159/000443549

Weinstein, M., Ryff, C., & Seeman, T. (2017). Midlife in the United States (MIDUS Refresher): Biomarker Project, 2012–2016. *Ann Arbor, MI: Interuniversity Consortium for Political and Social Research [Distributor]*, 12–21.

Williams, L. M. (2016). Precision psychiatry: A neural circuit taxonomy for depression and anxiety. *The Lancet Psychiatry*, *3*(5), 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

Zakharova, M., & Ziegler, H. K. (2005). Paradoxical Anti-Inflammatory Actions of TNF-α: Inhibition of IL-12 and IL-23 via TNF Receptor 1 in Macrophages and Dendritic Cells. *The Journal of Immunology*, *175*(8), 5024–5033. https://doi.org/10.4049/jimmunol.175.8.5024

Zhang, G., Jiang, G., Hattori, M., & Trichtinger, L. (2020). *EFAutilities: Utility Functions for Exploratory Factor Analysis. R package version 2.1.1.* https://cran.r-project.org/package=EFAutilities

CHAPTER 2

ASSOCIATED LITERATURE REVIEW

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 & 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.

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.

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 & 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).*

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 & 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 & 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.*

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).*

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.*

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.*

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 & 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 & 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.

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.

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 and colleagues (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 and colleagues (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 & 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.

Dimensionality

Recall the example of inflammation as a complex construct often indexed by several indicators (Segerstrom & 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 and colleagues 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.

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, Patrick, and Bernat (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 and colleagues (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 and colleagues (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.

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 $\acute{\eta}_p^2 = .001\text{-}.033$, $p > .05$), 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, pair-wise 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 and colleagues (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 (Moriarity, Ng, Curley, et al., 2020; 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 and colleagues (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 and colleagues (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 rs = .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 (rs = .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 rs = .10 - .54; saliva rs = .09 - .36) and for two-year reliability, for which saliva reliability was higher for four of the analytes (plasma rs = .16 - .57; saliva rs = .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). 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 and colleagues (2019) reported ICCs (in their supplemental material) for seven different salivary inflammatory proteins (CRP, IL-6, IL-8, IL-18, IL-1β, TNFα, MCP). 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. Elliot and colleagues (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 Elliot et al. (2020) (Wu et al., 2014). In response to Elliot and colleagues (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 Elliot 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 & 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 and colleagues (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.

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, Rand, Muldoon, and Kamarck (2009; six-year follow-up) found significant associations between IL-6 and future depression symptoms, but van der Biggelaar and colleagues found that CRP predicted future depression. Further, Copeland and colleagues (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 and colleagues (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., Figure 4). Conversely, IL-8 had a stronger association with change in depression symptoms for males as time to follow-up increased,

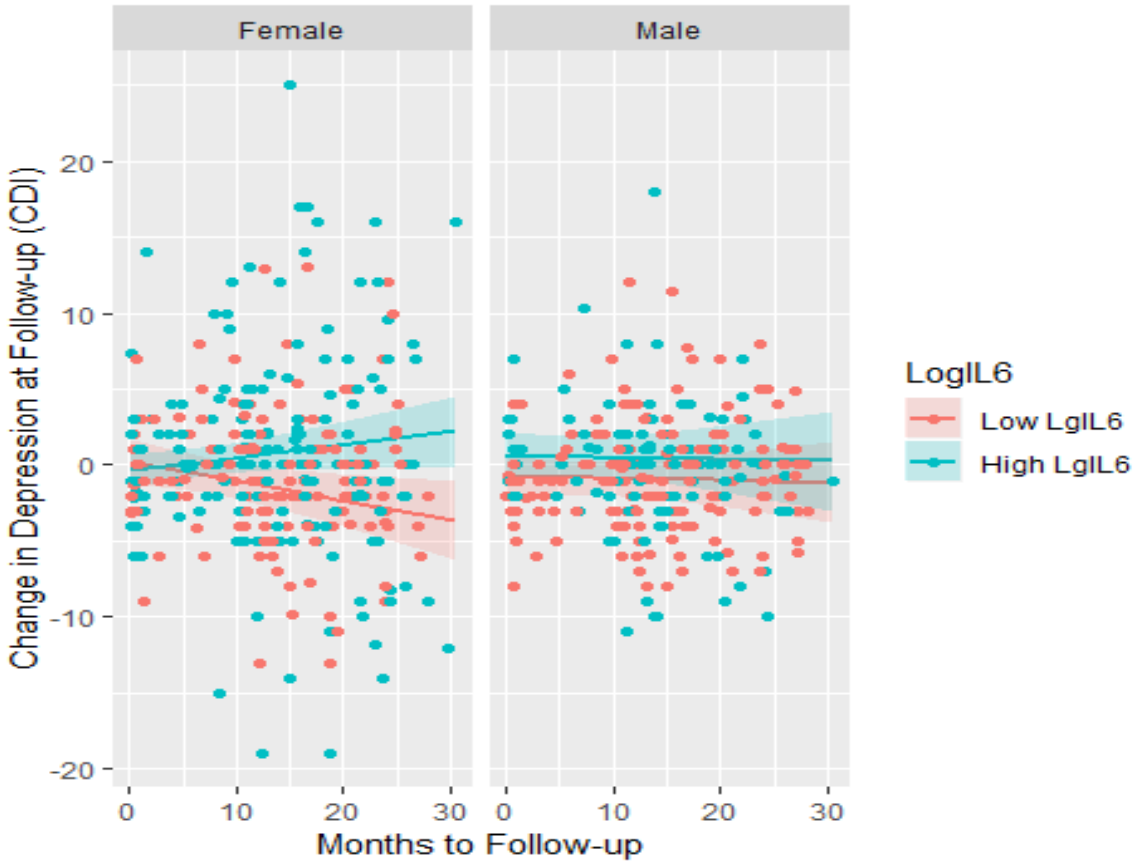

Figure 4. Temporal Specificity of Log IL-6 Predicting Change in Depression Symptoms by Sex. This figure was first presented in Moriarity et al. (2019). Note: IL = interleukin, CDI = Children's Depression Inventory. Shaded regions indicate 95% confidence intervals.

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 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 and colleagues (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.

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 and colleagues' (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 and colleagues (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.

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; G. A. Miller, 2010; Venkatasubramanian & 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 and colleagues (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 and colleagues (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 & 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 & Watson, 2019; Cronbach & 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.

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; G. A. Miller, 2010; Venkatasubramanian & 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.

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

References

Alloy, L. B., Black, S. K., Young, M. E., Goldstein, K. E., Shapero, B. G., Stange, J. P., Boccia, A. S., Matt, L. M., Boland, E. M., Moore, L. C., & Abramson, L. Y. (2012). Cognitive vulnerabilities and depression versus other psychopathology symptoms and diagnoses in early adolescence. *Journal of Clinical Child and Adolescent Psychology*, *41*(5), 539–560. https://doi.org/10.1080/15374416.2012.703123

Amrani, D. L. (1990). Regulation of fibrinogen biosynthesis: glucocorticoid and interleukin-6 control. *Blood Coagulation & Fibrinolysis*, *1*, 443–446. https://doi.org/10.1097/00001721-199010000-00013

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01>.

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. *Biological Psychology*, *129*(January), 207–230. https://doi.org/10.1016/j.biopsycho.2017.08.058

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

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*(12), 1669–1681. https://doi.org/10.1111/psyp.12545

Breen, E. C., Reynolds, S. M., Cox, C., Jacobson, L. P., Magpantay, L., Mulder, C. B., Dibben, O., Margolick, J. B., Bream, J. H., Sambrano, E., Martínez-Maza, O., Sinclair, E., Borrow,

P., Landay, A. L., Rinaldo, C. R., & Norris, P. J. (2011). Multisite comparison of high-sensitivity multiplex cytokine assays. *Clinical and Vaccine Immunology*, *18*(8), 1229–1242. https://doi.org/10.1128/CVI.05032-11

Burke-Gaffney, A., & Hellewell, P. G. (1996). Tumour necrosis factor-ac-induced ICAM-1 expression in human vascular endothelial and lung epithelial cells : modulation by tyrosine kinase inhibitors. *British Journal of Pharmacology*, *119*, 1149–1158.

Cavaillon, J. M., & Adib-Conquy, M. (2002). The Pro-Inflammatory Cytokine Cascade. *Immune Response in the Critically Ill*, 37–66. https://doi.org/10.1007/978-3-642-57210-4_4

Cicchetti, D. V. (1993). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://search.proquest.com/docview/1756276764?accountid=9630%0Ahttp://pmt-eu.hosted.exlibrisgroup.com/openurl/44LSE/44LSE_services_page?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=unknown&sid=ProQ:ProQ%3Abusinesspremium&atitle=Stat

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Connor, B. P. (2020). *EFA.dimensions: Exploratory Factor Analysis Functions for Assessing Dimensionality. R package version 0.1.6.* https://cran.r-project.org/package=EFA.dimensions

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. *Biological Psychiatry*, *71*(1), 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. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037//0021-9010.78.1.98

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cummings, C., Caporino, N., & Kendall, P. C. (2014). Comorbidity of anxiety and depression in children and adolescents: 20 Years After. *Psychological Bulletin*, *140*(3), 816–845. https://doi.org/10.1037/a0034733

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology*, *122*(3), 928–937. https://doi.org/10.1037/a0034028

Dabitao, D., Margolick, J. B., Lopez, J., & Bream, J. H. (2011). Multiplex measurement of proinflammatory cytokines in human serum: comparison of the Meso Scale Discovery electrochemiluminescence assay and the Cytometric Bead Array. *Journal of Immunological Methods*, *372*(1–2), 71–77. https://doi.org/10.1038/jid.2014.371

Davidshofer, K. R., & Murphy, C. O. (2005). *Psychological testing: principles and applications*.

Davidson, S. J. (2013). Inflammation and acute phase proteins in haemostasis. In *Acute Phase Proteins* (pp. 31–54).

Divietro, J. A., Smith, M. J., Smith, B. R. E., Petruzzelli, L., Larson, R. S., Lawrence, M. B., Larson, R. S., & Lawrence, M. B. (2001). Immobilized IL-8 Triggers Progressive Activation of Neutrophils Rolling In Vitro on P-Selectin and Intercellular Adhesion

Molecule-1. *Journal of Immunology*, *167*, 4017–4025. https://doi.org/10.4049/jimmunol.167.7.4017

Dixit, N., & Simon, S. I. (2012). Chemokines, selectins and intracellular calcium flux: Temporal and spatial cues for leukocyte arrest. *Frontiers in Immunology*, *3*, 1–9. https://doi.org/10.3389/fimmu.2012.00188

Dominguez-Rodriguez, A., Abreu-Gonzalex, P., & Kaski, J. C. (2009). Inflammatory systemic biomarkers in setting acute coronary syndromes - effects of the diurnal variation. *Current Drug Targets*, *10*(10), 1001–1008.

Dooley, L. N., Kuhlman, K. R., Robles, T. F., Eisenberger, N. I., Craske, M. G., & Bower, J. E. (2018). The role of inflammation in core features of depression: Insights from paradigms using exogenously-induced inflammation. *Neuroscience and Biobehavioral Reviews*, *94*(March), 219–237. https://doi.org/10.1016/j.neubiorev.2018.09.006

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, *21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*(3), 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.

Thrombosis Research, *164*(October 2017), 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. *Psychological Science*, 681–700. https://doi.org/10.1101/681700

Etienne-Manneville, S., Chaverot, N., Strosberg, A. D., & Couraud, P. O. (1999). ICAM-1-coupled signaling pathways in astrocytes converge to cyclic AMP response element-binding protein phosphorylation and TNF-alpha secretion. *Journal of Immunology*, *163*(2), 668–674. http://www.ncbi.nlm.nih.gov/pubmed/10395656

Fielding, C. A., Mcloughlin, R. M., Mcleod, L., Colmont, C. S., Najdovska, M., Grail, D., Jones, S. A., Topley, N., Brendan, J., Stat, I., Fielding, C. A., Mcloughlin, R. M., Mcleod, L., Colmont, C. S., Najdovska, M., Grail, D., Ernst, M., Jones, S. A., Topley, N., & Jenkins, B. J. (2008). IL-6 regulates neutrophil trafficking during acute inflammation via STAT3. *Journal of Immunology*, *181*, 2189–2195. https://doi.org/10.4049/jimmunol.181.3.2189

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*, *50*, 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. In *Handbook of Research Methods in Health Psychology* (pp. 127–140). Routledge.

Goldberg, L. R. (2006). Doing it all Bass-Ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, *40*(4), 347–358. https://doi.org/10.1016/j.jrp.2006.01.001

Gough, P., & Myles, I. A. (2020). Tumor Necrosis Factor Receptors: Pleiotropic Signaling Complexes and Their Differential Effects. *Frontiers in Immunology*, *11*(November), 1–14. https://doi.org/10.3389/fimmu.2020.585880

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, Behavior, and Immunity*, *74*(August), 222–230. https://doi.org/10.1016/j.bbi.2018.09.011

Gruys, E., Toussaint, M. J. M., Niewold, T. A., & Koopmans, S. J. (2005). Acute phase reaction and acute phase proteins. *Journal of Zhejiang University: Science*, *6B*(11), 1045–1056. https://doi.org/10.1631/jzus.2005.B1045

Guadagnoli, E., & Velicer, W. F. (1988). Relation of Sample Size to the Stability of Component Patterns. *Psychological Bulletin*, *103*(2), 265–275. https://doi.org/10.1037/0033-2909.103.2.265

Hager, K., Machein, U., Krieger, S., Platt, D., Seefried, G., & Bauer, J. (1994). Interleukin-6 and Selected Plasma Proteins in Healthy Persons of Different Ages. *Neurobiology of Aging*, *15*(6), 771–772.

Hajcak, G., Meyer, A., & Kotov, R. (2017). Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. *Journal of Abnormal Psychology*, *126*(6), 823–834. https://doi.org/10.1037/abn0000274

Hajcak, G., & Patrick, C. J. (2015). Situating psychophysiological science within the Research Domain Criteria (RDoC) framework. *International Journal of Psychophysiology*, *98*(2), 223–226. https://doi.org/10.1016/j.ijpsycho.2015.11.001

Hedges, J. C., Singer, C. A., & Gerthoffer, W. T. (2000). Mitogen-activated protein kinases

regulate cytokine gene expression in human airway myocytes. *American Journal of Respiratory Cell and Molecular Biology*, *23*(1), 86–94. https://doi.org/10.1165/ajrcmb.23.1.4014

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*(11), 1–16. https://doi.org/10.1371/journal.pone.0206583

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L., Bentler, P. M., & Hu, L. (2009). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Imhof, B. A., & Dunon, D. (1995). Leukocyte migration and adhesion. *Advances in Immunology*, *58*, 345–416. https://doi.org/10.1016/s0065-2776(08)60623-9

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling. R package version 0.5-4* (p. 2021).

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. *Translational Psychiatry*, *10*(1). 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. *Molecular Psychiatry*, *17*(12),

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*(8), 1241–1255. https://doi.org/10.1111/psyp.12663

Kessler, B., Rinchai, D., Kewcharoenwong, C., Nithichanon, A., Biggart, R., Hawrylowicz, C. M., & Bancroft, G. J. (2017). Interleukin 10 inhibits pro- inflammatory cytokine responses and killing of Burkholderia pseudomallei. *Nature Publishing Group, February*, 1–11. https://doi.org/10.1038/srep42791

Klein, D., Dougherty, L. R., & Olino, T. M. (2005). Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *34*(3), 412, 432. https://doi.org/10.1207/s15374424jccp3403

Korthuis, J., Anderson, C., & Granger, D. N. (1994). Role of Neutrophil-Endothelial Cell Adhesion Inflammatory Disorders. *Journal of Critical Care*, *9*(1), 47–71.

Koukkunen, H., Penttilä, K., Kemppainen, A., Halinen, M., Penttilä, I., Rantanen, T., & Pyörälä, K. (2001). C-reactive protein, fibrinogen, interleukin-6 and tumour necrosis factor-α in the prognostic classification of unstable angina pectoris. *Annals of Medicine*, *33*(1), 37–47. https://doi.org/10.3109/07853890109002058

Kovacs, M. (1985). The Chidlren's Depression Inventory (CDI). *Psychopharmacology Bulletin*, *21*(4), 995–998.

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*(August 2018), 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. *Human Brain Mapping*, *28*(11), 1251–1266. https://doi.org/10.1002/hbm.20359

Loehlin, J. C., & Goldberg, L. R. (2014). Do personality traits conform to lists or hierarchies? *Personality and Individual Differences*, *70*, 51–56. https://doi.org/10.1016/j.paid.2014.06.018

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(60), 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. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(3), 289–297. https://doi.org/10.1016/j.bpsc.2016.12.004

Luscinskas, F. W., Kiely, J. M., Ding, H., Obin, M. S., Hebert, C. A., Baker, J. B., & Gimbrone, M. A. (1992). In vitro inhibitory effect of IL-8 and other chemoattractants on neutrophil-endothelial adhesive interactions. *The Journal of Immunology*, *149*, 2163–2171.

Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The Perils of Partialling Cautionary Tales From Aggression and Psychopathy. *Assessment*, *13*(3), 328–341. https://doi.org/10.1177/1073191106290562

Mayr, F. B., Spiel, A. O., Leitner, J. M., Firbas, C., Kliegel, T., & & Jilma, B. (2008). Ethnic

differences in plasma levels of interleukin-8 (IL-8) and granulocyte colony stimulating factor (G-CSF). *Translational Research*, *149*(1), 10–14.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Miller, A. H., Maletic, V., & Raison, C. L. (2009). Inflammation and its discontents: the role of cytokines in the pathophysiology of major depression. *Biological Psychiatry*, *65*(9), 732–741. https://doi.org/10.1016/j.biopsych.2008.11.029

Miller, G. A. (2010). Mistreating psychology in the decades of the brain. *Perspectives on Psychological Science2*, *5*(6), 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. *Biological Psychiatry*, *72*(1), 34–40. https://doi.org/10.1016/j.biopsych.2012.02.034.

Moriarity, D. P., & Alloy, L. B. (2020). Beyond diagnoses and total symptom scores: Diversifying the level of analysis in psychoneuroimmunology research. *Brain, Behavior, and Immunity*, *89*, 1–2. https://doi.org/10.1016/j.bbi.2020.07.002

Moriarity, D. P., & Alloy, L. B. (2021). Back to basics: The importance of measurement properties in biological psychiatry. *Neuroscience and Biobehavioral Reviews*, *123*, 72–82. https://doi.org/10.1016/j.neubiorev.2021.01.008

Moriarity, D. P., Kautz, M. M., Mac Giollabhui, N., Klugman, J., Coe, C. L., Ellman, L. M., Abramson, L. Y., & Alloy, L. B. (2020). Bidirectional associations between inflammatory biomarkers and depressive symptoms in adolescents: Potential causal relationships. *Clinical Psychological Science*, *8*(4), 690–703. https://doi.org/10.1017/CBO9781107415324.004

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. *Clinical Psychological Science, 7*(4), 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. *Journal of Youth and Adolescence*, *47*(12), 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. (2020). Reward sensitivity, cognitive response style, and inflammatory response to an acute stressor in adolescents. *Journal of Youth and Adolescence*, *49*, 2149–2159.

Moriarity, D. P., Ng, T., Titone, M. K., Chat, I. K., Nusslock, R., Miller, G. E., & Alloy, L. B. (2020). Reward sensitivity and ruminative response styles for positive and negative affect interact to predict inflammation and mood symptomatology. *Behavior Therapy*, *51*(5), 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, Behavior, and Immunity*, *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*(1), 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? *Translational*

Psychiatry, *9*(293), 2–10.

Olszewski, M. B., Groot, A. J., Dastych, J., & Knol, E. F. (2007). TNF Trafficking to Human Mast Cell Granules: Mature Chain-Dependent Endocytosis. *The Journal of Immunology*, *178*(9), 5701–5709. https://doi.org/10.4049/jimmunol.178.9.5701

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, Behavior, and Immunity*, *26*(4), 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. *Psychological Assessment*, *31*(12), 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. *Journal of Abnormal Psychology*, *122*(3), 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. *International Journal of Psychophysiology*, *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*(3), 1746–1758.

https://doi.org/10.1016/j.neuroimage.2012.01.129

Raykov, T. (2001). Estimation of congeneric scale reliablity using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54*, 315–323.

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. *Developmental Psychobiology*, *57*(2), 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. *Developmental Psychobiology*, *56*(4), 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. *Journal of Educational Measurement*, *20*(4), 335–343.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Ryff, C. D., Seeman, T., & Weinstein, M. (2017). Midlife in the United States (MIDUS 2): Biomarker Project, 2004-2009. *Ann Arbor, MI: Inter-University Consortium for Political and Social Research [Distributor]*, 10.

Sakkinen, P. A., Wahl, P., Cushman, M., Lewis, M. R., & Tracy, R. P. (2000). Clustering of procoagulation, inflammation, and fibrinolysis variables with metabolic factors in insulin resistance syndrome. *American Journal of Epidemiology*, *152*(10), 891–907.

Samejima, F. (1973). Homogenous case of the continuous response model. *Psychometrika*,

38(2), 203–2019.

Scheller, J., Chalaris, A., Schmidt-Arras, D., & Rose-John, S. (2011). The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1813*(5), 878–888. https://doi.org/10.1016/j.bbamcr.2011.01.034

Segerstrom, S. C. (2020). Physiometrics in Salivary Bioscience. *International Journal of Behavioral Medicine*, *27*, 262–266.

Segerstrom, S. C., & Boggero, I. A. (2020). Expected Estimation Errors in Studies of the Cortisol Awakening Response: A Simulation. *Psychosomatic Medicine*, *82*(8), 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*(1), 299–309. https://doi.org/10.1016/j.psyneuen.2014.07.022

Segerstrom, S. C., & Miller, G. E. (2004). Psychological Stress and the Human Immune System : A Meta-Analytic Study of 30 Years of Inquiry. *Psychological Bulletin*, *130*(4), 601–630. https://doi.org/10.1037/0033-2909.130.4.601

Segerstrom, S. C., & Smith, G. T. (2012). Methods, variance, and error in psychoneuroimmunology research: The good, the bad, and the ugly. In S. C. Segerstrom (Ed.), *Oxford Handbook of Psychoneuroimmunology* (pp. 421–432). Oxford U Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*.

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, Behavior, and Immunity*, *81*(January 2020), 650–654. https://doi.org/10.1016/j.bbi.2019.06.007

Slavich, G. M. (2020). Social Safety Theory: A Biologically Based Evolutionary Perspective on Life Stress, Health, and Behavior. *Annual Review of Clinical Psychology*, *16*, 265–295. https://doi.org/10.1146/annurev-clinpsy-032816-045159

Slavich, G. M., & Irwin, M. R. (2014). From stress to inflammation and major depressive disorder: A social signal transduction theory of depression. *Psychological Bulletin*, *140*(3), 774–815. https://doi.org/10.1037/a0035302

Smart, S. J., & Casal, T. B. (1994). Pulmonary epithelial cells facilitate TNF-alpha-induced neutrophil chemotaxis. A role for cytokine networking. *Journal of Immunology*, *152*, 4087–4094.

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, Behavior, and Immunity*, *23*(7), 936–944. https://doi.org/10.1016/j.bbi.2009.04.011

Stumper, A., Olino, T. M., Abramson, L. Y., & Alloy, L. B. (2019). A factor analysis and test of longitudinal measurement invariance of the Children's Depression Inventory (CDI) across adolescence. *Journal of Psychopathology and Behavioral Assessment*, *41*, 692–698.

Sun, L., Guo, R., Newstead, M. W., Standiford, T. J., Macariola, D. R., & Shanley, T. P. (2009). *Effect of IL-10 on Neutrophil Recruitment and Survival after Pseudomonas aeruginosa Challenge*. *11*. https://doi.org/10.1165/rcmb.2008-0202OC

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (Sixth). Pearson.

Team, R. C. (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.r-project.org

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.

Educational and Psychological Measurement, *72*(1), 37–43. https://doi.org/10.1177/0013164411409929

Tziakas, D. N., Chalikias, G. K., Kaski, J. C., Kekes, A., Hatzinikolaou, E. I., Stakos, D. A., Tentes, I. K., Kortsaris, A. X., & Hatseras, D. I. (2007). Inflammatory and anti-inflammatory variable clusters and risk prediction in acute coronary syndrome patients: A factor analysis approach. *Atherosclerosis*, *193*(1), 196–203. https://doi.org/10.1016/j.atherosclerosis.2006.06.016

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. *Experimental Gerontology*, *42*(7), 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. *Clinical Psychological Science*, *6*(4), 561–580. https://doi.org/10.1177/2167702618757690

Venkatasubramanian, G., & Keshavan, M. S. (2016). Biomarkers in psychiatry – A critique. *Annals of Neurosciences*, *23*(1), 3–5. https://doi.org/10.1159/000443549

Weinstein, M., Ryff, C., & Seeman, T. (2017). Midlife in the United States (MIDUS Refresher): Biomarker Project, 2012–2016. *Ann Arbor, MI: Interuniversity Consortium for Political and Social Research [Distributor]*, 12–21.

Williams, L. M. (2016). Precision psychiatry: A neural circuit taxonomy for depression and

anxiety. *The Lancet Psychiatry*, *3*(5), 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

Zakharova, M., & Ziegler, H. K. (2005). Paradoxical Anti-Inflammatory Actions of TNF-α: Inhibition of IL-12 and IL-23 via TNF Receptor 1 in Macrophages and Dendritic Cells. *The Journal of Immunology*, *175*(8), 5024–5033. https://doi.org/10.4049/jimmunol.175.8.5024

Zhang, G., Jiang, G., Hattori, M., & Trichtinger, L. (2020). *EFAutilities: Utility Functions for Exploratory Factor Analysis. R package version 2.1.1.* https://cran.r-project.org/package=EFAutilities