Running Head: PSYCHIATRIC PHENOTYPES & MEASUREMENT INVARIANCE

This paper is now published in *Molecular Psychiatry*
(https://www.nature.com/articles/s41380-021-01414-5). Please cite as

Moriarity, D.P., Joyner, K.J., Slavich, G.M. *et al.* Unconsidered issues of measurement noninvariance in biological psychiatry: A focus on biological phenotypes of psychopathology. *Mol Psychiatry* (2022). https://doi.org/10.1038/s41380-021-01414-5

Unconsidered Issues of Measurement Noninvariance in Biological Psychiatry: A Focus on Biological Phenotypes of Psychopathology

Daniel P. Moriarity, M.A.[1]; Keanan J. Joyner, M.A.[2];
George M. Slavich, Ph.D.[3]; & Lauren B. Alloy, Ph.D.[1]

[1]Department of Psychology, Temple University

[2]Department of Psychology, Florida State University

[3]Cousins Center for Psychoneuroimmunology and Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles

Correspondence concerning this article should be addressed to Daniel P. Moriarity, Department of Psychology, Temple University, Weiss Hall, 1701 N. 13th St., Philadelphia, PA 19122, United States of America. E-mail: Daniel.moriarity@temple.edu

Abstract

There is increasing appreciation that certain biological mechanisms may not be equally related to all psychiatric symptoms in a given diagnostic category. Research on the biological phenotyping of psychopathology has begun examining the etiological and treatment implications of identified biotypes; however, little attention has been paid to a critical methodological implication of these results: measurement noninvariance. Measurement invariance is the ability of an instrument to measure the same construct across different people or across different time points for the same individual. If what a measure quantifies differs across different people (e.g., those with or without a particular biotype) or time points, it is invalid to directly compare means on said measure. Using a running example of inflammatory phenotypes of depression, we first describe the biological phenotyping of psychopathology. Second, we discuss three types of measurement invariance. Third, we demonstrate how differential biology-symptom associations invariably creates measurement noninvariance using a theoretical example and simulated data (for which code is provided), and how this issue can lead to false conclusions about the broader diagnostic construct. Finally, we provide several suggestions for addressing these important issues to help advance the field of biological psychiatry.

Introduction

Many research questions in biological psychiatry use biological variables such as inflammatory physiology, grey matter volume, or gene expression as predictors of differences on an aggregate measure of psychopathology. An underlying assumption of these tests, as commonly performed, is that the psychopathology measure used assesses the same construct each time it is administered, either across different people or across different time points for the same individual. This assumption might be untenable in light of growing evidence that some biological risk factors have differential associations with symptoms within a diagnostic construct (e.g., inflammatory proteins being most robustly associated with neurovegetative symptoms of depression [1]). In this article, we first briefly describe the concept of biological phenotypes. Second, we discuss the concept of measurement invariance. Third, we illustrate both how the presence of biological phenotypes of psychopathology induces measurement noninvariance and how this results in inappropriate conclusions about the broader diagnostic construct using a theoretical example and statistical simulation. Finally, we provide some recommendations for moving forward.

Biological Phenotypes of Psychopathology

There is accumulating evidence that different psychiatric symptoms within some diagnostic categories (e.g., depression) may have different risk factors [2]. Such findings have prompted interest in the symptom-level biological phenotyping of psychopathology. The thorough characterization of which specific symptoms within a disorder are associated with a given mechanism may in turn help advance biological psychiatry and precision medicine more generally. For example, understanding that inflammation is associated primarily with neurovegetative depression symptoms [1] can help clinicians identify patients who may possess

an underlying atypical inflammatory phenotype, and this information can in turn guide decisions about who might benefit most from adjunctive anti-inflammatory treatments [3].

Studying biological phenotypes of psychopathology will also improve the replicability of psychiatric research. For example, consider that the effect sizes between C-reactive protein (CRP) and depression symptoms in published research are highly variable across studies [4]. Given differential relations between CRP and depression symptoms [5–7], the mixed findings between sum scores of depression symptomatology or diagnostic groups and CRP is likely influenced by the sampling variability of symptom profiles across studies. Guided by phenotyping research, refining the psychiatric outcomes to more atomic levels (i.e., specific symptoms or subscales consistently associated with CRP) might increase replicability and shorten the research to practice timeline for syndromes characterized by high degrees of heterogeneity [8].

The implications of differential associations between a risk factor and the symptoms of a disorder extend beyond etiology, nosology, and treatment. Below, we examine an important methodological concern that has been largely ignored in extant discourse on phenotyping: measurement noninvariance. The running example of inflammation and depression will be continued throughout to contextualize the issue of measurement noninvariance, its consequences, and appropriate courses of action to ameliorate this concern. However, the issue of measurement invariance is universally applicable to all risk factors that are unequally associated with different symptoms on a measure.

Measurement Invariance: A Brief Overview

In the context of psychological questionnaires, measurement invariance is the ability of a questionnaire to measure the same construct regardless of who takes it (e.g., people from two different groups) or when it is completed (e.g., same person at multiple points in a longitudinal

study). To keep the language consistent, we will focus on measurement invariance across groups. Without measurement invariance, it is inappropriate to compare means, the most common level of analysis in psychiatric research. The three most commonly discussed types of measurement invariance are configural, metric (sometimes referred to as "weak" invariance), and scalar invariance (sometimes referred to as "strong" invariance). We will briefly discuss configural and metric invariance but focus mostly on scalar invariance, for the reasons described below. See Fig. 1 for visualization of the three kinds of measurement invariance. For a more thorough review of measurement invariance and how to test it, see [9].

Configural invariance, the least strict form of measurement invariance, refers to equivalence of model form. That is, which variables (e.g., items) load onto which latent variables (e.g., depression) does not change as a function of a third variable (e.g., elevated inflammatory phenotype). Configural noninvariance can be handled in two ways: (a) omit the noninvariant items and retest the model, or (b) conclude the construct itself is noninvariant and forego group difference testing entirely [9]. If configural invariance is supported, the next form of invariance to check is metric. Metric invariance, in turn, refers to the equivalence of item loadings on factors. If metric invariance is unsupported, there are three options: (a) investigate the factor loadings driving the noninvariance by sequentially removing or adding factor loadings constraints and retesting the models until a partially invariant model is found (for a description of partial invariance, see [9]), (b) remove items with noninvariant factor loadings and retest the configural and metric invariance models, or (c) conclude the construct is noninvariant and forego group difference testing entirely.

If both configural and metric invariance are supported, the next step is to test for scalar invariance for the items with metric invariance. Scalar invariance refers to equality of item

intercepts/thresholds (i.e., what level of endorsement of an item to expect if the latent variable associated with the item is 0). If item intercepts differ between groups, then observed mean differences in the latent construct (e.g., depression) do not accurately capture true mean differences in the latent variable (see below for an illustration). Thus, if scalar invariance is not met between two groups, any statistical test comparing mean differences on the total number of depression symptoms would be confounded by this lack of scalar invariance, precluding interpretable group-difference analyses.

Similar to when metric noninvariance is found, finding scalar noninvariance leaves researchers with three options: (a) investigate the items driving the noninvariance by sequentially removing or adding item intercept constraints and retesting the models until a partially invariant model is found, (b) remove items with noninvariant intercepts and retest the configural, metric, and scalar invariance models, or (c) conclude the construct is noninvariant and forego group difference testing [9]. *As illustrated below, differential associations between a biological mechanism and the mean levels of individual symptoms on a measure invariably induces scalar noninvariance.* In fact, it is analogous to the definition of scalar noninvariance, highlighting a potential limitation of much extant research in biological psychiatry.

Demonstration Using a Theoretical Example and Simulation

Imagine a scenario in which a researcher tests whether individuals with atypically elevated CRP report more depression symptoms on the Patient Health Questionnaire (PHQ)-9 [10] as compared to individuals with normative levels of CRP. Findings suggest that CRP levels are specifically related to changes in appetite and increased fatigue and no other depression symptoms on the PHQ-9 [6]. If the researcher simply summed the items on the PHQ-9 and compared group differences, it is possible that they would find a statistically significant mean

difference that could, at least in part, be driven by differences in these two specific symptoms.

Further, because we would expect that the items measuring changes in appetite and fatigue would have a systematically higher rate of endorsement (i.e., higher intercepts/scalar noninvariance) in the elevated CRP group relative to the non-elevated CRP group, identical sum scores across groups likely reflect different symptom profiles. Consequently, although there might be a statistically significant difference between the group means, these means are reflective of different depression constructs (e.g., one where endorsement of all nine symptoms is approximately equal, and one where changes in appetite and fatigue are featured proportionally more than the other seven symptoms), confounding inferences about comparisons of total depression scores between groups. It is important to note that, although a group-differences design is used in this example, measurement noninvariance can exist as a function of a continuous variable (for a description of moderated nonlinear factor analyses, see [10]). Furthermore, although we have focused on scalar noninvariance because it is invariably induced by unequal associations between a risk factor of interest and mean levels of individual symptoms on a measure, it is possible that certain biological mechanisms also are associated with other types of noninvariance (e.g., configural or metric).

As a didactic resource, annotated R code has been provided in Supplemental Materials to simulate 100 versions each of two different datasets: one with group differences in a subset of variables (henceforth referred to as “symptoms”) and a second with group differences in all symptoms, along with tests of the three types of measurement invariance described above. Only the dataset with the group differences in a subset of symptoms consistently will have scalar noninvariance (in 100% of simulations, compared to only 2% when there was an equal group difference across all symptoms), and this will be the only type of noninvariance that

systematically differs between the datasets.

As a follow-up to illustrate how scalar invariance can lead to false conclusions about the broader construct the items measure, group differences in the latent symptom total score were tested in the datasets with the systematic group difference in just a subset of symptoms. Even though the simulated datasets were not simulated to have differences at the latent factor level—and, therefore, we would expect a false-positive group-difference in approximately 5% of samples given a conventional alpha of .05—a significant group-difference in the latent factor was observed in 63% of simulations. In addition to illustrating the issues considered in this article, the code can be adapted to test for measurement invariance in readers' own data.

Moving Forward

We have used evidence for inflammatory phenotypes of depression [1, 12] as an illustrative example of how unequal associations between a given biological mechanism and different symptoms on a measure induces scalar noninvariance; however, this is a relevant concern for several subfields in psychiatry. For example, polygenetic risk scores for schizophrenia are primarily associated with positive psychotic symptoms [13]. Additionally, symptom-level endorsement of depression in women varies as a function of early vs. late onset, presence/absence of a family history of major depressive disorder, and exposure to adversity [14]. Several reproductive biomarkers have shown unequal associations with perinatal depression symptoms [15]. Further, differences in brain matter volume have domain-specific associations with obsessive-compulsive traits (e.g., less right insula volume associated with higher "contamination/washing"; [16]), and symptom-specific associations with depression (e.g., hippocampal volume is positively associated with loss of interest and irritability, but negatively associated with changes in appetite and sadness; [17]). As a consequence, all of these

subfields might be affected by unconsidered issues of measurement noninvariance.

Unfortunately, there are still many biological mechanisms that have not been investigated using symptom-specific approaches; thus, the true breadth of this problem is unknown. However, given increasing evidence across psychopathologies and biological mechanisms that not all symptoms within a disorder have the same risk factors, it is plausible that measurement noninvariance is a pervasive issue in biological psychiatry. To this end, it is imperative that biological psychiatry tests units of measurement smaller than diagnoses and total symptom scores [8]. By diversifying the level of psychopathological measurement explored, it will be possible to determine at what level biology-psychopathology associations most consistently exist (i.e., diagnosis vs. subscale vs. symptom). Among other benefits, this approach can provide insight into which specific subfields might suffer from the measurement noninvariance induced by unequal associations between a given biological mechanism and the symptoms of a disorder [8].

With these points in mind, we conclude with some recommendations to facilitate the exploration of measurement noninvariance as a function of biological measures and strategies to navigate this issue should it be found: First, *test for measurement noninvariance of symptom measures as a function of biological mechanisms* to identify subfields for which this is a concern that needs to be addressed. Second, *when measurement noninvariance is found, modify analyses as appropriate*. For example, adjust model constraints as described above to create models with measurement invariance (for more detailed information, see [9]). Alternatively, one could adopt a symptom-level approach to avoid aggregating noninvariant items and explore differential biology-symptom associations. For example, it is possible to specify models where a risk factor is associated with individual symptoms in addition to/rather than a latent variable. Another

possible option is to use network analysis. Third, *when analyzing heterogenous psychopathological constructs, explore multiple levels of measurement* (e.g., total score vs. subscale vs. specific items of a symptom measure). This will help isolate at which level of measurement a biological mechanism is associated with a behavioral phenotype and at what level it might be appropriate to aggregate similarly associated components.

Conclusion

In conclusion, growing evidence suggests that many biological mechanisms are unequally associated with symptoms in a given diagnostic category. As demonstrated above, these biological phenotypes of psychopathology can induce measurement noninvariance, which precludes valid comparison of sum scores on a measure as a function of the associated biological construct. Looking forward, researchers should explicitly test the possibility for measurement noninvariance before analyzing aggregate symptom measures and continue the investigation of biological phenotypes of psychopathology using symptom-level techniques.

Acknowledgements: Daniel P. Moriarity was supported by National Research Service Award F31 MH122116 and an APF Visionary Grant. Keanan J. Joyner was supported by a Ford Foundation Predoctoral Fellowship administered by the National Academy of Sciences, Engineering, and Medicine and National Institute of Drug Abuse R36 DA050049. George M. Slavich was supported by National Institutes of Health grant K08 MH103443. Lauren B. Alloy was supported by National Institute of Mental Health R01 MH101168.

Conflicts of interest: None

References

1. Majd M, Saunders EFH, Engeland CG. Inflammation and the dimensions of depression: A review. Front Neuroendocrinol. 2020;56.
2. Fried EI, Nesse RM, Zivin K, Guille C, Sen S. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. Psychol Med. 2014;44:2067–2076.
3. Slavich GM, Irwin MR. From stress to inflammation and major depressive disorder: A social signal transduction theory of depression. Psychol Bull. 2014;140:774–815.
4. Mac Giollabhui N, Ng TH, Ellman LM, Alloy LB. The longitudinal associations of inflammatory biomarkers and depression revisited: Systematic review, meta-analysis, and meta-regression. Mol Psychiatry. 2020:1–13.
5. Fried EI, von Stockert S, Haslbeck JMB, Lamers F, Schoevers RA, Penninx BWJH. Using network analysis to examine links between individual depressive symptoms, inflammatory markers, and covariates. Psychol Med. 2019. 2019. https://doi.org/https://doi.org/10.31234/osf.io/84ske.
6. Moriarity DP, Horn SR, Kautz MM, Haslbeck JM, Alloy LB. How handling extreme C-reactive protein (CRP) values and regularization influences CRP and depression criteria associations in network analyses. Brain Behav Immun. 2021;91:393–403.
7. Milaneschi Y, Kappelmann N, Ye Z, Lamers F, Moser S, Jones PB, et al. Association of Inflammation with Depression and Anxiety: Evidence for Symptom-Specificity and Potential Causality from UK Biobank and NESDA Cohorts. Mol Psychiatry.
8. Moriarity DP, Alloy LB. Beyond diagnoses and total symptom scores: Diversifying the level of analysis in psychoneuroimmunology research. Brain Behav Immun. 2020;89:1–2.

9. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Dev Rev. 2016;41:71–90.

10. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606–613.

11. Bauer D. A More General Model for Testing Measurement Invariance and Differential Item Functioning. Psychol Methods. 2017;22:507–526.

12. Dooley LN, Kuhlman KR, Robles TF, Eisenberger NI, Craske MG, Bower JE. The role of inflammation in core features of depression: Insights from paradigms using exogenously-induced inflammation. Neurosci Biobehav Rev. 2018;94:219–237.

13. Isvoranu AM, Guloksuz S, Epskamp S, Van Os J, Borsboom D. Toward incorporating genetic risk scores into symptom networks of psychosis. Psychol Med. 2020;50:636–643.

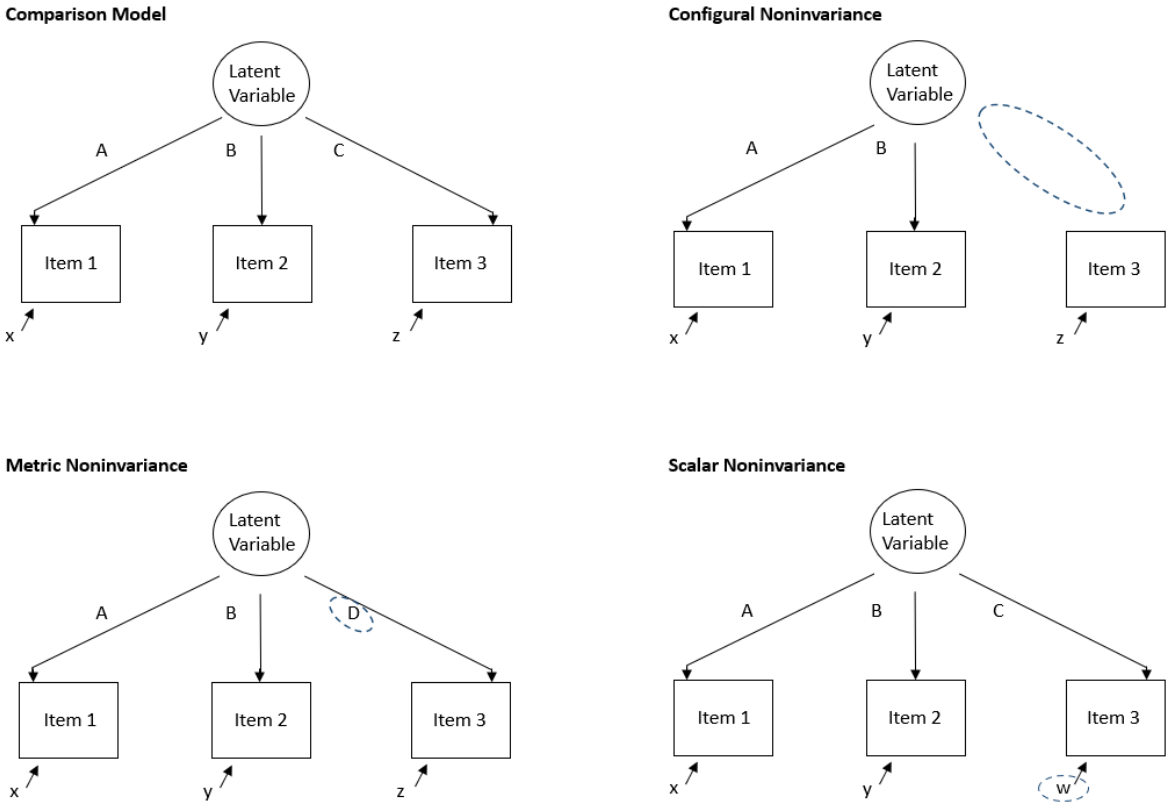
14. van Loo HM, Van Borkulo CD, Peterson RE, Fried EI, Aggen SH, Borsboom D, et al. Robust symptom networks in recurrent major depression across different levels of genetic and environmental risk. J Affect Disord. 2018;227:313–322.

15. Santos H, Fried EI, Asafu-Adjei J, Jeanne Ruiz R. Network structure of perinatal depressive symptoms in latinas: Relationship to stress and reproductive biomarkers. Res Nurs Heal. 2017;40:218–228.

16. Okada K, Nakao T, Sanematsu H, Murayama K, Honda S, Tomita M, et al. Biological heterogeneity of obsessive-compulsive disorder: A voxel-based morphometric study based on dimensional assessment. Psychiatry Clin Neurosci. 2015;69:411–421.

17. Hilland E, Landrø NI, Kraft B, Tamnes CK, Fried EI, Maglanoc LA, et al. Exploring the links between specific depression symptoms and brain structure: A network study. Psychiatry Clin Neurosci. 2020;74:220–221.

278



279

280 **Fig. 1 Visual representations of measurement noninvariance.** Top left panel = the comparison model. All other panels illustrate one form of measurement noninvariance relative to the comparison model. Focal differences associated with the specified type of noninvariance are highlighted by a dashed circle. Uppercase letters = factor loadings, lowercase letters = intercepts.