

1 This paper is now in press at *Molecular Psychiatry*. Please cite as

2

3 **Morarity, D.P.**, Joyner, K., Slavich, G. M., & Alloy, L. B. (in press). Unconsidered issues of measurement
4 noninvariance in biological psychiatry: A focus on biological phenotypes of psychopathology. *Molecular Psychiatry*.

5

6 Running Head: MEASUREMENT NONINVARIANCE IN PSYCHIATRY

7

8 **Unconsidered Issues of Measurement Noninvariance in Biological Psychiatry:**

9 **A Focus on Biological Phenotypes of Psychopathology**

10

11

12

13 Daniel P. Moriarity, M.A.¹; Keanan J. Joyner, M.A.²;

14 George M. Slavich, Ph.D.³; & Lauren B. Alloy, Ph.D.¹

15

16

17

18

19

20 ¹Department of Psychology, Temple University

21 ²Department of Psychology, Florida State University

22 ³Cousins Center for Psychoneuroimmunology and Department of Psychiatry and Biobehavioral

23 Sciences, University of California, Los Angeles

24

25

26 Correspondence concerning this article should be addressed to Daniel P. Moriarity, E-

27 mail: Daniel.moriarity@temple.edu

28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

Abstract

There is increasing appreciation that certain biological processes 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, the same way, 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, then it is invalid to directly compare means on that 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). We also show 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.

45 **Introduction**

46 Many research questions in biological psychiatry use variables that index processes such
47 as inflammatory activity, grey matter volume, and gene expression as predictors of an aggregate
48 measure of psychopathology. An underlying assumption of these tests, as commonly performed,
49 is that the psychopathology measure used assesses the same construct the same way each time it
50 is administered, either across different people or across different time points for the same
51 individual. However, this assumption might be untenable in light of growing evidence that some
52 biological risk factors have differential associations with symptoms within a diagnostic construct
53 (e.g., inflammatory proteins being most robustly associated with neurovegetative symptoms of
54 depression [1]).

55 In this article, we first briefly describe the concept of biological phenotypes. Second, we
56 discuss the concept of measurement invariance. Third, we use both a theoretical example and
57 statistical simulation to illustrate how the presence of biological phenotypes of psychopathology
58 induces measurement noninvariance. We also discuss how this issue can result in inappropriate
59 conclusions about the relations between biology and behavior. Finally, we provide some
60 recommendations for moving forward.

61 **Biological Phenotypes of Psychopathology**

62 There is accumulating evidence that different psychiatric symptoms within some
63 diagnostic categories (e.g., depression) may have different risk factors [2]. These findings have
64 prompted interest in the symptom-level biological phenotyping of psychopathology (see Fig. 1
65 for an example of a nine-item measure of depression for which a risk factor is only related to
66 three items). The thorough characterization of which specific symptoms of a disorder are
67 associated with a given process may in turn help advance biological psychiatry and, in addition,

68 precision medicine. For example, understanding that inflammation is associated primarily with
69 neurovegetative symptoms of depression [1] can help clinicians identify patients who may
70 possess an underlying atypical inflammatory phenotype, and this information can, in turn, guide
71 decisions about who might benefit most from adjunctive anti-inflammatory treatments [3].
72 Further, this level of specificity will improve insight into whether biology—behavior
73 associations are disorder specific or transdiagnostic in nature. For example, does irritability as an
74 indicator of depression have the same biological correlates as irritability as an indicator of
75 bipolar disorder or borderline personality disorder, and within non-clinical samples?

76 Studying biological phenotypes of psychopathology also has the potential to improve the
77 replicability of psychiatric research [4]. Again, using an immunologic example, consider that the
78 effect sizes between C-reactive protein (CRP) and depression symptoms across published studies
79 is highly variable across studies [5]. Given evidence that CRP is not equally associated with all
80 depression symptoms [6–8], inconsistent results between CRP and total depression symptoms are
81 likely influenced by the sampling variability of which symptoms are endorsed across studies.
82 Guided by phenotyping research, making psychiatric outcomes more nuanced or specific (i.e.,
83 specific symptoms or subtypes of depression) may increase replicability and shorten the research-
84 to-practice timeline for syndromes that are characterized by high degrees of heterogeneity [10].

85 The implications of differential associations between a risk factor and the symptoms of a
86 disorder extend beyond etiology, nosology, and treatment. Below, we examine an important
87 methodological concern that has been largely ignored in extant discourse on phenotyping:
88 measurement noninvariance. We will continue using the example of inflammation and
89 depression to contextualize the issue of measurement noninvariance, discuss its consequences,
90 and describe potential solutions. However, the issue of measurement invariance is universally

91 applicable to all risk factors that are unequally associated with different symptoms on a measure.

92 **Measurement Invariance: A Brief Overview**

93 In the context of psychological questionnaires, measurement invariance is the ability of a
94 questionnaire to measure the same construct in the same way regardless of who takes it (e.g., people
95 from two different groups) or when it is completed (e.g., same person at multiple points in a
96 longitudinal study). Measurement invariance also can exist as a function of continuous variables
97 (e.g., age). To keep the language consistent, we will focus on measurement invariance across groups.

98 Without measurement invariance, it is inappropriate to compare means—the most common
99 level of analysis in biological psychiatry—because identical scores might not reflect the same level
100 of a construct for both groups. As an example, consider if Person A steps on a scale on Earth and
101 their weight is displayed in pounds and Person B steps on the same scale and the weight is displayed
102 in pounds, but the scale is located on the moon. Despite using the same exact measurement
103 instrument, these two numbers cannot be directly compared because the weight registered by the
104 scale is influenced by a third factor—in this case, different levels of gravity.

105 The three most commonly discussed types of measurement invariance are configural, metric
106 (sometimes referred to as “weak” invariance), and scalar invariance (sometimes referred to as
107 “strong” invariance). We will briefly discuss configural and metric invariance, but focus on scalar
108 invariance, for reasons described below. See Fig. 2 for visualization of the three kinds of
109 measurement invariance and [11] for a more thorough review of measurement invariance and how to
110 test it. Also, lest researchers assume that because they do not model their psychopathology variables
111 in latent space and instead use sum scores, they are immune from the challenges raised herein, we
112 want to highlight that sum scores actually are themselves latent variables (for an in-depth explainer,
113 see [12]). Quoting from the authors of that article, “sum scoring corresponds to a statistical model

114 and is not a model-free arithmetic calculation.” Mathematically, a sum score from a set of items is a
115 latent variable model that fixes all loadings and error variances to equivalence across items (among
116 other assumptions). As such, all of the issues being discussed here are just as applicable to
117 differences in sum scores as they are to latent variable means.”

118 Configural invariance, the least strict form of measurement invariance, refers to
119 equivalence of model form. That is, which variables (e.g., items) load onto which latent variables
120 (e.g., depression) does not change as a function of a third variable (e.g., elevated inflammatory
121 levels). An example of configural noninvariance is if all nine items on a depression questionnaire
122 load onto the depression latent factor in a sample with normative inflammation, but only eight of
123 the items load onto the depression factor in a group with elevated inflammation. If configural
124 invariance is supported, the next form of invariance to check is metric. Metric invariance, in turn,
125 refers to the equivalence of item loadings (how much an item is associated with a factor). For
126 example, suppose item #9 had a loading of .3 on the depression factor in a sample with
127 normative inflammation, but had a loading of .6 in a group with elevated inflammation.

128 If both configural and metric invariance are supported, the next step is to test for scalar
129 invariance for the items with metric invariance. Scalar invariance refers to equality of item
130 intercepts/thresholds (i.e., what level of endorsement of an item to expect if the latent variable
131 associated with the item is 0). For example, consider a sample in which, when the true latent
132 score of depression is 0, none of the items are endorsed. An example of scalar noninvariance
133 would be if, in a different sample (e.g., one with elevated inflammation), when the true latent
134 score of depression was 0, there would still be a couple of items likely to be endorsed (because
135 they are attributable to inflammation instead of depression). If item intercepts differ between
136 groups, then observed mean differences in the construct (e.g., depression) do not accurately

137 capture true mean differences in the latent variable (see below for an illustration). Therefore, if
138 scalar invariance is not met, any statistical test comparing mean differences on the total number
139 of depression symptoms would be confounded by the lack of scalar invariance, precluding
140 interpretable group-difference analyses. As illustrated below, *unequal associations between a*
141 *variable and individual items on a measure will always induce scalar noninvariance*. In fact, it is
142 analogous to the definition of scalar noninvariance, highlighting a potential limitation of much
143 extant research in biological psychiatry.

144 **A Theoretical Example and Simulation of Measurement Noninvariance**

145 Imagine a scenario in which a researcher tests whether individuals with atypically
146 elevated CRP report more depression symptoms on the Patient Health Questionnaire (PHQ)-9
147 [13] as compared to individuals with normative levels of CRP. The findings suggest that CRP
148 levels are specifically related to changes in appetite and increased fatigue and no other
149 depression symptoms on the PHQ-9 [7]. If the researcher simply summed the items on the PHQ-
150 9 (or used them to load onto a single, latent variable of depression) and then tested group
151 differences, it is possible that they would find a statistically significant mean difference that
152 could, at least in part, be driven by actual differences in these two specific symptoms. Further,
153 because we would expect that the items measuring changes in appetite and fatigue would have a
154 systematically higher rate of endorsement (i.e., higher intercepts/scalar noninvariance) in the
155 elevated CRP group relative to the non-elevated CRP group, identical scores across these groups
156 likely reflect different symptom profiles. Consequently, although there might be a statistically
157 significant difference between the group means, these means are reflective of different
158 depression constructs (e.g., one where endorsement of all nine symptoms is approximately equal,
159 and one where changes in appetite and fatigue are featured proportionally more than the other

160 symptoms), confounding inferences about differences in total depression scores between groups.
161 It is important to reiterate that, although a group-differences design is used in this example,
162 measurement noninvariance can exist as a function of a continuous variable (for a description of
163 moderated nonlinear factor analyses, see [10]). We recommend using such approaches when
164 there are not clinically/theoretically meaningful cut-offs for biological variables of interest.
165 Furthermore, although we have focused on scalar noninvariance because it is invariably induced
166 by unequal associations between a risk factor of interest and mean levels of individual symptoms
167 on a measure, it is possible that certain biological processes also are associated with other types
168 of noninvariance (e.g., configural or metric).

169 As a didactic resource, annotated R code that can be used to simulate 100 versions each
170 of two different datasets, each consisting of two groups (representative of the theoretical elevated
171 and non-elevated CRP groups above) with 250 participants each, is provided in the Supplemental
172 Materials. The first dataset has group differences for only a subset of three variables (henceforth
173 referred to as “symptoms”); the second dataset has group differences for all of the symptoms
174 measured (i.e., the high-risk group only increased risk for 3/9 symptoms in the first dataset, but
175 equally increased risk for 9/9 symptoms in the second dataset). Tests of the three types of
176 measurement invariance described above also are provided. Only the dataset that yields group
177 differences in a subset of symptoms consistently has scalar noninvariance (i.e., in 100% of the
178 simulations conducted, compared to only 2% of the simulations when there was an equal group
179 difference across all symptoms). Notably, this is the only type of noninvariance that
180 systematically differs between the datasets.

181 As a follow-up to illustrate how scalar noninvariance can lead to false conclusions about
182 the broader construct that items measure, group differences in the latent symptom total score

183 were tested in the datasets available in the Supplemental Materials with the systematic group
184 difference present for just a subset of symptoms. Even though the simulated datasets were not
185 simulated to have differences at the latent factor level—and, therefore, we would expect a false-
186 positive group-difference for approximately 5% of the samples given a conventional alpha of
187 .05—a significant group-difference in the latent factor was observed in 63% of simulations.
188 Therefore, there was a greatly inflated risk of falsely concluding group-differences in the latent
189 factor when scalar noninvariance was present. In addition to illustrating the issues considered in
190 this article, the code can be adapted to test for measurement invariance in readers' own data.

191 **Moving Forward**

192 We have used the example of inflammatory phenotypes of depression [1, 15] to illustrate
193 how unequal associations between a given biological process and different symptoms on a
194 measure induces scalar noninvariance; however, this is a relevant concern for several subfields in
195 psychiatry. For example, polygenetic risk scores for schizophrenia are primarily associated with
196 positive psychotic symptoms [16]. Additionally, symptom-level endorsement of depression in
197 women varies as a function of early vs. late onset Major Depressive Disorder, presence/absence
198 of a family history of major depressive disorder, and exposure to adversity [17]. Several
199 reproductive biomarkers also have shown unequal associations with perinatal depression
200 symptoms [18]. Further, differences in grey matter volume have domain-specific associations
201 with obsessive-compulsive traits (e.g., less right insula volume associated with higher
202 "contamination/washing"; [18]), and symptom-specific associations with depression (e.g.,
203 hippocampal volume is positively associated with loss of interest and irritability, but negatively
204 associated with changes in appetite and sadness; [19]). Therefore, research on all of these topics
205 might be affected by unconsidered issues of measurement noninvariance.

206 Because there are many biological processes that have not yet been investigated using
207 symptom-specific approaches, the true breadth of this problem is unknown. However, given
208 increasing evidence across psychopathologies and biological processes that not all symptoms of
209 a disorder have the same risk factors, it is plausible that measurement noninvariance is pervasive
210 in biological psychiatry. Testing measurement invariance can provide insight into which specific
211 subfields of psychiatry—or areas in psychiatry and psychology more generally—may be missing
212 differential associations between biological processes and symptoms of specific disorders. To
213 this end, it is imperative that biological psychiatry tests units of measurement smaller than
214 diagnoses and total symptom scores [10]. By diversifying the level of psychopathological
215 measurement explored, it will be possible to determine at what level biology-psychopathology
216 associations most consistently exist (i.e., diagnosis vs. subscale vs. symptom).

217 With these points in mind, we conclude with some recommendations to facilitate the
218 exploration of measurement noninvariance as a function of biological measures and strategies to
219 navigate this issue should it be found: First, *test for measurement noninvariance of symptom*
220 *measures as a function of biological processes* to identify subfields for which this is a concern
221 that needs to be addressed. Second, *when measurement noninvariance is found, modify analytic*
222 *strategy as appropriate*. It is important to emphasize that ideal choice of analytic adjustment is
223 influenced by a few considerations including: the type(s) of noninvariance observed, sample size,
224 diagnostic philosophy, and the specific nuances of one's research question. For example, it is
225 possible to adjust model constraints to create latent models with measurement invariance (for
226 more details, see [11]). However, especially in the case of scalar noninvariance, noninvariance is
227 an indication to select a more detailed analytic approach. Options include: hierarchical models
228 (e.g., within the HiTOP framework [21]) where biology predicts multiple levels of measurement

229 (e.g., total score, subscales, specific symptoms), analyses taking a differential item functioning
230 approach (i.e., tests how the probability of endorsing an item might change as a function of a
231 biological variable), or symptom-level analyses. Third, *when analyzing heterogenous*
232 *psychopathological constructs, consider exploring multiple levels of measurement* (e.g., total
233 score vs. subscale vs. specific items of a symptom measure, as described above) *as an a priori*
234 *analytic strategy* [10]. This will help isolate at which level of measurement a biological process
235 is associated with a behavioral phenotype and at what level it might be appropriate to aggregate
236 similarly associated components. Further, this level of inquiry protects against problems with
237 diagnostic heterogeneity [9] and facilitates dimensional conceptualizations of mental illness
238 consistent with RDoC [22].

239 At the same time, it is important to note that many extant measures were explicitly
240 created with unidimensional sum scores/latent variables in mind and the psychometrics of the
241 constituent parts of the measures (subscales, individual items) must also be considered.
242 Additionally, many biological psychiatry studies have small sample sizes that may preclude tests
243 of measurement invariance. Approximating the necessary sample size for tests involves
244 understanding of the underlying factor structure and strength of the relation between the
245 biological and psychological variables of interest, and thus, is outside the scope of this paper. It
246 is also important to note that sample size influences ideal fit indices, a topic that is described in
247 more detail in [11]. Further, should a sample be underpowered for measurement invariance, it is
248 advisable to consider preliminary analyses in a larger dataset, perhaps one with openly-available
249 data (e.g., Midlife in the United States [23], UK Biobank [24], Adolescent Brain Cognitive
250 Development Study [25]).

251 **Conclusion**

252 In conclusion, growing evidence suggests that many biological processes are unequally
253 associated with symptoms in a given diagnostic category. As demonstrated above, these
254 biological phenotypes of psychopathology can induce measurement noninvariance, which
255 precludes valid comparison of unidimensional sum scores/latent variables on a measure as a
256 function of the associated biological construct being assessed. Looking forward, researchers
257 should explicitly test for measurement noninvariance before analyzing aggregate symptom
258 measures and continue investigating biological phenotypes of psychopathology using more
259 detailed analytic techniques.

260 **Acknowledgements:** Daniel P. Moriarity was supported by National Research Service Award
261 F31 MH122116 and an APF Visionary Grant. Keanan J. Joyner was supported by a Ford
262 Foundation Predoctoral Fellowship administered by the National Academy of Sciences,
263 Engineering, and Medicine and National Institute of Drug Abuse R36 DA050049. George M.
264 Slavich was supported by National Institutes of Health grant K08 MH103443 and by grant
265 OPR21101 from the California Initiative to Advance Precision Medicine. Lauren B. Alloy was
266 supported by National Institute of Mental Health R01 MH101168.

267 **Conflicts of interest:** None

268

269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291

References

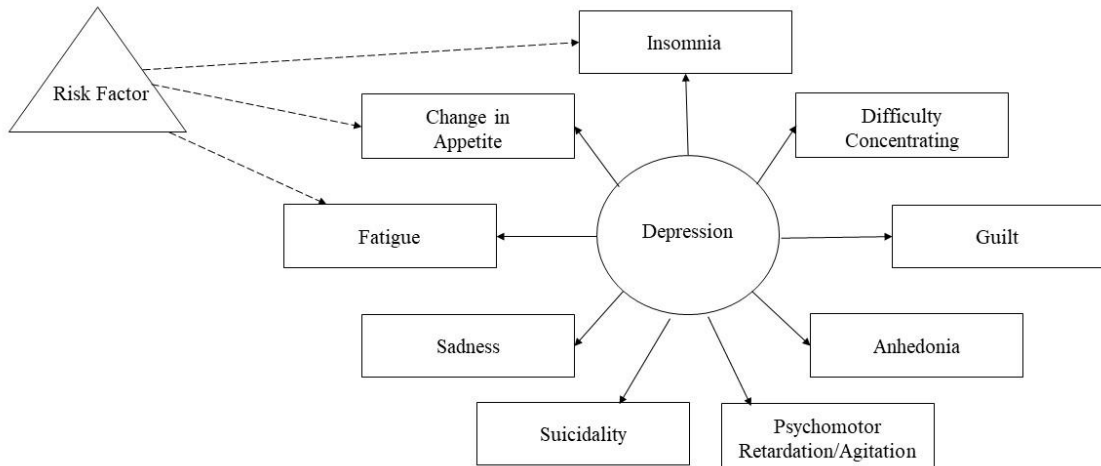
1. Majd M, Saunders EFH, Engeland CG. Inflammation and the dimensions of depression: A review. *Front Neuroendocrinol.* 2020;56.
2. Fried EI, Nesse RM, Zivin K, Guille C, Sen S. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychol Med.* 2014;44:2067–2076.
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. Moriarity DP. Building a replicable and clinically-impactful immunopsychiatry: Methods, phenotyping, and theory integration. *Brain, Behav Immun - Heal.* 2021;16.
5. 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.
6. 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>.
7. 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.
8. 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.*

- 292 <https://doi.org/10.1101/2021.01.08.20248710>.
- 293 9. Feczko E, Miranda-dominguez O, Marr M, Graham AM, Nigg JT, Fair DA. The
294 Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends Cogn Sci*.
295 2019;1–18.
- 296 10. Moriarity DP, Alloy LB. Beyond diagnoses and total symptom scores: Diversifying the
297 level of analysis in psychoneuroimmunology research. *Brain Behav Immun*. 2020;89:1–2.
- 298 11. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: The state
299 of the art and future directions for psychological research. *Dev Rev*. 2016;41:71–90.
- 300 12. McNeish D, Wolf MG. Thinking twice about sum scores. *Behav Res Methods*.
301 2020;52:2287–2305.
- 302 13. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity
303 measure. *J Gen Intern Med*. 2001;16:606–613.
- 304 14. Bauer D. A More General Model for Testing Measurement Invariance and Differential
305 Item Functioning. *Psychol Methods*. 2017;22:507–526.
- 306 15. Dooley LN, Kuhlman KR, Robles TF, Eisenberger NI, Craske MG, Bower JE. The role of
307 inflammation in core features of depression: Insights from paradigms using exogenously-
308 induced inflammation. *Neurosci Biobehav Rev*. 2018;94:219–237.
- 309 16. Isvoranu AM, Guloksuz S, Epskamp S, Van Os J, Borsboom D. Toward incorporating
310 genetic risk scores into symptom networks of psychosis. *Psychol Med*. 2020;50:636–643.
- 311 17. van Loo HM, Van Borkulo CD, Peterson RE, Fried EI, Aggen SH, Borsboom D, et al.
312 Robust symptom networks in recurrent major depression across different levels of genetic
313 and environmental risk. *J Affect Disord*. 2018;227:313–322.
- 314 18. Santos H, Fried EI, Asafu-Adjei J, Jeanne Ruiz R. Network structure of perinatal

- 315 depressive symptoms in latinas: Relationship to stress and reproductive biomarkers. *Res*
316 *Nurs Heal.* 2017;40:218–228.
- 317 19. Okada K, Nakao T, Sanematsu H, Murayama K, Honda S, Tomita M, et al. Biological
318 heterogeneity of obsessive-compulsive disorder: A voxel-based morphometric study based
319 on dimensional assessment. *Psychiatry Clin Neurosci.* 2015;69:411–421.
- 320 20. Hilland E, Landrø NI, Kraft B, Tamnes CK, Fried EI, Maglanoc LA, et al. Exploring the
321 links between specific depression symptoms and brain structure: A network study.
322 *Psychiatry Clin Neurosci.* 2020;74:220–221.
- 323 21. Kotov R, Krueger RF, Watson D, Bagby M, Carpenter WT, Caspi A. The Hierarchical
324 Taxonomy Of Psychopathology (HiTOP). *J Abnorm Psychol.* 2017:1–83.
- 325 22. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine D, Quinn K, et al. Research Domain
326 Criteria (RDoC): Toward a new classification framework for research on mental disorders.
327 *Am J Psychiatry Online.* 2010;167:748–751.
- 328 23. Ryff CD, Seeman T, Weinstein M. Midlife in the United States (MIDUS 2): Biomarker
329 Project, 2004-2009. Ann Arbor, MI Inter-University Consort Polit Soc Res [Distributor].
330 2017:10.
- 331 24. Allen NE, Sudlow C, Peakman T, Collins R. UK Biobank Data: Come and Get It. *Sci*
332 *Transl Med.* 2014;6:4–7.
- 333 25. Volkow ND, Koob GF, Croyle RT, Bianchi DW, Gordon JA, Koroshetz WJ, et al. The
334 conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev*
335 *Cogn Neurosci.* 2018;32:4–7.

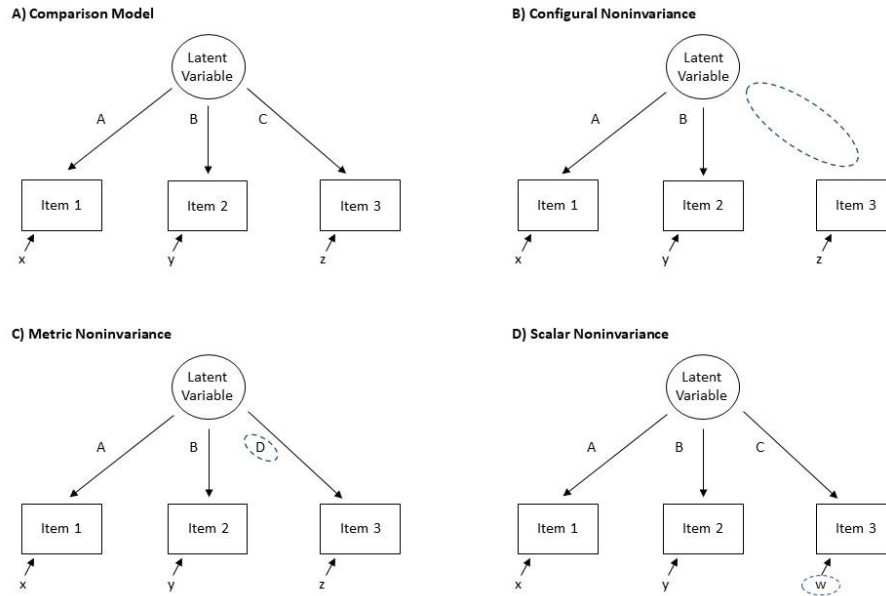
336

337



338

339 Fig. 1. Visual representation of a risk factor associated with a subset of symptoms. The triangle
340 represents a risk factor, rectangles indicate individual depression symptoms, and the circle
341 represents depression. Solid lines connect the individual items and depression. Dashed lines
342 represent relations between a risk factor and a subset of items.



343

344 Fig. 2. Visual representations of measurement noninvariance. **A)** The comparison model, **B)**345 configural noninvariance, **C)** metric noninvariance, and **D)** scalar noninvariance. Focal

346 differences associated with the specified type of noninvariance are highlighted by a dashed

347 circle. Uppercase letters = factor loadings; lowercase letters = intercepts.