Modeling and Ontology of Quantitative Dimensions of Psychopathology

Michael E. Aristodemou

Eiko I. Fried, Supervisor

Leiden University

August 15th, 2019

Author note

Correspondence concerning this article should be addressed to Michael E. Aristodemou, BSc, Department of Clinical Psychology, Leiden University, 2300 RA Leiden, the Netherlands. Contact: m.e.aristodemou@umail.leidenuniv.nl

Word count: 7712

Abstract

A hierarchical taxonomy of psychopathology aims to replace traditional disorder categories with a hierarchy of increasingly general dimensions that arise from the statistical covariation among symptoms. However, the ontogeny and ontology of these dimensions remain contentious. We analyzed two large longitudinal datasets to examine developmental changes in two dimensions of psychopathology, at two distinct levels of the hierarchical taxonomy, at two distinct developmental periods: the p-factor from early to late adolescence and major depressive disorder from middle adulthood to old age. We used latent change score models to directly compare the ability of two long-standing theories—the *common cause theory* and the *dynamic mutualism theory*—to explain the development of the two dimensions of psychopathology. A dynamic mutualism model best explained the development of the p-factor, while both models provided equally good explanations for the development of major depressive disorder. But neither model could provide a sufficient explanation for the development of either dimension. We show that computational models offer a promising tool to improve mechanistic theories of psychopathology and suggest that progress may lie at the interface between multiple causes.

Keywords: mutualism; p-factor; depression; longitudinal modeling; hierarchical taxonomy; nosology

Modeling and Ontology of Quantitative Dimensions of Psychopathology

Concern with the utility of traditional classification has inspired a data-driven movement that aims to reinvent psychiatric taxa. Proponents of this movement use factor analysis to describe empirical patterns of covariation among symptoms of psychopathology (Caspi et al., 2014; Kotov et al., 2017; Lahey et al., 2012; Lahey, Krueger, Rathouz, Waldman, & Zald, 2017). This results in a hierarchy of increasingly general dimensional entities (e.g. hierarchical taxonomy of psychopathology; Kotov et al., 2017). The most general of which, termed the p-factor due to its conceptual resemblance to the g-factor of intelligence, reflects variance shared by any and all types of mental illness (Caspi et al., 2014; Caspi & Moffitt, 2018).

Hierarchical factor models have been consistently shown to describe the pattern of symptom covariation well, 'but the statistical models are agnostic about—and certainly do not reveal—the causes of these correlations' (Caspi & Moffitt, 2018, p.3). There are multiple plausible data-generating mechanisms, and multiple models can statistically describe the positive covariation among symptoms, known as the *positive manifold*, equally well (Kruis & Maris, 2016; Marsman, Maris, Bechger, & Glas, 2015; van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017; Van Der Maas et al., 2006). This limits a purely data-driven search for the data-generating mechanism and supports the importance of substantive theory. Theories make distinct predictions about the mechanisms that drive the development of the positive manifold in psychopathology, and innovations in structural equation modeling (McArdle, Hamagami, Meredith, & Bradway, 2000; McArdle & Hamagami, 2001) make it possible to formalize these mechanisms and directly compare them. This allows for an unprecedented comparison of long-standing theoretical mechanisms that vie to be the foundational framework for psychiatric classification and opens an inroad to the ontology of mental disorders.

The present paper aims to advance knowledge on how to best model and understand the positive manifold in psychopathology. We limit the conceptual space to two theories that have garnered significant traction in recent literature to investigate which theory provides the most accurate explanation for the development of the positive manifold in psychopathology—the *common cause theory* or the *dynamic mutualism theory*? Each theory will be translated into a statistical model using different latent change score models (McArdle & Hamagami, 2001), and these models will be directly compared. To our knowledge, no study to date has *directly* compared common cause theory with dynamic mutualism theory, within the context of psychopathology. In the succeeding sections, we present the two theories and their corresponding models. We examine previous attempts to distinguish these developmental accounts and present our work that aims to further existing knowledge.

Common cause theory

The c*ommon cause theory* posits that symptoms within, but also across, traditional syndromes covary because they share a singular common cause (p-factor), on top of more specific causes shared by subsets of syndromes such as the internalizing and externalizing spectra (Caspi & Moffitt, 2018). The theory that symptoms covary because they share a common cause is a plausible candidate that represents the status quo when inferring the cause of traditional syndromes (e.g. Insel & Cuthbert, 2015; Reise & Waller, 2009). A common cause interpretation does not follow as a mathematical necessity from the models used to construct the hierarchical taxonomy (Jonas & Markon, 2016). But some have argued that it is the only defensible interpretation because common factor models treat variance that is *not* shared by a latent variable's items as error (Van Bork, Wijsen, & Rhemtulla, 2017).

The use of a common factor model when a different model is appropriate carries substantial implications. First, the misuse of factor models can lead to inappropriate inferences made from structural equation models (Rhemtulla, van Bork, & Borsboom, 2019).

Take the case where we want to measure the relationship between depression and academic performance (AP), using a factor model with sleeplessness as an indicator of depression. This relationship would be overestimated because the unique relationship between sleeplessness and AP would be wrongly attributed to the relationship between AP and depression. Second, if all symptoms used to measure a common factor are not interchangeable indicators of their root cause, then studies using different items may not assess the same construct (Watts, Poore, & Waldman, 2019). Third, a common factor that does not reflect a common cause is a vacuous construct, in the sense that the position on the latent variable is not informative of the process that led to the person's response (Borsboom, Mellenbergh, & Van Heerden, 2003). Vacuous common factors provide limited insight into mechanisms that can be targeted to improve treatment (Aristodemou & Fried, in press). Hence, research that aims to explain the development of psychopathology using hierarchical factor models, seems to implicitly assume that dimensions within the hierarchy are made up of congeneric items.

Dynamic mutualism theory

The *dynamic mutualism theory* offers a plausible alternative by proposing that the positive manifold arises through dynamic processes that occur throughout development (Van Der Maas et al., 2006). This theory originates from research on human intelligence where empirical evidence strongly suggests that mutualistic coupling between different cognitive domains forms an essential developmental mechanism (Kievit, Hofman, & Nation, 2019; Kievit et al., 2017).

Within the domain of psychopathology, dynamic mutualism theory explains the development of the positive manifold by stating that symptoms covary because they cause each other (*network theory*; Borsboom, 2008; Borsboom & Cramer, 2013; Cramer, Waldorp, Van der Maas, & Borsboom, 2010; Borsboom, 2017; Fried et al., 2017; Mcnally, 2016). Some have argued that network theory is incompatible with a common cause interpretation because no singular causal mechanism can explain the coherence of symptoms within a network (Borsboom, Cramer, & Kalis, 2019). Moreover, network theory embraces the assumption that multiple causes can lead to the manifestation of a symptom (i.e. multiple realizability; Fodor, 1974; Horgan, 1993; Putnam, 1967; Pylyshyn, 1984). If these assertions hold, faulty categories and limited technology may not be at fault for our disappointing track record in identifying disorder-specific etiology (Borsboom et al., 2019). Instead, we should blame the monocausal model.

The network perspective does not only inspire skepticism about the common cause theory but also offers a methodological tool to study mental disorders as complex systems of causally active symptoms (Borsboom & Cramer, 2013). Potentially meaningful causal interrelations can be formalized using network models (Borsboom & Cramer, 2013; Epskamp, Borsboom, & Fried, 2018; Epskamp & Fried, 2018; Fried & Cramer, 2017). This works by controlling for the shared variance among a set of symptoms, to estimate a weighted network of unique associations that represent possible causal associations (Epskamp & Fried, 2018). The substantive meaning of these empirical associations, however, depends on the contentious ontology of mental disorders.

Common cause theory versus dynamic mutualism theory

At present we do not know which theory is (more) correct, and this harbors uncertainty about the best direction for clinical psychology and psychiatry. Common factor models and network models can offer equivalent descriptions of data (Kruis & Maris, 2016; Marsman et al., 2015), but their respective theories make different predictions about dynamic behavior that can be exploited using longitudinal data. However, only a few studies have compared common cause theory and dynamic mutualism theory within a longitudinal context, and none found preferential evidence for either theory (Greene & Eaton, 2017;

McElroy, Belsky, Carragher, Fearon, & Patalay, 2018; Murray, Eisner, & Ribeaud, 2016; Snyder, Young, & Hankin, 2017).

We commend the authors for their efforts but given the complexity of the topic, several challenges point to the merit of further investigation. First, one study specified hypotheses that may only be congruent with either theory if all other developmental processes are held equal (Murray et al., 2016). For instance, the authors hypothesized that dynamic mutualism theory predicts increasing symptom covariation over time because symptoms interact in a mutually reinforcing manner. But mutually reinforcing symptom interactions may co-occur with other developmental processes that lead to increasingly specific mental illness (McElroy et al., 2018). Hence, the validity of the dynamic mutualism theory does not necessitate increasing symptom covariation. Second, two studies relied on two waves with limited temporal coverage (i.e. 18-24 months; Greene & Eaton, 2017; Snyder, Young, & Hankin, 2017). Third, one study sampled from a wide age-range that covers divergent life periods (Greene & Eaton, 2017), which may be associated with different developmental mechanisms (Kievit et al., 2017). Fourth, a different study used cross-lagged panel models (McElroy et al., 2018) which are not well-suited to examine change (Kievit et al., 2017); and relied on maternal reports that are not well aligned with children’s responses once they are able to self-report (Waters, Steward-Brown, & Fitzpatrick, 2003). Lastly, and most importantly, none of the abovementioned studies has *directly* compared the two developmental mechanisms.

We aim to supplement past efforts by conducting a *direct* comparison between common cause theory and dynamic mutualism theory to find out which provides the most accurate account for the development of the positive manifold in psychopathology. We translate fundamental predictions made by each theory into latent change-score models (Kievit et al., 2018; McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; McArdle &

Hamagami, 2001) that are well-suited to study temporal dynamics and allow us to directly compare the two theories. We further extend prior work by comparing the two theoretical accounts at two distinct levels of the hierarchical taxonomy, at two distinct developmental periods, using two distinct longitudinal datasets. We have two main research questions, each corresponding to a different dataset:

(1) Which theory provides the most accurate account for the development of the positive manifold among dimensions that constitute the p-factor from early to late adolescence—common cause theory or dynamic mutualism theory?

(2) Which theory provides the most accurate account for the development of the positive manifold among symptoms that constitute major depressive disorder from middle adulthood to old age—common cause theory or dynamic mutualism theory?

Method

Dataset 1: z-proso

Participants. The sample was obtained from the Zurich Project on the Social Development of Children and Youths (z-proso). The z-proso is a longitudinal cohort and intervention study with a focus on the development of adaptive and maladaptive social behaviors. However, the data are treated as observational since early interventions had no substantial effects on children (e.g. Averdijk, Zirk-Sadowski, Ribeaud, & Eisner, 2016; Malti, Ribeaud, & Eisner, 2011). The study population consists of all children that started in the first grade of primary school in the academic year of 2004/5, in Zurich. The target sample was chosen using a stratified random sampling procedure that considered school size and location. This consists of 1675 children from 56 public primary schools. In the present study, we will assess data from the four most recent measurement waves collected to date. This includes data from 1,482 children and composes 88 percent of the original target sample. The median age of children at each wave is approximately 13, 15, 17, and 20 years. The gender ratio is roughly equal (51% male) and the sample is ethnically diverse, with the Swiss majority constituting only 38.4% of the sample. Approximately 11 percent of the data in the study sample is missing and assumed to be missing at random. More detailed information regarding data collection and sample characteristics can be found on the z-proso website (https://www.jacobscenter.uzh.ch/en/research/zproso/aboutus/inst_erheb.html).

Measures. Psychopathology symptoms were measured using an adaption of the self-report version of the Social Behavior Questionnaire (SBQ; Tremblay et al., 1991), which was administered in paper-and-pencil format. The z-proso version includes the addition of several items to enhance the measurement of psychopathology and sustain developmental pertinence as children move through different life periods. Moreover, the original 3-point scale was converted to a 5-point Likert scale (*Never* to *Very often*) and the questionnaire was administered in German. Prior psychometric analyses have ‘generally supported the factorial validity, criterion validity, and reliability of the SBQ items’ (Murray, Obsuth, Eisner, & Ribeaud, 2017, p.3). In the current study, we examined 42 items measured throughout four waves. This includes all SBQ items that have been recorded within the first three waves (age 13-17), which assess the constructs of prosociality, aggression, oppositionality, depression, anxiety, and attention deficit hyperactivity disorder. In the final wave (age 20), an additional 19 items were included to assess the constructs of anger and psychotic experiences. All measured domains refer to the frequency of the behavior in the past year, except for anxiety and depression items which refer to the frequency in the last month. We made two omissions to ensure comparability of item content across waves. First, items that are unique to the final wave (age 20) were excluded from our analyses. Second, four out of five self-report waves were analyzed because the first wave (age 11) measured fewer items (32 out of 42) than the succeeding waves. Lastly, prosociality items were recoded so that higher scores indicate lower levels of prosociality.

Statistical analyses (dataset 1: z-proso)

The following segment describes the specification, estimation, and assessment of the structural equation models used to compare the common cause theory with the dynamic mutualism theory. First, we describe the specification of the measurement models, followed by the specification of the structural models for each theory. Thereafter, we report our choice of estimator and the fit indices used to assess and compare models. To end we describe how we tested for measurement invariance.

Measurement models. The hierarchical structure of psychopathology, which includes a causal p-factor, is modeled in two ways. The first is the *bifactor model*, which structures psychopathology using a general factor (p-factor) and multiple (usually orthogonal) specific factors (Caspi et al., 2014; Lahey et al., 2012). The p-factor explains most of the variance in symptoms of psychopathology, while the residual variance is explained by a set of specific factors. The second is the *higher-order factor model*, which structures psychopathology through a general factor that arises from the correlations among its subordinate dimensions (e.g. internalizing and externalizing factors). In other words, the p-factor in a higher-order factor model *explains* the covariance between its subordinate dimensions (Markon, 2019).

We examined the developmental p-factor that arises from a higher-order factor model. Our rationale is based on theoretical and psychometric concerns about bifactor models (Bonifay, Lane, & Reise, 2017; Greene et al., 2019; Morgan, Hodge, Wells, & Watkins, 2015; Watts et al., 2019) and practical limitations (see Supplement 1, Appendix C).

***Exploratory factor analysis*.** We used an exploratory process to estimate the measurement models. First, we selected four first-order factors (internalizing, externalizing/aggression, pro-sociality, ADHD) based on previous work (Murray et al., 2016, 2017) and conceptual interpretability. Second, to determine the item content of each of the four factors we used exploratory factor analysis (EFA) on each wave and chose the most replicable solution. EFA was conducted using the *psych* package in R (Revelle, 2019) and factors were extracted using minimal residual extraction with oblique rotation. Framing our measurement models as exploratory was deemed most defensible because prior exploratory work on the factor structure of the SBQ, albeit targeting different measurement waves, has already been published using the z-proso data (e.g. Murray et al., 2016).

***Confirmatory factor analysis*.** All analyses were conducted using the *lavaan* package in R (Rosseel, 2012). The four specific factors, estimated through the preceding EFA, were used to specify the confirmatory factor models. First, for the common cause model, confirmatory factor models were estimated at each time point. These models incorporate a five-factor structure composed of four, mutually correlated, first-order factors (internalizing, externalizing/aggression, prosociality, ADHD) and a second-order factor (p-factor). The second-order factor is a summary of the variance shared by the four first-order factors. Second, confirmatory factor models for the dynamic mutualism model were estimated at each time point. These models are composed of four, mutually correlated, first-order factors (internalizing, externalizing/aggression, pro-sociality, ADHD).

Structural models. To compare competing theoretical mechanisms, we specified different latent change score (LCS) models (Kievit et al., 2018; McArdle & Hamagami, 2001; McArdle et al., 2000). The key notion in LCS models is that we can use successive differences between measures to calculate change scores. If we have a basic autoregression (1), where the scores of person *i*, for construct *y*, at time *t* are a function of an autoregressive component and some residual ζ.

$$y_{i,t} = \beta_{\mathrm{t,t-1}} y_{\mathrm{t-1},i} + \zeta_{t,i} \qquad (1)$$

Setting the regression slope ($\beta_{\mathrm{t,t-1}}$) to equal 1 (2), allowed us to conceptualize the residual as the difference between $y_{t,\mathrm{i}}$ and $y_{\mathrm{t-1},i}$ (3), representing the change score $\Delta y_{t,i}$ (4).

$$y_{\mathrm{t}i} = (1) y_{\mathrm{t-1},i} + \zeta_{t,i} \qquad (2)$$

$$\zeta_{t,i} = y_{t,i} - y_{t-1,i} \qquad (3)$$

$$\Delta y_{t,i} = y_{t,i} - y_{t-1,i} \qquad (4)$$

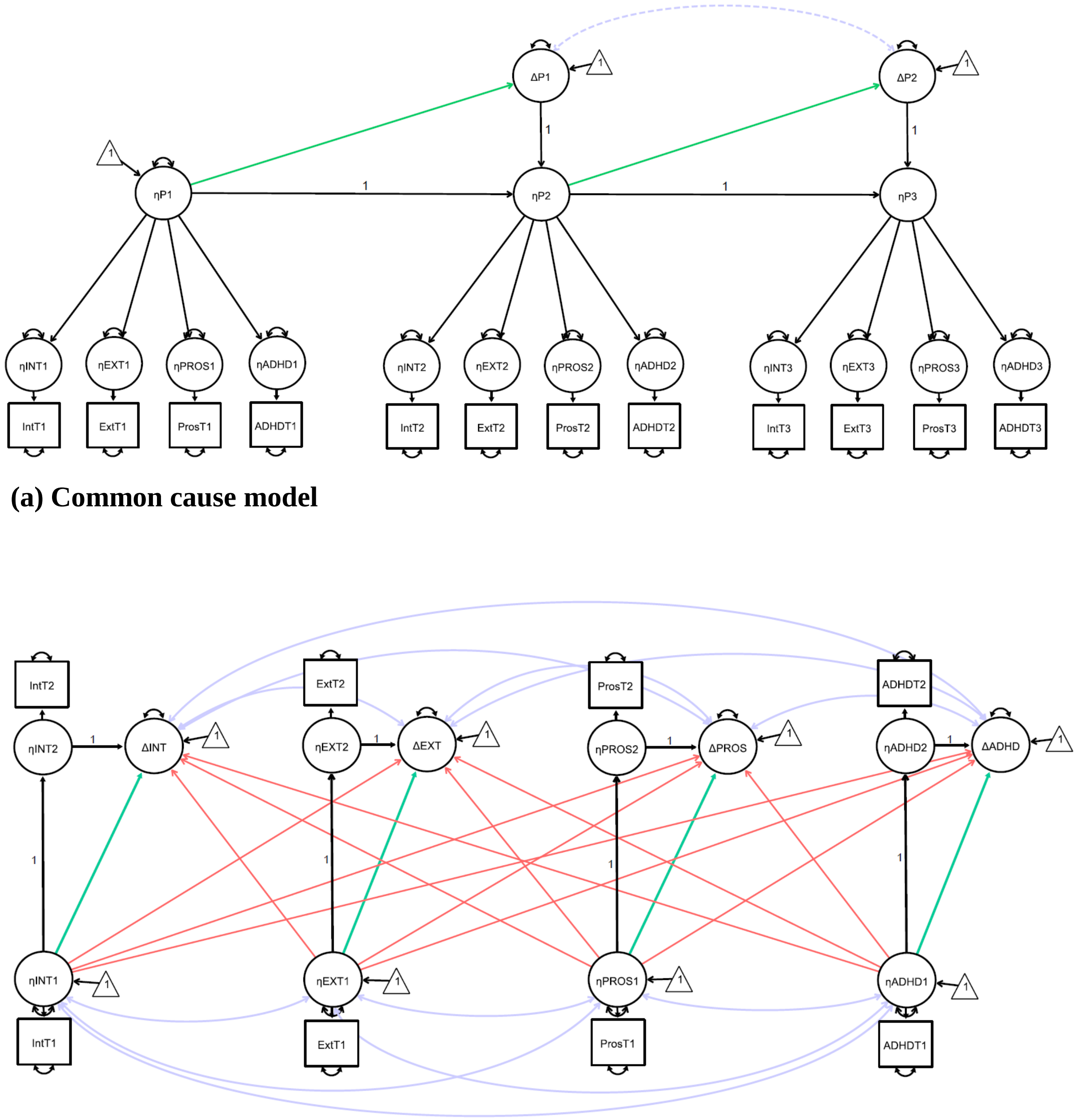
We then defined a latent change score factor $\Delta\eta_{t,i}$, with a factor loading equal to 1, which allowed us to measure change between two time-points. Thereafter, we added a regression parameter to the latent change score factor to estimate how much of the change is due to scores at the previous time point (self-feedback process (β)). Since we are interested in the average change at each time point, we estimated the mean of the difference factor. The variance in the change factor was also estimated, representing how much individuals differ in their change score across time points. Next, we extended the univariate LCS model to a multivariate latent change score model. This gave us the ability to model change scores in multiple domains (McArdle et al., 2002). Change scores in a multivariate LCS model are modeled as the product of two parameters (5): a self-feedback process (β) and a coupling parameter (γ), with the latter capturing the extent to which change in one domain Δy_1 at time t, is dependent on the score of another domain y_2 at the preceding time point t-1.

$$\Delta y_{t,i} = \beta 1y1_{t-1,i} + \gamma 1y2_{t-1,i} \qquad (5)$$

Common cause model. We conceptualized symptom scores as a function of the p-factor score at each time point (Figure 1a). The p-factor influences its own change through a self-feedback parameter (β). We estimated the mean and variance of the latent change score factor at each time point. While the mean and variance of the general factor were estimated at time 1 and equality constrained over time. We allowed residual terms to covary between time points for each observed variable with itself, to allow indicator specific variance (Kievit et al., 2017). Lastly, we imposed measurement invariance over time.

Dynamic mutualism model. In this model, there is no common factor that causes the covariation among the four first-order factors. Instead, the first-order factors influence each other's change (Figure 1b). Higher scores in domain y_1 lead to greater changes in domain y_2, and vice versa. This relationship was modeled for all domains via the use of coupling parameters (*γ*). Moreover, each domain was allowed to influence its own change via a self-feedback process (*β*). All four domains could correlate at time 1. Additionally, to allow indicator specific variance, we allowed residual terms to covary between time points for each observed variable with itself (Kievit et al., 2017). We estimated the mean and variance of the latent change score factors at each time point. The mean and variance of the four latent domains were estimated at time 1 and equality constrained over time. Latent change factors were allowed to correlate within and between time points, portraying the relationship between them after any possible coupling effects. Lastly, we imposed measurement invariance over time.

Model fit and comparison. All models were estimated using the *lavaan* package in R (Rosseel, 2012). Full information maximum likelihood with robust standard errors was used to deal with missingness and nonnormality. We relied on the following indices for the assessment of overall model fit: the root-mean-square error of approximation (RMSEA; acceptable fit: < .08, good fit: < .05), the comparative fit index (CFI; acceptable fit: .95-97, good fit: >.97), and the standardized root-mean-square residual (SRMR; acceptable fit: .05-.10, good fit: < .05; Schermelleh-Engel, Moosbrugger, & Müller, 2003). The chi-square test was reported but not taken into consideration when assessing model fit, because it is oversensitive to sample size and would almost certainly indicate inadequate model fit (Meade, Johnson, & Braddy, 2008). Model comparison was based on the overall model fit indices, the Akaike's information criterion (AIC) and Bayesian information criterion (BIC), and the Akaike weights (Wagenmakers & Farrell, 2004).



(a) Common cause model

(b) Dynamic mutualism model

Figure 1. Illustration of common cause model **(a)** and dynamic mutualism model **(b)**, for z-proso dataset. Circles indicate latent variables, rectangles indicate observed variables, and triangles indicate intercepts. Double-headed arrows indicate covariances (purple) and variances (black). Dashed lines show the parameters that were only included in the exploratory analyses. Single-headed arrows denote regressions. Green single-headed arrows indicate self-feedback parameters (β). Orange single-headed arrows indicate coupling parameters (γ). A "1" shows that the parameter has been constrained to one. The illustration only depicts a limited number of waves and one observed variable per factor, for visual clarity.

Measurement invariance. We tested increasingly strict assumptions for longitudinal measurement invariance. For inferences about changes in factor means over time, strong factorial invariance must hold (Meredith, 1993). To test invariance, we sequentially established equality constraints over time (Widaman, Ferrer, & Conger, 2010). We constrained factor loadings, intercepts, and error terms in that sequence, across time points. Changes in the comparative factor index (ΔCFI) were used to test for measurement invariance (Cheung and Rensvold, 2002). When strong invariance was violated, we loosened intercept constraints for each noninvariant factor separately. Item intercepts were freed sequentially, starting from the item with the largest modification index, until partial invariance was achieved. We compared the results from the fully invariant models with those from the partially invariant models, to test the practical significance of assuming strong invariance (Widaman et al., 2010).

Exploratory analyses. To test the hypothesis that the p-factor can exhaustively explain its development, we estimated a model with freely estimated covariances among change scores over time. We then used a likelihood ratio test to compare it to the common cause model with covariances constrained to zero.

Dataset 2: SHARE

Participants. The data were acquired from the Survey of Health, Ageing and Retirement in Europe (SHARE). SHARE is a European multinational longitudinal project. The study population consists of all persons 33 years or older that have their regular residency at a SHARE country, at the time of sampling. The target sample was acquired using probability sampling with maximum population coverage in each country. This ensured that every person within the population had a probability greater than zero to be selected into the sample. To enable valid inferences for the target population, weighted sample statistics were used to mitigate bias resulting from the unequal probability each individual has to be selected in the target sample. In the current study, the sample consists of 3,969 persons who had at least one measure of interest (i.e. one item on the EURO-D scale), throughout the five waves of interest (waves 1, 2, 4, 5, and 6). Each respective measurement wave is at a two-year distance from its predecessor, apart from the collection at wave 4 which commenced four years after the previous measurement at wave 2. All missing data (0.04%) was assumed to be missing at random. For further information regarding data collection and sample characteristics, we refer the reader to the SHARE website (http://www.share-project.org).

Measures. Data collection was conducted using computer-assisted personal interviewing (CAPI). All questionnaires were translated into the participants' native language using an online translation tool. The present study utilized the EURO-D scale to assess symptoms of major depressive disorder (Prince et al., 1999). This scale includes items covering the symptoms of depression, pessimism, suicidality, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment, and tearfulness. All symptoms assessed for prevalence within the last month. Each symptom is measured using one item on a binary scale, with 0 corresponding to "not present" and 1 to "present". Thus, the total score is measured on an ordinal scale with a maximum score of 12. Several studies have investigated the psychometric properties of the EURO-D scale. The internal consistency of the EURO-D was found to range from 0.58 to 0.80 across countries, as indexed by the standardized alpha. The criterion validity of the scale was deemed sufficient. Associations with different continuous measures of depression ranged from r = 0.70 to 0.93 across sites and the area under the ROC curve indicated strong associations between the EURO-D and other dichotomous measures (0.83-0.93; Prince et al., 1999). In a different study, internal consistency was measured at *a* = 0.75 and test-retest reliability at kappa = 0.60. Moreover, criterion validity was indexed at 0.92 by the area under the ROC curve when predicting DSM-III-R major depression diagnosis by psychiatrists (Larraga et al., 2006). All items were (re)coded so that “1” indicates the presence of symptoms and “0” their absence.

Item parceling. To aid distributional assumptions we allocated the binary EURO-D symptom items to parcels (Bandalos, 2002; Hau & Marsh, 2004; MacCallum, Widaman, Zhang, & Hong, 1999; Matsunaga, 2008; Nunnally, 1978). The two parcels we created, mirror the two factors that have been identified in previous psychometric analyses using the EURO-D scale (Castro-Costa et al., 2008; Guerra, et al., 2015; Prince et al., 1999). The first parcel representing the “affective suffering” construct, included the items of sadness, suicidality, guilt, sleeplessness, irritability, appetite, fatigue, and tearfulness. The second parcel representing the “motivation” construct, included the items of pessimism, interest, concentration, and enjoyment. Both parcels were assumed to be continuous.

Statistical analyses (dataset 2: SHARE)

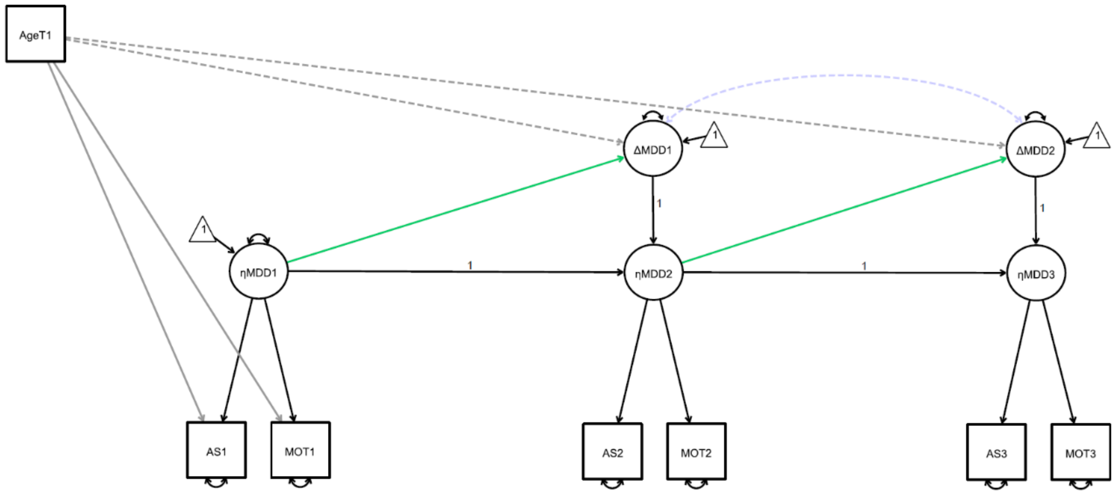
Below we will describe the specification, estimation, and assessment of the models used to compare common cause theory with dynamic mutualism. This section follows the same structure as the z-proso statistical analysis section (measurement models → structural models → estimation and fit). The assessment of measurement invariance was the same for both datasets. Thus, the procedure will not be repeated.

Measurement model for common cause theory. To model major depressive disorder as a reflective latent variable we constructed a one-factor model, which conceptualizes the latent factor of major depression as the direct cause for the covariation among the two parcels. Confirmatory factor models were estimated at each time point.

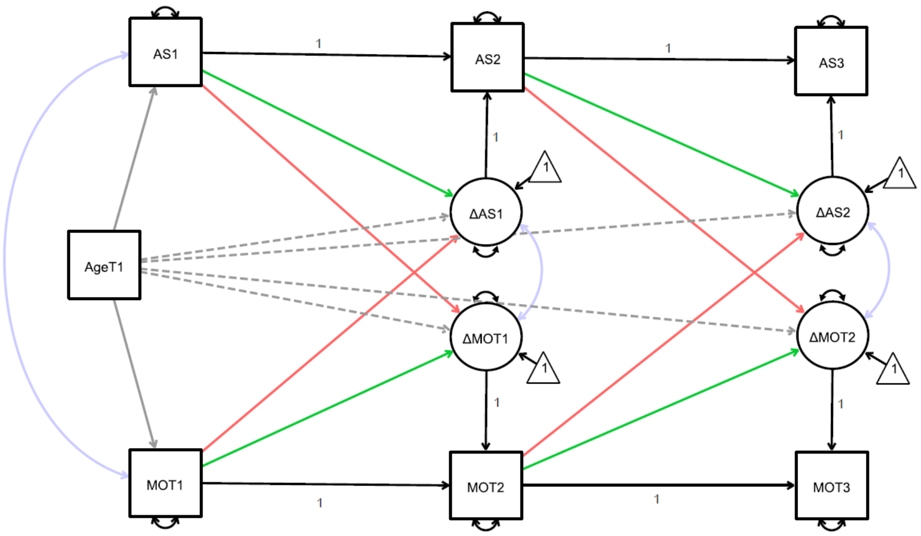
Measurement model for dynamic mutualism theory. A measurement model was not specified for the dynamic mutualism model, because the structural model for dynamic mutualism theory specified the direct interrelations between the two parcel indicators.

Structural model for common cause theory. At each wave, we conceptualized parcel scores as a function of the depression factor score (Figure 2a). The depression factor influenced its own change through a self-feedback parameter (*β*). We estimated the mean and variance of the latent change score factor at each time point. The mean and variance of the depression factor were measured at time 1 and equality constrained over time. To allow indicator specific variance, we allowed residual terms to covary between time points for each indicator with itself (Kievit et al., 2017). Moreover, age at time 1 was included as a covariate to control for the influence of age on the baseline score of the depression factor. Lastly, we imposed measurement invariance over time.

Structural model for dynamic mutualism theory. In this model, no common factor that causes the covariation among the two parcels was specified (Figure 2b). Instead, we conceptualized latent change scores as the function of a coupling process (*γ*) and a self-feedback process (*β*). The two parcels directly influenced each other's change and their own change over time. All observed variables were allowed to correlate at time 1. We estimated the mean and variance of the latent change score factors at each time point. The mean and variance of the two parcels were measured at time 1 and equality constrained over time. We allowed latent change factors to correlate with each other within and between time points, to capture the relationship between them after any possible coupling effects. Lastly, age at time 1 was included as a covariate to control for the influence of age on baseline parcel scores.



(a) Common cause model



(b) Dynamic mutualism model

Figure 2. Illustration of common cause model **(a)** and dynamic mutualism model **(b)** for SHARE dataset. Circles indicate latent variables, rectangles indicate parcels, and triangles indicate intercepts. Double-headed arrows indicate covariances (purple) and variances (black). Dashed lines indicate parameters that were only included in exploratory analyses. Single-headed arrows denote regressions. Green single-headed arrows indicate self-feedback parameters (β). Orange single-headed arrows indicate coupling parameters (γ). Gray single-headed arrows indicate associations with age at time 1 (T1). A “1” shows that the parameter has been constrained to one. The illustration only depicts three out of five waves (both models) and does not depict covariances between change scores across time (dynamic mutualism model only), for visual clarity.

Exploratory statistical analyses. We extended our models in two exploratory analyses. First, to test the hypothesis that the depression factor was the sole influence of its change, we allowed the change scores of the common cause model to correlate over time. This allowed us to capture their relationship after any possible self-feedback effects. Second, to test whether the rate of developmental change in psychopathology was solely explained by the dynamics within the two models, we estimated the direct effect of age on change scores (Kievit et al., 2017). We conducted likelihood ratio tests to assess whether the added parameters significantly improved the models.

Preregistration

All confirmatory analyses were conducted according to our preregistered plan (https://osf.io/a4ywe/?view_only=498f5640c18847bea3ac6a9b0b596821), deviations from the initial plan are reported in Table C1, Appendix C.

Results

Dataset 1: z-proso

Factor specification. Exploratory factor analyses showed that the optimal four-factor solution varied over the four waves. Hence, we chose to theoretically specify the item content of the four factors. Factor loadings for the 42 symptom items are presented in Table 1.

Table 1

Factor loadings for four first-order factors

Items	Abbreviated content	Wave 1	Wave 2	Wave 3	Wave 4
Prosociality					
K5_614	Feel sympathy	0.780	0.788	0.801	0.775
K5_604	Understand feelings	0.528	0.546	0.575	0.505
K5_607	Share with others	0.433	0.424	0.432	0.403
K5_611	Settle dispute	0.512	0.479	0.506	0.455
K5_620	Try to comfort	0.402	0.365	0.376	0.295
K5_617	Try to help injured	0.678	0.633	0.627	0.690
K5_601	Help clear up	0.792	0.772	0.776	0.773
K5_625	Sympathy for feel bad	0.840	0.833	0.843	0.818
K5_623	Listen to other opinion	0.491	0.478	0.509	0.463
K5_626	Sympathy for bullied	0.701	0.687	0.711	0.603
Externalizing					
K5_618	Aggressive if something taken	0.688	0.677	0.685	0.524
K5_603	Aggressive when teased	0.446	0.421	0.360	0.317

K5_605	Bad things behind back	0.444	0.431	0.392	0.370
K5_602	Hit parent	0.676	0.640	0.604	0.590
K5_608	Violent attack	0.756	0.750	0.770	0.742
K5_609	Boss others around	0.507	0.474	0.381	0.299
K5_629	Aggressive when insulted	0.741	0.749	0.749	0.405
K5_610	Lie to parent	0.360	0.299	0.263	0.235
K5_612	Incite other to dislike	0.524	0.476	0.453	0.437
K5_613	Hit, bite, kick others	0.725	0.733	0.726	0.759
K5_630	Humiliate others	0.671	0.653	0.653	0.663
K5_615	Yell at parent	0.350	0.305	0.278	0.249
K5_616	Active exclusion	0.472	0.461	0.452	0.514
K5_633	Told secrets when mad	0.442	0.446	0.431	0.415
K5_606	Scare to force others	0.371	0.350	0.407	0.396
K5_619	Threat others to get something	0.596	0.574	0.580	0.595
K5_621	Throw things at parent	0.402	0.365	0.344	0.380
K5_622	Engage in brawl	0.666	0.644	0.665	0.675
K5_624	Mad not getting something	0.513	0.475	0.446	0.405
Internalizing					
K5_657	Sad without reason	0.657	0.656	0.676	0.708
K5_652	Cried	0.644	0.661	0.662	0.664
K5_653	Fear	0.636	0.655	0.639	0.650
K5_654	Unhappy	0.749	0.773	0.786	0.790
K5_651	Bored	0.743	0.761	0.755	0.763
K5_656	Couldn’t fall asleep	0.481	0.512	0.534	0.547
K5_655	Felt alone	0.272	0.307	0.311	0.310

K5_658	Worried	0.662	0.699	0.723	0.761
K5_659	Self-injury	0.309	0.352	0.376	0.453
ADHD					
K5_627	Restless	0.640	0.687	0.717	0.698
K5_628	Difficulties to concentrate	0.591	0.665	0.658	0.654
K5_631	Inattentive	0.525	0.567	0.598	0.651
K5_632	Hectic and fidgety	0.682	0.730	0.751	0.760

Note: Cross-loadings in confirmatory factor model were constrained to zero.

Measurement invariance. Temporal invariance did not hold for both the common cause model and the dynamic mutualism model. Imposing weak measurement invariance led to a negligible drop in fit for both models (Common cause: ΔCFI = 0.005; Mutualism: ΔCFI = 0.002; Cheung & Rensvold, 2002). Conversely, constraining intercepts to be equal over time led to a substantial drop in model fit for both models (Common cause: ΔCFI = 0.031; Mutualism: ΔCFI = 0.029). The violation of temporal invariance is in line with dynamic mutualism theory but does not necessarily imply mutualism, because many alternative causes of non-invariance exist (e.g. response shift bias; Fokkema, Smits, Kelderman, & Cuijpers, 2013). Next, we compared the results from the fully invariant models with the partially

Table 2

Model comparison fit statistics for z-proso models

Model	χ2	df	RMSEA	CFI	SRMR
Common cause	< 0.001	13835	0.034 [0.034, 0.035]	0.749	0.108
Dynamic mutualism	< 0.001	13716	0.031 [0.031, 0.032]	0.795	0.067
Exploratory common cause*	< 0.001	13832	0.034 [0.034, 0.035]	0.749	0.108

Note: *Common cause model with residual change score covariances.

invariant models, to test the practical significance of assuming strong measurement invariance (Widaman et al., 2010).

Model comparison. The dynamic mutualism model fit best according to all preregistered fit statistics, and both models fit the data well according to all fit indices except the comparative fit index (Table 2). This conclusion is mirrored by the information criteria (AIC and BIC; Figure 3), which were used to control for complexity in terms of the number of freely estimated parameters. The Akaike weights show that given our data the mutualism model is 99.99% more likely to be the better model (Figure 3). The partially invariant models mirrored these conclusions. All fit indices indicated that the partially invariant mutualism model was preferable over the partially invariant common cause model (Table A1, Appendix A). Hence, the violation of temporal invariance was deemed to be of little practical significance (Widaman et al., p.13). Further investigations were carried out using the fully invariant models.

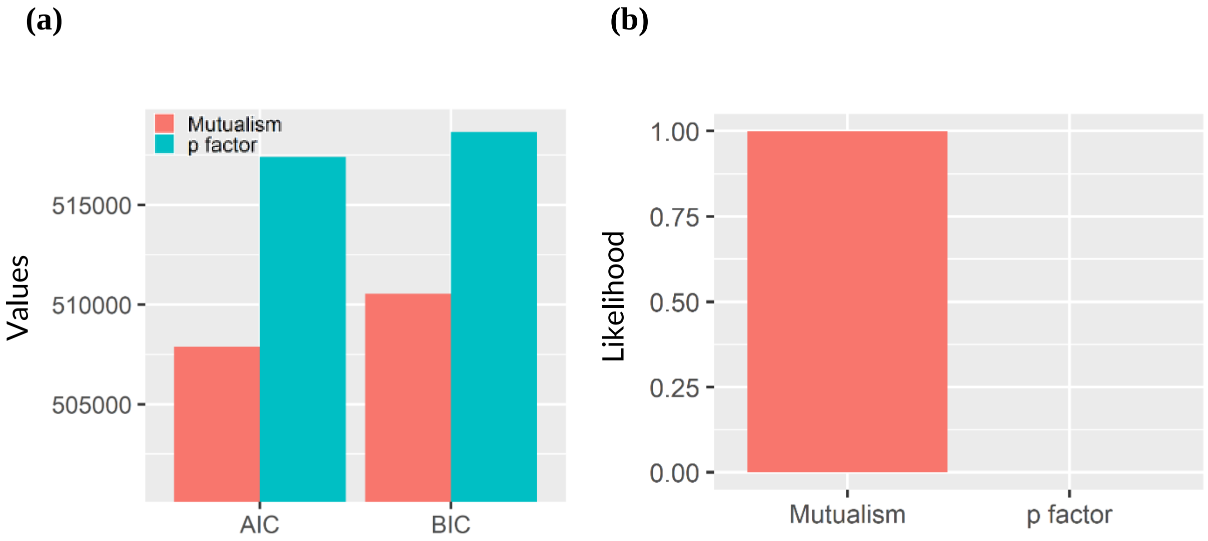


Figure 3. Information criteria AIC and BIC **(a)** and Akaike weights indicating normalized probabilties **(b)**.

Model parameters. We proceeded to closely examine the item content of the best fitting model, the dynamic mutualism model. All regression parameters are reported in Table A2, Appendix A. Psychopathology domains were significantly correlated at baseline, except for ADHD and prosociality that were not significantly related (b = -0.48, p = 0.15). Persons varied substantially in their rate of change, as indexed by the significant variance in change scores in all measured domains (Table A3, Appendix A).

On all measurement occasions, higher scores on all domains predicted a greater decrease in themselves at the succeeding wave. The self-feedback parameters ranged from small to moderate (r = -0.197 – -0.550, r^2 = 3.88 – 30.25%). Most coupling effects were not significant; of the significant ones, there was a mixture of negative and positive coupling parameters. The externalizing dimension was only associated with change in one domain at one time-point. Higher scores in the externalizing domain at time 1 were associated with a greater decrease in internalizing at time 2 (r = -0.125, r^2 = 1.56%). Prosociality influenced the internalizing and externalizing dimensions, though not consistently. Higher prosociality scores at time 3 predicted a greater increase in internalizing at time 4 (r = 0.213, r^2 = 4.54%). Higher scores in prosociality at time 1 predicted a larger increase in externalizing at time 2 (r = 0.092, r^2 = 0.85%). ADHD scores were not significantly associated with change in any dimension. Lastly, higher scores on the internalizing dimension at time 1, were associated with a greater increase in ADHD at time 2 (r = 0.146, r^2 = 2.13%) and a greater decrease in prosociality at time 2 (r = -0.067, r^2 = 0.45%).

Exploratory analyses. In line with the hypothesis that a causal p-factor is the sole influence of its own change, an exploratory common cause model that allowed change scores to be related after self-feedback effects did not fit better than the preregistered common cause model ($\Delta\chi^2(3)$ = 6.61, p = 0.09).

Dataset 2: SHARE

Measurement invariance. Temporal invariance was not violated for the common cause model. Imposing weak measurement invariance did not change model fit (ΔCFI = 0.000), and neither did imposing strong measurement invariance (ΔCFI = 0.000; Cheung & Resvold, 2002).

Table 3

Model comparison fit statistics SHARE data

Model	χ2	df	RMSEA	CFI	SRMR
Common cause	< 0.001	28	0.039 [0.034, 0.045]	0.980	0.026
Dynamic mutualism	< 0.001	8	0.066 [0.057, 0.076]	0.983	0.031

Model comparison. We fitted both models to the data to determine which best explains the development of the positive manifold within major depressive disorder, over the measured waves. None of the preregistered fit indices showed preferential support for either model, except for the root-mean-square error of approximation (RMSEA) which supported the common cause model (Table 3). The information criteria (AIC and BIC; Figure 4), which were used to control for complexity in terms of the number of freely estimated parameters, portrayed a similarly, mixed picture. The Akaike weights show that given our data the dynamic mutualism model is 99.83% more likely to be the best model, while the Schwarz weights support the opposite conclusion (the common cause model is 99.99% more likely to be the best model; Figure 4). In comparison to the AIC, the BIC criterion places a higher penalty on complexity that scales with sample size. This may explain the antithetical conclusions.

Model parameters. Since we were unable to establish the superiority of either model, we closely examined the parameters of both models.

Common cause model. All regression parameters are presented in Table B1, Appendix B. We found considerable interindividual variability in the rate of change of the common factor (Table B2, Appendix B). Age at baseline predicted common factor scores at baseline, as evidenced by the substantial drop in fit after we fixed the effect of age at baseline on the common factor at baseline, to 0 ($\Delta\chi 2(1) = 24.43$, $p < .001$). Over most measurement waves higher scores on the depression factor were associated with a greater decrease in depression at the next time point. The negative self-feedback parameters were of moderate size (r = -0.227—0.362, r^2 = 5.29—12.96%). The relationship between the depression factor at time 3 and change at time 4 was the exception to this pattern, with higher scores on the depression factor predicting a greater increase in depression scores. The positive self-feedback parameter was in the small range (r = .131, r^2 = 1.71%).

Dynamic mutualism model. All regression parameters are reported in Table B3, Appendix B. We found evidence for individual differences in the rate of change within both domains (Table B2, Appendix B). Age at baseline predicted parcel scores at baseline ($\Delta\chi 2(2) = 19.23$, $p < .001$) and residual change score covariances substantially impacted model fit ($\Delta\chi 2(26) = 1690.5$, $p < .001$). This indicates the existence of unmeasured influences on change within the two domains and/or temporal mismatch between measurement and the natural pace of change (Hofman et al., in review).

Self-feedback effects. The affective suffering domain only substantially influenced its own change at the second measurement wave. Higher affective suffering scores at time 1 were associated with a greater decrease in affective suffering at time 2 (r = -0.562, r^2 = 31.58%). The same was evident in the motivation domain. Higher scores on motivation at time 1 were associated with a greater decrease in the motivation domain at time 2 (r = -0.640, r^2 = 40.96%).

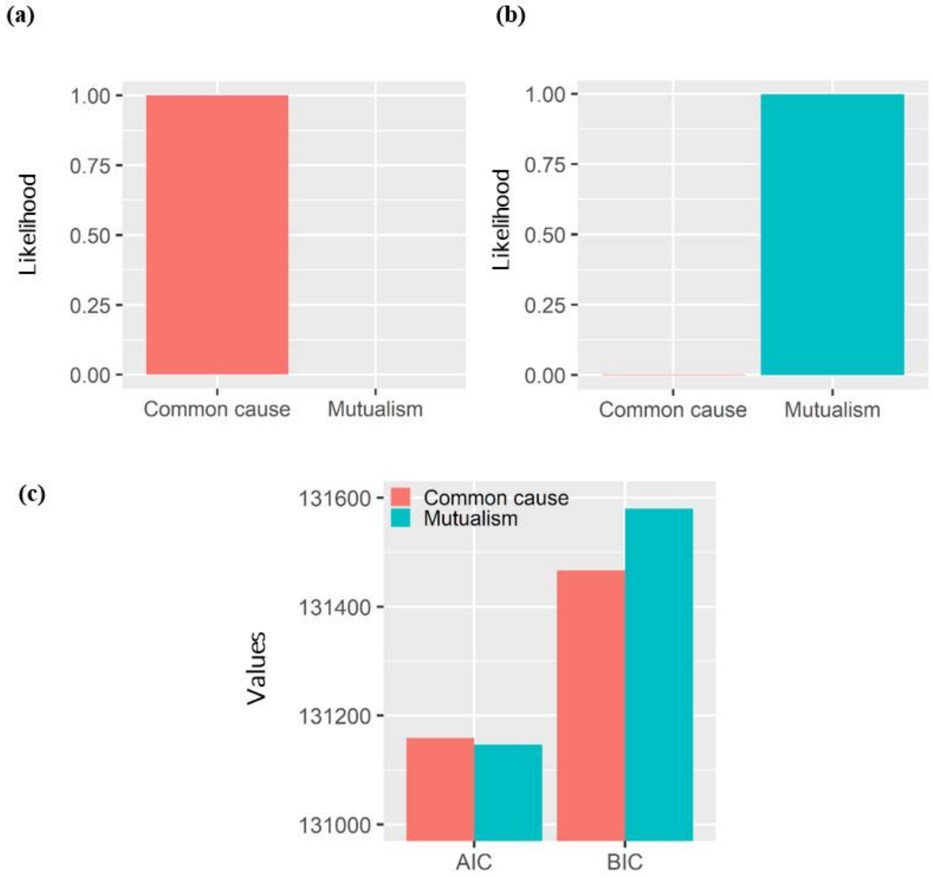


Figure 4. Normalized probabilities indicated by Schwarz weights **(a)** and Akaike weights **(b)**, and AIC and BIC information criteria **(c)** for each model.

Coupling effects. Most coupling effects indicated that scores on the motivation domain were not significantly associated with change in the affective suffering domain. The exception was higher scores on motivation at time 1, which predicted a greater increase in affective suffering at time 2 (r = 0.063, r^2 = 0.40%). The affective suffering domain was significantly associated with change in motivation in half of the measurement occasions. Higher scores in affective suffering at time 1 and time 3, were associated with a greater increase in the motivation domain at time 2 (r = 0.049, r^2 = 0.24%) and time 4 (r = 0.101, r^2 =1.02%), respectively.

Exploratory analyses. Fit statistics for the exploratory models are presented in Table B4, Appendix B. First, our results did not support the hypothesis that the common cause factor is solely responsible for its own change, since allowing change scores to covary over time led to substantial improvement in model fit ($\Delta\chi^2(6) = 65.98$, $p < 0.001$). Second, both models failed to capture all age-related dynamics, since allowing age to directly affect change scores led to a significant improvement in model fit (Common cause: $\Delta\chi^2(4) = 110.71$, $p < 0.001$; Mutualism: $\Delta\chi^2(8) = 202.93$, $p < 0.001$).

Discussion

To gain insight into the ontogeny and ontology of psychiatric dimensions, we directly compared the ability of dynamic mutualism theory and common cause theory to explain the development of the positive manifold at multiple levels of the hierarchical taxonomy—the p-factor from early to late adolescence and major depressive disorder from middle adulthood to old age.

Interpretation of major findings

At the level of the p-factor, the dynamic mutualism model provided a better account of development than the common cause model and a hybrid (common cause) model that allowed unmeasured factors to influence the development of the p-factor. However, all three models fit the data poorly according to the confirmatory factor index. At the level of major depressive disorder, it was not clear which model provided a better explanation of development, as different fit indices supported a different model. Both models fit the data well, but neither model could exhaustively explain all age-related developmental dynamics and exploratory analyses supported the possibility that unmeasured factors (e.g. life events) affected the rate of change.

Our findings strongly question the assumption that symptoms cohere due to a singular causal mechanism. But do not invalidate the predictive utility of constructs such as the p-factor, which has been correlated with numerous risk factors and deleterious outcomes (Caspi et al., 2014; Lahey et al., 2012, 2015; Martel et al., 2017; Snyder et al., 2017; Waldman, Poore, van Hulle, Rathouz, & Lahey, 2016). In many cases, coupling between symptoms and/or syndromes may explain the development of psychopathology. But mutualistic coupling may not be sufficient either.

Multiple external factors may compose key drivers of developmental change. Our models did not include such factors, which might explain the unexpected finding that multiple coupling parameters were negative. This is especially likely given the large amount of time between measurements. That is, within one year many unmeasured factors may have influenced change in a given domain, which may have biased its statistical associations with other domains. This limitation is an artifact of currently available longitudinal data. But also expose the vagueness of dynamic mutualism theory, which does not explicate the time it takes for developmental processes to unfold, nor does it state exactly how internal and external factors might interact. Future studies should use measurements with greater temporal density to empirically inform theory about the temporal pace of hypothesized developmental mechanisms and seek to specify functional associations between symptoms and relevant external factors.

A hybrid approach to understanding psychopathology

In line with previous conclusions, neither theory could fully explain the development of psychopathology (Greene & Eaton, 2017; McElroy et al., 2018; Murray et al., 2016; Snyder et al., 2017). Hence, it may be better to start looking at what percentage of variance in the developmental dynamics of psychopathology each theory can explain. Hybrid models may provide promising multicausal explanations for the development of psychopathology (Fried & Cramer, 2017). For instance, general vulnerability to psychopathology may reinforce causal interactions between symptoms by lowering their activation threshold. This can increase the probability that symptoms are caused by environmental events and other symptoms. In turn, symptoms may exacerbate this general vulnerability (e.g. effect of sleep on stress response system; Koss & Gunnar, 2018; Ly, McGrath, & Gouin, 2015).

Alternatively, specific types of latent causes may lead to specific disorders and interactions among the presenting symptoms and environmental factors may lead to comorbidity. This would be consistent with an interpretation of the p-factor as an amalgamation of distinct causes (Krueger et al., 2018; Watts et al., 2019) that may cohere due to the causal interrelations among their outcomes. Both scenarios, and multiple others (see Fried & Cramer, 2017 for more examples) where latent pathophysiology co-exists, and possibly interacts, with mechanisms at the level of manifest psychopathology, may explain the development of the positive manifold. Moreover, different mechanisms and their interactions may lead to the development of different disorders and the same disorder may be explained by several mechanisms (Borsboom, Cramer, & Kalis, 2019). Thus, a multicausal framework may be most suited to understand mental illness (Kendler, 2019).

Implications for current practices

Our understanding of the mechanisms that drive the development of psychopathology is still in its infancy, and the lack of evidence supporting a monocausal explanation should inspire skepticism about practices predicated on the assumption that disorders reflect a common cause. First, a monocausal interpretation of disorders neglects heterogeneity in symptoms and their causes. Neglect of symptom heterogeneity is commonly seen in diagnostic schemes that use symptoms as interchangeable indicators of a disorder (e.g. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; American Psychiatric Association, 2013) and in the widespread use of diverse measurement instruments to measure the same disorder (Fried, 2017). Consequently, we may be lumping persons with meaningfully diverse pathologies at every stage of clinical research and practice (Fried, 2015). Second, it is uncertain what (reflective) latent variables that summarize the shared variance between items represent if this shared variance is not solely due to a common cause. Presently, reflective latent variables are used to represent all dimensions of psychopathology.

The shared variance that composes the variance of some dimensions may represent the concert action of several mechanisms, which calls to question the explanatory utility of research that aims to explain the variance in plausibly confounded reflective latent variables. For explanatory research, reflective latent variables should be constructed at the right level of resolution to capture a discrete source of shared variance. Hence, research initiatives that wish to describe, explain, and modify psychopathology by leveraging a hierarchical factor model (e.g. HiTOP; Caspi et al., 2014; Kotov et al., 2017; Lahey et al., 2012, 2017), could largely benefit from establishing the ontology of quantitative dimensions to ascertain which dimensions are suitable for explanatory work and which are better suited for prediction.

Limitations and future recommendations

The results of the current study should be viewed considering several limitations. First, in the dynamic mutualism models, we specified causal interrelations between all the constituent parts of the model. However, it is more likely that a psychopathology network includes both direct and indirect associations. A fruitful avenue for future research could be to directly compare mutualism models that include different causal paths to improve our understanding of causal associations and improve the specificity of theory.

Second, we used item parcels to normalize data, but it is questionable whether the parcels reflect substantive constructs. Although the content of the parcels was based on constructs that have been recurrently identified in prior studies using the EURO-D scale (Castro-Costa et al., 2008; Guerra et al., 2016; Prince et al., 1999), the constructs are quantitative creations and the theoretical coherence between the items is questionable. Additionally, parcels attribute equal weight to all items, which can lead to bias proportional to the difference between the item coefficient in the true model and the weight specified by the sum score (Bollen & Bauldry, 2011). We would ideally use multi-item continuous measures of individual symptoms to create a dynamic mutualism model that specifies interrelations among symptoms directly.

Third, for the means of testing the ontology of the p-factor, we assumed that broad transdiagnostic constructs are adequately described using reflective latent variables. However, it may be that these dimensions are, at least partly, the product of mutualistic coupling among symptoms. Future work would benefit from assessing the mechanisms that explain the coherence among symptoms starting from the lowest levels of abstraction. This would prevent the misuse of reflective latent variables to summarize variance when a different model is appropriate (Rhemtulla, van Bork, & Borsboom, 2019).

Fourth, the findings of the present study are limited to the populations assessed in the analyzed samples and to the measured dimensions of psychopathology. Further research focusing on different developmental periods could assess homogeneity in psychological processes throughout development.

Generally, future work should aim to cumulatively build a theoretical foundation for psychiatry (Kendler, 2009). Recent technological innovations provide us the privilege to formalize theory via computational models and accelerate cumulative theory construction (Robinaugh et al., in review). Computational models promote the precise specification of functional associations among elements and constitute a valuable continuation to necessarily imprecise verbal theories (Smaldino, 2017). Precision ensures that hypotheses are falsifiable, and theories are amenable to cumulative modification that can edge us closer to understanding the complex phenomenon of mental illness.

Conclusion

We echo recent calls for the abandonment of a monocausal framework to explain mental illness (Borsboom et al., 2019; Kendler, 2019) and reinforce this argument through the first direct comparison between dynamic mutualism theory and common cause theory.

Neither of our models provides the best possible explanation for the development of psychopathology, but rather our findings illustrate the abstractness of currently dominant theories and expose gaps in our understanding. We hope that our models will provide the necessary stimulation to start a conversation around formalized theory, to build a solid base for the future of psychiatry.

References

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (5th Edition). In *American Journal of Psychiatry*. https://doi.org/10.1176/appi.books.9780890425596.744053

Aristodemou, M. E., & Fried, E. I. (in press). Common Factors and Interpretation of the P Factor of Psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry*.

Averdijk, M., Zirk-Sadowski, J., Ribeaud, D., & Eisner, M. (2016). Long-term effects of two childhood psychosocial interventions on adolescent delinquency, substance use, and antisocial behavior: a cluster randomized controlled trial. *Journal of Experimental Criminology*, *12*(1), 21–47. https://doi.org/10.1007/s11292-015-9249-4

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*(1), 78-102. https://doi.org/10.1207/S15328007SEM0901_5

Bollen, K. A., & Bauldry, S. (2011). Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science*, *5*(1), 184-186. https://doi.org/10.1177/2167702616657069

Borsboom, D. (2008). Psychometric Perspectives on Diagnostic Systems. *Journal of Clinical Psychology*, *64*(9), 1089–1108. https://doi.org/10.1002/jclp.20503

Borsboom, D., & Cramer, A. O. J. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*, 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*(e2), 1-63. https://doi.org/10.1017/S0140525X17002266

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The Theoretical Status of Latent Variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., … Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119-137. https://doi.org/10.1177/2167702613497473

Caspi, A., & Moffitt, T. E. (2018). All for One and One for All: Mental Disorders in One Dimension. *American Journal of Psychiatry*. *175*(9), 831-844. https://doi.org/10.1176/appi.ajp.2018.17121383

Castro-Costa, E., Dewey, M., Stewart, R., Benerjee, S., Huppert, F., Mondonca-Lima, C., … Prince, M. (2008). Ascertaining late-life depressive symptoms in Europe: an evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE project. *International Journal of Methods in Psychiatric Research*, *17*(1), 12–29. https://doi.org/10.1002/mpr

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2-3), 137–150. https://doi.org/10.1017/S0140525X09991567

Borsboom, D. (2017). A network theory of mental disorders. *World psychiatry*, *16*(1), 5-13.

https://doi.org/10.1002/wps.20375

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617–634. https://doi.org/10.1037/met0000167

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97-115. *https*://doi.org/10.1007/BF00485230

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. Psychological Assessment, 25, 520–531. http://dx.doi.org/10.1037/a0031669

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6*, 309. https://doi.org/10.3389/fpsyg.2015.00309

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I., & Cramer, A. O. J. (2017). Moving Forward: Challenges and Directions for Psychopathological Network Theory and Methodology. *Perspectives on Psychological Science*, *12*(6), 999–1020. https://doi.org/10.1177/1745691617705892

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017a). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1–10. https://doi.org/10.1007/s00127-016-1319-z

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom,

D. (2017b). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology, 52*(1), 1-10. https://doi.org/10.1007/s00127-016-1319-z

Greene, A. L., & Eaton, N. R. (2017). The temporal stability of the bifactor model of comorbidity: An examination of moderated continuity pathways. *Comprehensive Psychiatry, 72,* 74-82. https://doi.org/10.1016/j.comppsych.2016.09.010

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Waldman, I., … Patrick, C. J. (2019). Are Fit Indices to Test Psychopathology Structure Biased? A Simulation Study. *Journal of Abnormal Psychology*. https://doi.org/10.1037/abn0000434

Guerra, M., Ferri, C., Llibre, J., Prina, A. M., & Prince, M. (2015). Psychometric properties of EURO-D, a geriatric depression scale: A cross-cultural validation study. *BMC Psychiatry, 15*(1), 12. https://doi.org/10.1186/s12888-015-0390-4

Hau, K. T., & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology, 57*, 327-251. https://doi.org/10.1111/j.2044-8317.2004.tb00142.x

Hofman, A. D., Kievit, R., Stevenson, C., Molenaar, D., Visser, I., & van der Maas, H. (2018, February 28). The dynamics of the development of mathematics skills: A comparison of theories of developing intelligence. https://doi.org/10.31219/osf.io/xa2ft

Horgan, T. (1993). Nonreductive Materialism and the Explanatory Autonomy of Psychology. In *Naturalism: A Critical Appraisal*.

Insel, B. T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science, 348*(6234), 499–500. https://doi.org/10.1126/science.aab2358

Jonas, K. G., & Markon, K. E. (2016). A descriptivist approach to trait conceptualization and inference. *Psychological Review, 123*(1), 90–96. https://doi.org/10.1037/a0039542

Kendler, K. S., (2009). An historical framework for psychiatric nosology. *Psychological*

Medicine, *39*(12), 1935–1941. https://doi.org/10.1017/S0033291709005753

Kendler, K. S., (2019). From Many to One to Many-The Search for Causes of Psychiatric Illness. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2019.1200

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A. L., de Mooij, S. M. M., Moutoussis, M., … Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99–117. https://doi.org/10.1016/j.dcn.2017.11.007

Kievit, R. A., Hofman, A. D., & Nation, K. (2019). Mutualistic Coupling Between Vocabulary and Reasoning in Young Children: A Replication and Extension of the Study by Kievit et al. (2017). *Psychological Science*, *30*(8), 1245-1252. https://doi.org/10.1177/0956797619841265

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., & Dolan, R. J. (2017). Mutualistic Coupling Between Vocabulary and Reasoning Supports Cognitive Development During Late Adolescence and Early Adulthood. *Psychological Science*, *28*(10), 1419-1431. https://doi.org/10.1177/0956797617710785

Koss, K. J., & Gunnar, M. R. (2018). Annual Research Review: Early adversity, the hypothalamic–pituitary–adrenocortical axis, and child psychopathology. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *59*(4), 327–346. https://doi.org/10.1111/jcpp.12784

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., … Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*(3), 282-293. https://doi.org/10.1002/wps.20566

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, *6*, 1–11. https://doi.org/10.1038/srep34175

Larraga, L., Saz, P., Dewey, M. E., Marcos, G., & Lobo, A. (2006). Validation of the Spanish version of the EURO-D scale: an instrument for detecting depression in older people. *International Journal of Geriatric Psychiatry*, *21*(12), 1199-1205. https://doi.org/10.1002/gps.1642

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is There a general factor of prevalent psychopathology during adulthood? *Journal of Abnormal Psychology*, *121*(4), 971. https://doi.org/10.1037/a0028355

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin*, *143*(2), 142-186. https://doi.org/10.1037/bul0000069

Lahey, B. B., Rathouz, P. J., Keenan, K., Stepp, S. D., Loeber, R., & Hipwell, A. E. (2015). Criterion validity of the general factor of psychopathology in a prospective study of girls. *Journal of Child Psychology and Psychiatry*, *56*(4), 415-422, https://doi.org/10.1111/jcpp.12300

Ly, J., McGrath, J. J., & Gouin, J. P. (2015). Poor sleep as a pathophysiological pathway underlying the association between stressful experiences and the diurnal cortisol profile among children and adolescents. *Psychoneuroendocrinology*, *57*, 51-60. https://doi.org/10.1016/j.psyneuen.2015.03.006

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84. https://doi.org/10.1037/1082-989X.4.1.84

Malti, T., Ribeaud, D., & Eisner, M. P. (2011). The effectiveness of two universal preventive interventions in reducing children’s externalizing behavior: A cluster randomized controlled trial. *Journal of Clinical Child and Adolescent Psychology*, *40*(5), 677–692. https://doi.org/10.1080/15374416.2011.597084

Markon, K. E. (2019). Bifactor and Hierarchical Models: Specification, Inference, and Interpretation. *Annual Review of Clinical Psychology*, *15*(1), 1–19. https://doi.org/10.1146/annurev-clinpsy-050718-095522

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports*, *5*, 1–7. https://doi.org/10.1038/srep09050

Martel, M. M., Pan, P. M., Hoffmann, M. S., Gadelha, A., do Rosário, M. C., Mari, J. J., … Salum, G. A. (2017). A general psychopathology factor (P Factor) in children: Structural model analysis and external validation through familial risk and child global executive function. *Journal of Abnormal Psychology*, *126*(1), 137. https://doi.org/10.1037/abn0000205

Matsunaga, M. (2008). Item Parceling in Structural Equation Modeling: A Primer. *Communication Methods and Measures*. *2*(4), 260-293. https://doi.org/10.1080/19312450802458935

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, *38*(1), 115–142. https://doi.org/10.1037/0012-1649.38.1.115

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In *New methods for the analysis of change.* https://doi.org/10.1037/10409-005

McArdle, J. J., Hamagami, F., Meredith, W., & Bradway, K. P. (2000). Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learning and Individual Differences*, *12*(1), 53–79. https://doi.org/10.1016/S1041-6080(00)00036-4

McElroy, E., Belsky, J., Carragher, N., Fearon, P., & Patalay, P. (2018). Developmental stability of general and specific factors of psychopathology from early childhood to adolescence: dynamic mutualism or p-differentiation? *Journal of Child Psychology and Psychiatry*, *59*(6), 667-675. https://doi.org/10.1111/jcpp.12849

Mcnally, R. J. (2016). Can network analysis transform psychopathology?*. *Behaviour Research and Therapy*, *86*, 95–104. https://doi.org/10.1016/j.brat.2016.06.006

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and Sensitivity of Alternative Fit Indices in Tests of Measurement Invariance. *Journal of Applied Psychology*, *93*(3), 568. https://doi.org/10.1037/0021-9010.93.3.568

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, *58*(4), 525-543.

Morgan, G., Hodge, K., Wells, K., & Watkins, M. (2015). Are Fit Indices Biased in Favor of Bi-Factor Models in Cognitive Ability Research?: A Comparison of Fit in Correlated Factors, Higher-Order, and Bi-Factor Models via Monte Carlo Simulations. *Journal of Intelligence*, *3*(1), 2–20. https://doi.org/10.3390/jintelligence3010002

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The Development of the General Factor of Psychopathology ‘p Factor’ Through Childhood and Adolescence. *Journal of Abnormal Child Psychology*, *44*(8), 1573–1586. https://doi.org/10.1007/s10802-016-0132-1

Murray, A. L., Obsuth, I., Eisner, M., & Ribeaud, D. (2017). Evaluating Longitudinal Invariance in Dimensions of Mental Health Across Adolescence: An Analysis of the Social Behavior Questionnaire. *Assessment*. https://doi.org/10.1177/1073191117721741

Nunnally, J. C. (1978). Psychometric Theory: Second Edition. *Applied Psychological Measurement*.

Prince, M. J., Beekman, A. T. F., Deeg, D. J. H., Fuhrer, R., Kivela, S.-L., Lawlor, B. A., … Copeland, J. R. M. (1999). Depression symptoms in late life assessed using the EURO–D scale. *British Journal of Psychiatry*, *174*(4), 339–345. https://doi.org/10.1192/bjp.174.4.339

Putnam, H. (1967). *Psychological Predicates*. University of Pittsburgh Press.

Pylyshyn, Z. W. (1984). Computation and Cognition. *Dialogue*, *23*(292), 811–814. https://doi.org/10.1017/S0012217300049854

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Revelle, W. R. (2019). psych: Procedures for psychological, psychometric, and personality research, R package 1.8. 4.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2019). Worse than Measurement Error: Consequences of Inappropriate Latent Variable Measurement Models. *Psychological Methods*, *8*(5), 55. http://dx.doi.org/10.1037/met0000220

Robinaugh, D., Haslbeck, J. M. B., Waldorp, L., Kossakowski, J. J., Fried, E. I., Millner, A., … Borsboom, D. (2019, May 29). Advancing the Network Theory of Mental Disorders: A Computational Model of Panic Disorder. https://doi.org/10.31234/osf.io/km37w

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. R package version 0.5-15. *Journal of Statistical Software*, *48*(2), 1–36.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *MPR-Online*, *8*(2), 23–74.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In *Computational Social Psychology*. https://doi.org/10.4324/9781315173726

Snyder, H. R., Young, J. F., & Hankin, B. L. (2017). Strong Homotypic Continuity in Common Psychopathology-, Internalizing-, and Externalizing-Specific Factors Over Time in

Adolescents. *Clinical Psychological Science*, *5*(1), 98-110. https://doi.org/10.1177/2167702616651076

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivée, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *Journal of Abnormal Child Psychology*, *19*(3), 285–300. https://doi.org/10.1007/BF00911232

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory and Psychology*, *27*(6), 759–773. https://doi.org/10.1177/0959354317737185

Van Bork, R., Wijsen, L., & Rhemtulla, M. (2017). Toward a causal interpretation of the common factor model. *Disputatio*, *9*(47), 581–601. https://doi.org/10.1515/disp-2017-0019

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychological Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.1021/ef300604q

Waldman, I. D., Poore, H. E., van Hulle, C., Rathouz, P. J., & Lahey, B. B. (2016). External validity of a hierarchical dimensional model of child and adolescent psychopathology: Tests using confirmatory factor analyses and multivariate behavior genetic analyses. *Journal of Abnormal Psychology*, *125*(8), 1053. https://doi.org/10.1037/abn0000183

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier Tests of the Validity of the Bifactor Model of Psychopathology. *Clinical Psychological Science*, *5*(1), 3–13. https://doi.org/10.1177/2167702616673363

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Appendix A: Supplementary material for z-proso dataset

Table A1

Partially invariant model comparison (z-proso)

Exploratory models					
Model	χ2	df	RMSEA	CFI	SRMR
Common cause	< 0.001	13820	0.038 [0.037, 0.038]	0.704	0.113
Dynamic mutualism	< 0.001	13704	0.035 [0.034, 0.035]	0.751	0.088

Table A2

Regression parameters for dynamic mutualism model (z-proso)

Regressions	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper	Std.all
dINT1 ~							
INTlv_T1	-0.385	0.044	-8.822	0.000	-0.470	-0.299	-0.400
EXTlv_T1	-0.146	0.050	-2.916	0.004	-0.244	-0.048	-0.125
PSlv_T1	-0.032	0.031	-1.044	0.297	-0.093	0.028	-0.037
ADHDlv_T1	0.094	0.049	1.927	0.054	-0.002	0.189	0.092
dINT2 ~							
INTlv_T2	-0.168	0.077	-2.185	0.029	-0.318	-0.017	-0.197
EXTlv_T2	0.011	0.087	0.122	0.903	-0.160	0.181	0.009
PSlv_T2	-0.093	0.064	-1.446	0.148	-0.219	0.033	-0.103
ADHDlv_T2	0.020	0.082	0.243	0.808	-0.142	0.182	0.021
dINT3 ~							
INTlv_T3	-0.188	0.109	-1.727	0.084	-0.401	0.025	-0.214
EXTlv_T3	-0.136	0.143	-0.955	0.340	-0.416	0.143	-0.089
PSlv_T3	0.205	0.101	2.039	0.041	0.008	0.402	0.213
ADHDlv_T3	0.079	0.111	0.714	0.475	-0.139	0.298	0.083
dEXT1 ~							
EXTlv_T1	-0.464	0.042	11.099	0.000	-0.546	-0.382	-0.539
INTlv_T1	-0.002	0.027	-0.079	0.937	-0.054	0.050	-0.003
PSlv_T1	0.059	0.022	2.668	0.008	0.016	0.102	0.092
ADHDlv_T1	-0.036	0.032	-1.129	0.259	-0.100	0.027	-0.049
dEXT2 ~							

EXTlv_T2	-0.273	0.071	-3.871	0.000	-0.411	-0.135	-0.341
INTlv_T2	0.023	0.047	0.483	0.629	-0.070	0.115	0.042
PSlv_T2	0.049	0.044	1.107	0.268	-0.038	0.135	0.084
ADHDlv_T2	-0.008	0.047	-0.161	0.872	-0.100	0.085	-0.013
dEXT3 ~							
EXTlv_T3	-0.433	0.072	-5.992	0.000	-0.575	-0.292	-0.550
INTlv_T3	0.021	0.045	0.481	0.631	-0.066	0.109	0.047
PSlv_T3	0.064	0.050	1.296	0.195	-0.033	0.162	0.130
ADHDlv_T3	-0.038	0.047	-0.799	0.424	-0.130	0.055	-0.077
dPS1 ~							
PSlv_T1	-0.454	0.029	15.702	0.000	-0.511	-0.397	-0.530
INTlv_T1	-0.064	0.032	-1.995	0.046	-0.127	-0.001	-0.067
EXTlv_T1	0.041	0.047	0.862	0.389	-0.052	0.134	0.035
ADHDlv_T1	0.023	0.038	0.621	0.534	-0.051	0.097	0.023
dPS2 ~							
PSlv_T2	-0.208	0.065	-3.198	0.001	-0.336	-0.081	-0.248
INTlv_T2	-0.030	0.068	-0.447	0.655	-0.163	0.103	-0.038
EXTlv_T2	-0.030	0.085	-0.348	0.728	-0.197	0.138	-0.026
ADHDlv_T2	0.024	0.072	0.329	0.742	-0.118	0.165	0.027
dPS3 ~							
PSlv_T3	-0.446	0.085	-5.247	0.000	-0.613	-0.280	-0.525
INTlv_T3	-0.162	0.083	-1.953	0.051	-0.324	0.001	-0.209
EXTlv_T3	0.094	0.120	0.786	0.432	-0.141	0.329	0.070
ADHDlv_T3	0.004	0.092	0.043	0.966	-0.176	0.184	0.005
dADHD1 ~							

ADHDlv_T1	-0.392	0.047	-8.294	0.000	-0.485	-0.299	-0.422
INTlv_T1	0.128	0.038	3.380	0.001	0.054	0.203	0.146
EXTlv_T1	-0.095	0.049	-1.956	0.051	-0.190	0.000	-0.089
PSlv_T1	0.002	0.032	0.071	0.943	-0.060	0.065	0.003
dADHD2 ~							
ADHDlv_T2	-0.204	0.087	-2.336	0.019	-0.375	-0.033	-0.233
INTlv_T2	0.033	0.077	0.430	0.667	-0.117	0.183	0.041
EXTlv_T2	0.119	0.096	1.240	0.215	-0.069	0.306	0.102
PSlv_T2	0.026	0.072	0.366	0.715	-0.114	0.167	0.031
dADHD3 ~							
ADHDlv_T3	-0.281	0.104	-2.696	0.007	-0.485	-0.077	-0.340
INTlv_T3	0.156	0.099	1.578	0.114	-0.038	0.350	0.206
EXTlv_T3	-0.061	0.128	-0.477	0.634	-0.313	0.190	-0.046
PSlv_T3	-0.001	0.100	-0.006	0.995	-0.196	0.195	-0.001

Table A3

Change score variances for dynamic mutualism model (z-proso)

Change scores	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper	Std.all
dINT1	0.416	0.030	13.805	0.000	0.357	0.475	0.839
dINT2	0.415	0.033	12.561	0.000	0.350	0.480	0.891
dINT3	0.464	0.040	11.461	0.000	0.385	0.543	0.880
dEXT1	0.192	0.017	11.226	0.000	0.158	0.225	0.714
dEXT2	0.145	0.018	8.093	0.000	0.110	0.180	0.744
dEXT3	0.086	0.012	7.160	0.000	0.062	0.109	0.610
dPS1	0.360	0.023	15.392	0.000	0.314	0.406	0.742
dPS2	0.344	0.025	13.496	0.000	0.294	0.394	0.848
dPS3	0.272	0.021	12.826	0.000	0.231	0.314	0.660
dADHD1	0.337	0.026	13.039	0.000	0.286	0.388	0.818
dADHD2	0.366	0.031	11.710	0.000	0.305	0.428	0.886
dADHD3	0.311	0.032	9.829	0.000	0.249	0.373	0.787

Appendix B: Supplementary material for SHARE dataset

Table B1

Self-feedback parameters for common cause model (SHARE)

Regressions	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper	Std.all
dpft2 ~							
pft1	-0.231	0.086	-2.694	0.007	-0.400	-0.063	-0.362
dpft3 ~							
pft2	-0.153	0.094	-1.625	0.104	-0.339	0.032	-0.227
dpft4 ~							
pft3	0.109	0.088	1.243	0.214	0.063	0.281	0.131
dpft5 ~							
pft4	-0.234	0.041	-5.670	0.000	-0.315	-0.153	-0.335

Table B2

Change score variances for common cause mode and dynamic mutualism model

Common cause model

Change score	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper	Std.all
dpft2	0.143	0.045	3.170	0.002	0.055	0.232	0.869
dpft3	0.165	0.036	4.514	0.000	0.093	0.236	0.948
dpft4	0.299	0.056	5.326	0.000	0.189	0.410	0.983
dpft5	0.364	0.057	6.350	0.000	0.251	0.476	0.888

Dynamic mutualism model

Change scores	Estimate	Std. Err	z-value	P(>|z|)	ci.lower	ci.upper	Std.all
dsad1	1.623	0.044	36.856	0.000	1.537	1.709	0.696
dsad2	2.317	0.114	20.302	0.000	2.093	2.541	0.948
dsad3	2.401	0.137	17.566	0.000	2.133	2.669	1.014
dsad4	2.221	0.135	16.417	0.000	1.955	2.486	0.944
dsleep1	0.319	0.014	22.639	0.000	0.291	0.346	0.602
dsleep2	0.610	0.062	9.848	0.000	0.489	0.731	1.004
dsleep3	0.582	0.055	10.665	0.000	0.475	0.689	0.866
dsleep4	0.712	0.099	7.166	0.000	0.518	0.907	1.125

Table B3

Self-feedback and coupling parameters for dynamic mutualism model (SHARE)

Regression	Estimate	SE	z-value	p-value	CI_{lower}	CI_{upper}	Std.lv	Std.all
daff1 ~								
affect1	-0.596	0.017	-35.614	0.000	-0.629	-0.564	-0.391	-0.562
mot1	0.154	0.042	3.705	0.000	0.073	0.236	0.101	0.063
daff2 ~								
affect2	-0.058	0.053	-1.101	0.271	-0.161	0.045	-0.037	-0.052
mot2	-0.056	0.192	-0.295	0.768	-0.432	0.319	-0.036	-0.021
daff3 ~								
affect3	0.016	0.058	0.280	0.780	-0.098	0.131	0.011	0.016
mot3	-0.037	0.194	-0.189	0.850	-0.418	0.344	-0.024	-0.016
daff4 ~								
affect4	-0.053	0.061	-0.875	0.382	-0.173	0.066	-0.035	-0.055
mot4	-0.232	0.225	-1.033	0.302	-0.672	0.208	-0.151	-0.105
dmot1 ~								
mot1	-0.749	0.022	-33.452	0.000	-0.793	-0.705	-1.029	-0.640
affect1	0.025	0.007	3.662	0.000	0.012	0.038	0.034	0.049
dmot2 ~								
mot2	0.001	0.116	0.011	0.991	-0.227	0.230	0.002	0.001
affect2	0.013	0.027	0.479	0.632	-0.039	0.065	0.016	0.023
dmot3 ~								
mot3	-0.171	0.106	-1.617	0.106	-0.378	0.036	-0.208	-0.139
affect3	0.055	0.029	1.934	0.053	-0.001	0.111	0.067	0.101
dmot4 ~								

mot4	0.139	0.146	0.953	0.341	-0.147	0.425	0.175	0.122
affect4	-0.042	0.036	-1.171	0.242	-0.113	0.028	-0.053	-0.084

Table B4

Model comparison fit statistics SHARE data

Exploratory models					
Model	χ2	df	RMSEA	CFI	SRMR
Common cause[1]	< 0.001	22	0.033 [0.028, 0.040]	0.988	0.021
Common cause[2]	<0.001	18	0.027 [0.021, 0.033]	0.992	0.015
Dynamic mutualism[3]	< 0.001	6	0.019 [0.004, 0.034]	0.999	0.006

Note: 1 = Common model with residual change score covariance over time. 2= Common cause model with residual change score covariance over time and direct age effects on change. 3 = Dynamic mutualism model with age directly influencing change scores and coupling parameters constrained to equality to over-identify model.

Appendix C – Preregistration and supplementary information

Supplement 1. We examined the developmental p-factor that arises from a higher-order factor model and not the p-factor that arises from bifactor model. Our rationale is twofold. First, several concerns have been voiced regarding the use of bifactor models. These include concerns about the theoretical interpretability of specific factors (Bonifay, Lane, & Reise, 2017) and a propensity to overfit data, which urges caution when interpreting model fit indices (Murray & Johnson, 2013; Morgan et al., 2015). Second, it is not possible to estimate a dynamic mutualism model that can be used for comparison with a bifactor model. In a bifactor model the p-factor directly explains a large component of the shared variance between all symptoms. A competing dynamic mutualism model needs to explain this shared variance through the causal interrelations between all symptoms. This needs more regression parameters than are possible to estimate with the degrees of freedom we have available. In a higher-order factor model, the p-factor explains the shared variance between specific factors. A competing dynamic mutualism model needs to explain this shared variance through the causal interrelations between specific factors (not symptoms). This is possible. Directly comparing a dynamic mutualism model that specifies causal interrelations between specific factors with a bifactor common cause model, would be like comparing two different explanations for two different phenomena. As the bifactor model would explain the correlations between symptoms via a causal p-factor, while the dynamic mutualism model would explain the correlations between specific factors via the causal interrelations between them. Hence, we will directly compare the higher-order factor model with the (only possible) dynamic mutualism model, because they provide different explanations for the same phenomenon.

Table C1

Deviations from preregistration

Dataset	Planned to	Deviation	Rationale
SHARE	Use four symptom items as indicators.	Used item parcels as indicators.	To normalize data for maximum likelihood estimation.
	Use data from all persons in sample.	Used only the persons that had at least 1 measurement on the EURO-D scale across all waves.	We wanted to assess developmental changes over time and most cases only had data on one wave. This would have resulted in more than 70 percent missing data. Hence, only analyzing a subset of this data was deemed most defensible.
z-proso	We reported that the sample size for the 4 waves used was 1532.	The actual sample size was 1482.	—
	Use a four-factor EFA to identify item content of factors.	We specified the item content of the four factors based on theory.	The four-factor EFA showed a divergent optimal structure throughout the four waves. To specify a

homogeneous factor model that is most generalizable, we relied on the conceptual congruence of the items instead.