Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood

Rogier A. Kievit*,[a,b], Ulman Lindenberger[a,c,d], Ian Goodyer,[e,f]. Peter B. Jones,f[e,f], Peter Fonagy[h], Ed T. Bullmore[e,f,i,j], the NSPN Consortium[k] & Raymond J. Dolan[a,l]

Abstract

One of the most replicable findings in psychology is that individual differences in cognitive abilities are universally positively correlated. The developmental origin of this positive manifold is crucial to its understanding. In a large (N = 785) longitudinal (566 both waves, mean interval 1.48 years) cohort of adolescents and young adults (age range 14-25) we examined developmental changes in two core cognitive domains, fluid reasoning and vocabulary. We use bivariate latent change score models to compare three leading accounts of cognitive development: g factor theory, investment theory and mutualism. We show that a mutualism model, which proposes that basic cognitive abilities directly and positively interact during development, provides the best account of age related changes. We find that individuals with higher scores in vocabulary show greater gains on matrix reasoning and vice versa. These dynamic coupling pathways are not predicted by other accounts, and provide a novel mechanistic window into cognitive development.

Keywords

Cognitive development, mutualism, vocabulary, fluid reasoning, longitudinal modelling

*corresponding author: rogier.kievit@mrc-cbu.cam.ac.uk
a Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London WC1B 5EH, UK
b MRC Cognition and Brain Sciences Unit, Cambridge, 15 Chaucer Rd, Cambridge CB2 7EF
c Max Planck Institute for Human Development, Berlin, Germany
d Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany
e Department of Psychiatry, University of Cambridge, UK
f Cambridgeshire and Peterborough National Health Service Foundation Trust, Cambridge, CB21 5EF, United Kingdom
i ImmunoPsychiatry, GlaxoSmithKline Research and Development, Stevenage SG1 2NY, United Kingdom
j Medical Research Council/Wellcome Trust Behavioural and Clinical Neuroscience Institute, University of Cambridge
k See supplementary information for a full list of contributors
l Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, UK

Introduction

Among the most reproducible findings in the literature on general cognitive ability is the positive manifold, which captures a pervasive positive correlation between distinct cognitive abilities (Deary, 2012; Spearman, 1927). This positive manifold allows the extraction of a single factor, often called '*g*' (for 'general intelligence') that summarizes a considerable proportion of shared variance across abilities within a single index. *g* has remarkable predictive ability for a variety of life outcomes including health, income, mortality, mental health, educational attainment and socio-economic status (Aichele, Rabbitt, & Ghisletta, 2015; Gottfredson & Deary, 2004; Penke et al., 2012). Although the presence of a positive manifold and the *g* factor as a statistical entity is beyond question, its ontology and ontogeny are more contentious.

One challenge arises out of the fact that the *g* factor is almost always based on cross-sectional data, and this can obscure developmental patterns that are not adequately accounted for in many influential theories. For instance, van der Maas (2006) has noted that one of the most influential modern works on the *g* factor (Jensen, 1998) fails to address the issue of development. This is despite observations of a relatively rapid increase in higher cognitive abilities such as reasoning, knowledge and mental speed during adolescence, a trajectory mirrored by an increasingly steep decline in old age (Schaie, 1994). Moreover, very different hypotheses regarding the underlying nature of *g* can give rise to mathematically equivalent statistical patterns in cross-sectional data (van der Maas et al., 2006).

Here we ask whether a lack of attention to development has limited a comprehensive understanding both of the *g* factor, as well as its development over time. Lifespan changes in cognitive abilities provide a crucial inroad into the ontological status of *g,* enabling one to ask whether there truly is an underlying general factor that plays a causal role during cognitive development or, alternatively, whether a positive manifold arises out of a more complex developmental process. We consider three possible accounts of cognitive development: *g* factor theory, investment theory, and mutualism, each of which provides distinct causal accounts of the

2

emergence of cognitive abilities during development. Crucially, developments in structural equation modelling (McArdle, 2009) allow each of these accounts to be translated into psychometric models, enabling us to compare them directly using the same longitudinal dataset.

g-factor theory (Gignac, 2014; Jensen, 1998) posits a single underlying general ability that is used in various domains. For example, Gottfredson (2002) states '*g* is a highly general capability for processing complex information of any type'. A simple developmental perspective based on the *g* factor proposes that during (early) development, an individual's general ability increases over time, which in turn affects (increases) scores across a variety of abilities that depend directly or indirectly on *g*. A defining feature of this account is an absence of direct causal links *between* cognitive abilities. Evidence for this *g* factor account comes from Gignac (Gignac, 2014, 2016), who suggested that the *g* factor structure is relatively stable between ages of 2.5 and 10 (Gignac, 2014) and that the residual structure of lower cognitive factors (Gignac, 2016) is more compatible with *g* factor theory than competing accounts such as mutualism. Contrary evidence comes from McArdle (McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002), who showed that developmental trajectories across abilities vary considerably not just in their developmental order, but also in their shape, concluding '….a single *g* factor yields an overly simplistic view of growth and change over age'.

A second influential account is Cattell's investment theory (Cattell, 1971). This is based on a division of cognitive abilities into crystallized (knowledge-based) and fluid abilities (flexible skills not dependent on acquired knowledge or skills). The theory makes a central developmental claim, namely that fluid abilities are *invested* in order to *acquire* crystallized abilities. Recent work (Weiland, Barata, & Yoshikawa, 2014) suggests that executive function scores at the beginning of a preschool year predict improvements in vocabulary performance at the end of the year but not vice versa. A large cross-sectional sample studied (Valentin Kvist & Gustafsson, 2008) found that the factor structure of general and fluid abilities within, and across, groups was compatible with investment theory. However, findings are ambiguous (Valentin Kvist & Gustafsson, 2008), with

others finding no effect (Christensen, Batterham, & Mackinnon, 2013), only the reverse pattern (Fuhs & Day, 2011), or an effect only in one cohort (Ferrer & McArdle, 2004).

A third developmental account is the *mutualism* model. This model suggests causal interactions between multiple basic cognitive abilities across developmental time, such that cognitive abilities mutually facilitate growth over time. Under this assumption, developmental change will yield a positive manifold even from a starting point of completely uncorrelated cognitive abilities. The model predicts positive coupling between multiple basic cognitive abilities across (early) development. The strongest empirical evidence for mutualistic processes comes from a lifespan cohort study that observed coupling effects between speed and block design, memory and vocabulary and digit span and block design and forward digit span (McArdle, Hamagami, Meredith, & Bradway, 2000). Similarly, Schmidt and Crano (1974) used cross-lagged panel analysis to test investment theory, but found evidence that both crystallised and fluid abilities are related over time, concluding investment theory cannot account for this pattern. Contrary evidence from a cross-sectional sample suggests that an increase in g factor strength expected in the strongest version of mutualism is not unambiguously observed (Gignac, 2014).
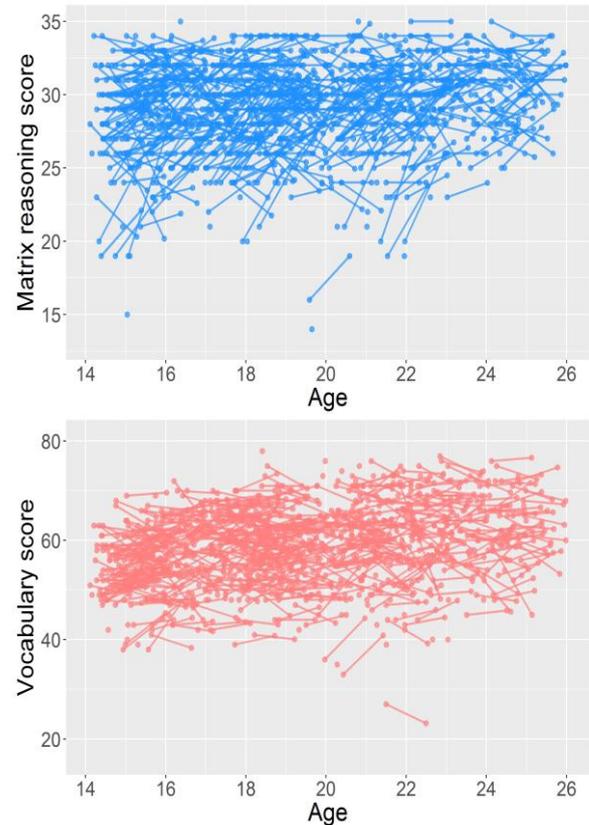
Method

Several challenges preclude strong inferences regarding the best model of cognitive development. First, the studies discussed above sample from various points in the lifespan, which may be governed by different developmental mechanisms. Second, several reports have relied on statistical techniques such as cross-lagged panel models (Schmidt & Crano, 1974) not ideally suited to study change (McArdle, 2009). Third, other studies have relied on cross-sectional cohorts which limit the range of inferences that can be made (e.g. (Gignac, 2014; Valentin Kvist & Gustafsson, 2008). Most importantly, although several studies test specific theories, or compared a subset (Ferrer & McArdle, 2004; Ghisletta & Lindenberger, 2003; McArdle et al., 2002, 2000) to the best of our knowledge no study has yet directly compared these three prominent accounts of development. Our aim in this study was to fill this gap by exploiting innovations in structural equation modelling (McArdle, 2009)

4

that are uniquely suited to directly compare these three developmental accounts. To do this we

exploit data from a large (*N* = 785, age 14-25) developmental cohort measured on two domain-

representative (crystallized and fluid), standardized tests (WASI matrix reasoning and WASI

vocabulary). Raw scores are shown in Figure 1,

descriptive statistics given in Table 1.

Using a latent change score modelling

framework we modelled the three theoretical

accounts of change in cognitive abilities as three

different versions of the (B)LCS, shown in Figure 2A-

2C, with key parameters in red. First, for the *g*-factor

model (Fig 2A) we conceptualize model scores on

vocabulary and matrix reasoning as a function of an

underlying *g* score for each time point. To ensure

comparability of factor change scores across T1 and

T2 for the *g* factor model, we tested for longitudinal

measurement invariance (Widaman, Ferrer, &

Conger, 2010). We found that imposing weak

Fig. 1. Raw scores on WASI-II matrix reasoning (top) and Vocabulary scores (bottom), lines connecting the same individual across testing sessions.

invariance across time points (factor loadings) led to negligible decrease in model fit (ΔCFI = 0.004)

(Cheung & Rensvold, 2002). Imposing strong invariance (equality of both factor loadings and

thresholds) also led to acceptable decrease in model fit (ΔCFI = 0.014). This suggests longitudinal

measurement invariance is tenable, and we can interpret changes in factor scores accordingly.

Second, investment theory implies that scores in fluid abilities (here indexed by matrix

reasoning) should positively influence the degree of change in crystallized abilities (vocabulary), such

that individuals with greater fluid ability will, on average, improve more in crystallized abilities than

peers with lower matrix reasoning scores on T1. This process is modelled by a single coupling

parameter from matrix scores at T1 on the vocabulary change factor at T2 (Figure 2B in red). Finally,

the mutualism model (2C) predicts bivariate coupling between both cognitive abilities, such that

higher starting points in vocabulary would lead to larger gains in matrix reasoning and vice versa. In

all models we add age as a covariate to account for differences in baseline scores, but do not include

age anywhere else in the model (i.e. we hypothesize that the dynamics of change are fully captured

by the change dynamics proposed by each theory).

Task	N	Mean	Min	Max	SD	Skewness	Excess Kurtosis
Matrix reasoning T1	785	29.04	14	35	3.18	-0.87	1.33
Matrix reasoning T2	565	29.63	17	35	2.88	-0.84	0.85
Vocabulary T1	785	58.57	27	78	7.85	-0.26	0.05
Vocabulary T2	566	58.99	20	77	7.74	-0.56	1.17

Table 1. Raw scores and key moments for WASI-II matrix reasoning and Vocabulary scores across two waves.

Sample

As part of the Cambridge-UCL Neuroscience in Psychiatry (NSPN) cohort, we enrolled 785

participants (402 female; mean age: 19.05, range 14.1 to 24.99 years), shown to be a sufficient

sample size to fit moderately complex SEM's with adequate power (e.g. Rast & Hofer, 2014; Wolf,

Harrington, Clark, & Miller, 2013), with 566 participants tested a second time, on average 1.48 years

later (range: 0.65 – 2.62 years). Those who returned for a second wave did not differ from those who

did not on time 1 vocabulary scores ($t(366.5)=.27$, $BF_{01} = 10.86$), time 1 matrix reasoning scores

($t(361.57)= 0.54$, $BF_{01} = 9.64$) or sex ($\chi^2 (1)= 0.7254$, $BF_{01} = 8.11$), current or past treatments for

emotional, behavioural or mental health problems (current: $t(271.6)= -1.47$, $BF_{01} = 2.08$), past:

current: $t(348.04)= -0.95$, $BF_{01} = 6.8$) or parental education (mothers school leaving age: $t(156.51)= -0.85$, $BF_{01} = 4.93$; fathers school departure age: $t(159.4)= -0.49$, $BF_{01} = 4.93$). Those with complete

data were slightly younger at the time of first testing ($M = 18.81$) than those who did not (M =

19.67), $t(415.62)= -3.77$, $BF_{10} = 64.7$, and had slightly higher Barratt Impulsivity Scores (BIS-11;

Stanford et al., 2009)) ($M = 60.52$ vs. 63.30, $t(389.9)= -3.58$, $BF_{10} = 46.77$). Implementing either

Complete Case Analysis or excluding individuals with above cut-off (72) BIS scores did not

meaningfully affect model parameters or model comparisons reported below. The role of age is discussed in more detail below. Full ethical approval was provided prior to the study (reference: 12/EE/0250). Data and code needed to reproduce analyses are available online[1].

Participants were tested on the Wechsler Abbreviated Scale of Intelligence-II (Wechsler, 2011), consisting of two subtests: Matrix reasoning and Vocabulary knowledge. Matrix reasoning measures fluid and visual intelligence by means of a series of incomplete visual matrices, requiring participants to pick one out of five options that best completes the matrix. The vocabulary subtest measures the breadth of word knowledge and verbal concepts by asking participants to verbally define words and describe words or concepts orally presented by the examiner. Both subtests have excellent inter-rater reliability (.98 and .95), split half reliability (.90-.92) and concurrent validity (.71-.92) with comparable tests such as the WISC-IV and WAIS-IV (Wechsler, 2011, key points summarized in McCrimmon & Smith, 2013, p. 339). The highly similar reliabilities of the measures ensure comparable interpretation of cross-domain effects. Prior to further modeling scores on both tests at time 2 were rescaled to equate intervals across individuals using the score difference and the inter-test interval, as proposed by Ferrer & McArdle (2004).

Modelling framework

To tease apart candidate mechanisms of development we fit a series of Latent Change Score (LCS) models (Kievit et al., 2017; McArdle & Hamagami, 2001; McArdle et al., 2000). These models conceptualize differences between successive measurements as a latent change factor. Crucially, this allows us to directly model within subject changes as a function of structural parameters, making these models more suitable for our purposes than latent growth curve models or cross-lagged regressions (McArdle, 2009). The basic equation of the latent change score model specifies the score of person *i* on test *Y* at time *t* as a sum of score at time t-1 and a change, or difference, to the score at t-1 as follows:

---

[1]https://osf.io/rvcph/

$$(1) \ Y_{i,t} = \beta_{t,t-1} Y_{i,t-1} + \Delta_{i,t}$$

A key step in the LCS specification is to set the regression weight $\beta_{t,t-1}$ to 1 (McArdle & Hamagami, 2001), allowing us to rewrite the change scores as follows:

$$(2) \ \Delta_{i,t} = Y_{i,t} - Y_{i,t-1}$$

These change scores are then modelled as perfect indicators of a latent change score factor. In case of only one observed variable, or indicator, per construct being available, the latent change factor is construed as the difference between these indicators over time. In the absence of coupling the intercept of the simple LCS gives approximately identical results as a paired t-test when testing for differences across two measurement occasions, but it allows a modelling of two additional parameters of considerable theoretical importance: the variance in change scores (i.e. do individuals change homogeneously, or not, over time) and the covariance between scores at t-1 and change scores. We can extend the basic univariate LCS to a Bivariate Latent Change Score model with abilities Y1 and Y2 (McArdle et al., 2002) by modelling the change scores on two domains Y1 and Y2 (here vocabulary and matrix reasoning) as the function of two processes: a self-feedback process (beta) and a coupling process (gamma), as follows:

$$(3) \ \Delta Y1_{i,t} = \beta 1 * Y1_{i,t-1} + \gamma 12 * Y2_{i,t-1}$$
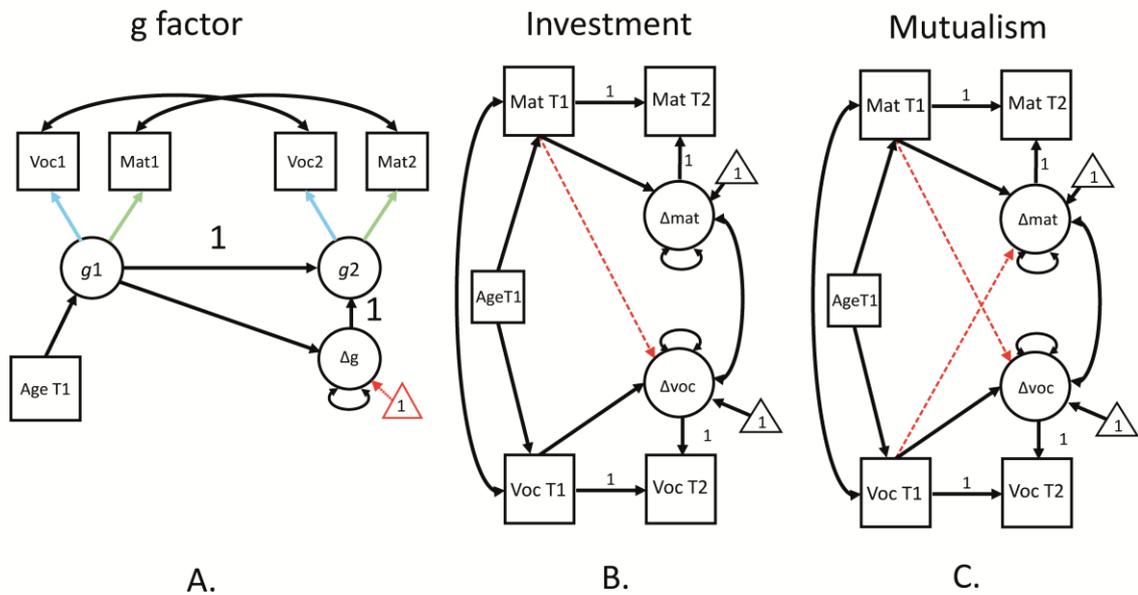
Fig. 2A-2C. Three developmental accounts as Latent Change Score models. Defining features include the presence of a general factor (A) and the presence of univariate (C) or bivariate (D) coupling parameters. Voc: WASI vocabulary test (capturing crystallized abilities), mat: WASI matrix reasoning test (capturing fluid abilities). Key parameters shown dashed.

The self-feedback parameter beta is thought to reflect a combination of effects including regression to the mean and a dampening effect induced by an end horizon for rapid development (i.e. individuals reaching their performance ceiling). The coupling parameter gamma is of special importance for several developmental accounts. It captures whether the *change* in Y1 is determined by the t-1 scores in Y2 (and vice versa for Y2), and thus captures the degree to which cross-domain change is affected by the level of a cognitive ability in some other domain, above and beyond the self-feedback parameter. These gammas are conceptually similar to the M matrix in the mutualism model (van der Maas et al., 2006)

Model fit and comparison

Models were estimated in Lavaan version 5.22 (Rosseel, 2012) using Full Information Maximum Likelihood with robust standard errors to account for missingness and non-normality. No observations were excluded. We assess overall model fit via the chi-square test, the RMSEA (acceptable fit<0.08, good fit <0.05), the CFI (acceptable fit .95-.97, good fit >0.97), the SRMR (acceptable fit .05-.10, good fit <0.05) (Schermelleh-Engel, Moosbrugger, & Müller, 2003). We

compare model fit using chi-square test (for nested models), information criteria (AIC and BIC) and

Akaike Weights, which express the relative likelihood of a set of models given the data

(Wagenmakers & Farrell, 2004).

Results

Before fitting the models shown in Figure 2, we fit two univariate latent change score models to

Vocabulary and Matrix Reasoning scores in order to quantify change within each domain. Both

models fit the data well: Matrix Reasoning: $\chi^2(1)= 2.59$, $p = 0.108$, RMSEA = 0.045 [0.00 0.114], CFI =

0.996, SRMR = 0.013, Yuan-Bentler scaling factor= .917; Vocabulary: $\chi^2(1) = 0.033$,

$p = 0.85$, RMSEA = 0.00 [0.00 0.049], CFI = 1.0, SRMR = 0.001, scaling factor= 1.052). Both models

show evidence for change over time (unstandardized change score intercepts[2]: Matrix reasoning =

10.171, $SE = 0.769$, $z = 13.22$, Vocabulary: 9.0, $SE = 1.22$, $z = 7.36$), evidence for negative feedback

(higher scores at T1 are associated with less improvement, compatible with regression to the mean

and/or developmental ceiling effects; Matrix: -0.331, $SE = 0.026$, $z = -12.82$; Vocabulary: -0.147, $SE =$

0.21, $z = -7.15$), and significant evidence for individual differences in change scores (variance of

Matrix reasoning change scores: 2.85, $SE = 0.23$, $z = 12.73$, Vocabulary change scores = 11.67, $SE =$

1.11, $z = 10.47$).

 Having shown, as expected, a growth in scores in both domains, we next fit all three models

(*g* factor, investment and mutualism) to determine which provides the best account of longitudinal

development in these two cognitive domains across two measurement occasions. We use model

comparison to compare the three models in three ways: overall model fit (cf. Schermelleh-engel et

al., 2003), by comparing Information Criteria (AIC and BIC) and by computing Akaike weights

(Wagenmakers & Farrell, 2004), which use differences in AIC to quantify the relative likelihood of a

model being the best among the set of competitors, given the data. Next, we fit Model B

---

[2]Note that these intercept parameters can only be interpreted in the context of the full latent change score model that includes age as covariate and the self-feedback pathway. The model implied score increases in the absence of coupling are 0.370 (Vocabulary) and 0.559 (Matrix Reasoning), raw scores are shown in Table 1.

(investment) and Model C (Mutualism), differing only in the presence of a vocabulary to reasoning coupling parameter.

In Table 2 we report the fit statistics for each of the three competing models. This comparison suggests that the mutualism model fits the data best, showing excellent model fit on all indices. The two alternative models show comparable model fit among each other, and any difference is marginal according to conventional guidelines. As the mutualism model is also the most complex model, we plot information criteria (AIC and BIC) for each of the three models to explicitly weigh parsimony, as shown in Figure 3A. This comparison shows a superior fit on both indices for the mutualism model. Finally, we compute Akaike weights. This measure is based on the difference in AIC, and allows us to quantify how likely a model is to be best among a set of competitors given the data (cf. 36). These are shown in Figure 3B, illustrating that the mutualism model has by far the highest normalized probability (>99.99%) of being the best model given our data. Compared to the other two models, the mutualism model is $1.98*10^7$ times more likely to be best model. As the investment model is nested within the Mutualism model, we can compare the two with a chi-square test, which again shows the mutualism model outperforming the investment model ($\chi^2\Delta$ = 22.75, *df* Δ =1, *p* < 0.001).

|	χ^2	*df*	RMSEA	CFI	SRMR
g factor	30.078	3	0.107 [0.077 0.140]	0.979	0.029
Investment	26.28	3	0.099 [0.068 0.135]	0.982	0.039
Mutualism	0.132	2	0.000 [0.000 0.020]	1	0.001

Table 2. Fit statistics for each of the three models.

Having established the superior fit of the mutualism model, we next investigated its estimated parameters in more detail. The full model with all estimated parameters is shown in Figure 4. Supplementary Table 1 contains all parameters estimates and 95% confidence intervals. As expected, reasoning and vocabulary at T1 are positively correlated, and age at first testing predicts scores on both tasks at T1. In addition to significant intercepts (i.e. increasing scores), fixing the variance of change scores to 0 led to a substantial drop in model fit ($\chi^2\Delta$ = 83.16, *df*Δ =1, *p* < 0.001

11

for matrix reasoning, $\chi^2 \Delta$ = 13.44, $df \Delta$ = 1, $p < 0.001$ for vocabulary), suggesting considerable individual differences in change between T1 and T2. Crucially, as predicted by the mutualism model, both coupling parameters are positive, such that individuals who start out with a higher matrix reasoning score improve more on vocabulary, and vice versa. The coupling effect from vocabulary T1 scores on gains in reasoning score is r = 0.203, for an r^2 of 4.1%, and the (fully standardized) estimate of reasoning on vocabulary gains is 0.144, for r^2 of 2.1%, corresponding to 'typical' and 'small to typical' effects respectively (Gignac & Szodorai, 2016). Together the self-feedback and coupling parameters account for 30.8% of the individual differences in matrix reasoning score changes, and for 11.7% of the individual differences in vocabulary score changes, illustrating the considerable importance of longitudinal kinematics in cognitive development. Even in the presence of the bivariate coupling parameters the residual change scores are still positively correlated. This is compatible with (although not direct evidence for) additional, unmeasured, cognitive abilities driving change in both vocabulary and matrix reasoning ability. Further control analyses suggested the mutualism model could be equality constrained across sexes without notable drop in model fit ($\chi^2 \Delta$ = 17.184, $df \Delta$ =18, p = 0.51).
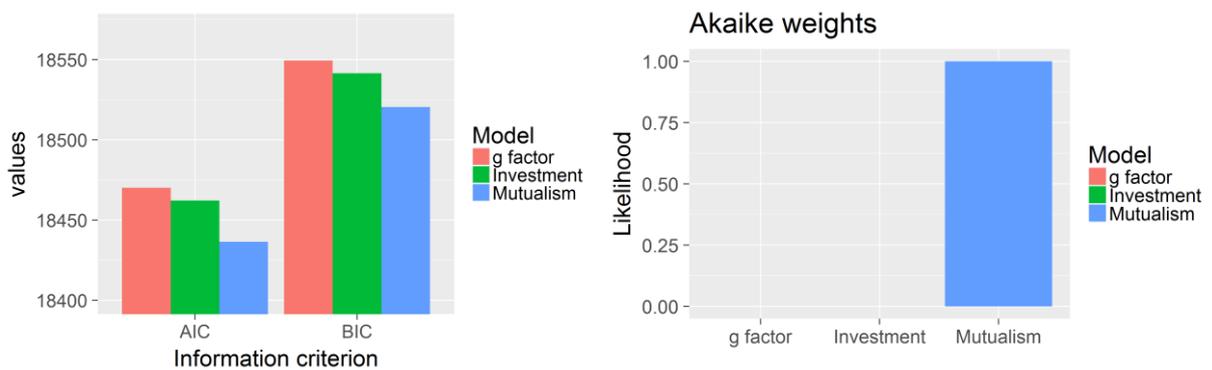

Fig. 3. Information criteria (AIC and BIC) for each of the three models (left), and normalized probabilities for each of the three models using Akaike weights (right).
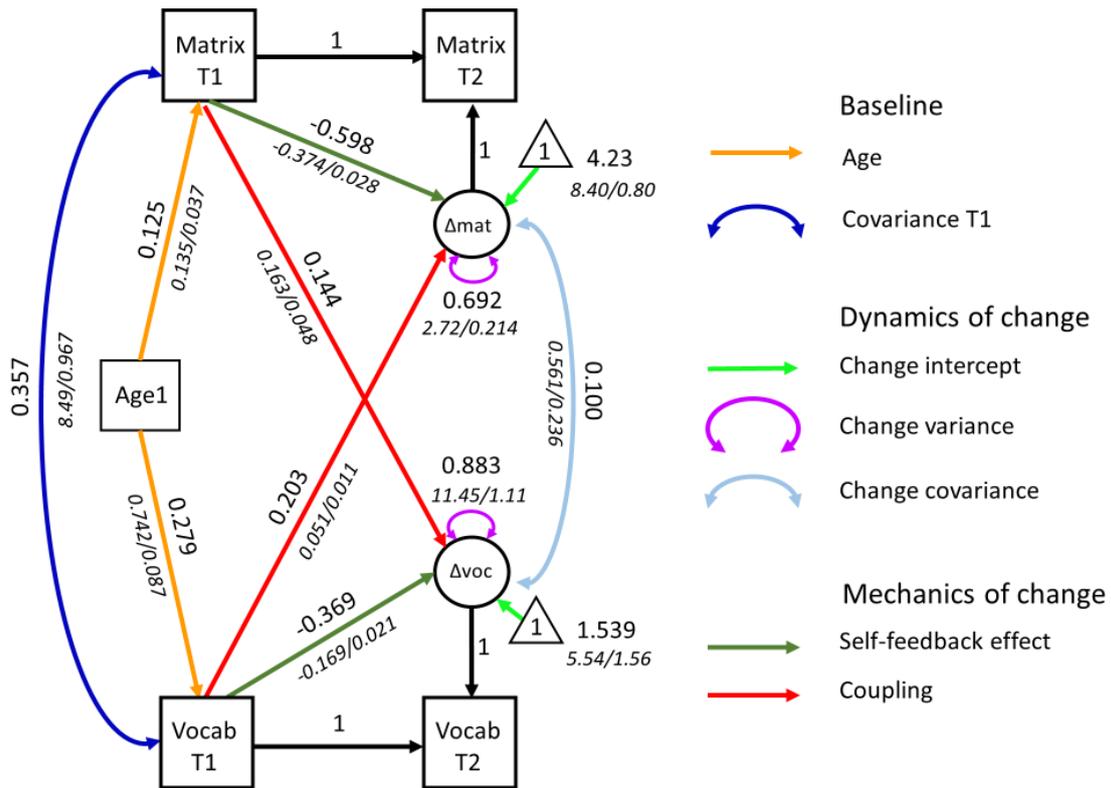
Fig. 4. Estimated parameters for best fitting model (mutualism). Values show fully standardized parameter estimate above and unstandardized parameter estimates/standard errors below paths. Further details are given in Supplementary Table 1.

Using equation 3 and the estimated parameters of the full model (Figure 4) we can visualize the expected change between T1 and T2. Inspired by vector field plots (e.g. Petscher et al. 2016), each arrow represents a (hypothetical) bivariate score at T1 (base of each arrow) and model-implied expected score at T2 (end of arrow) across a range of possible scores. Figure 5 shows the vector field plot and highlights regions where the mutualistic effects are easiest to see.
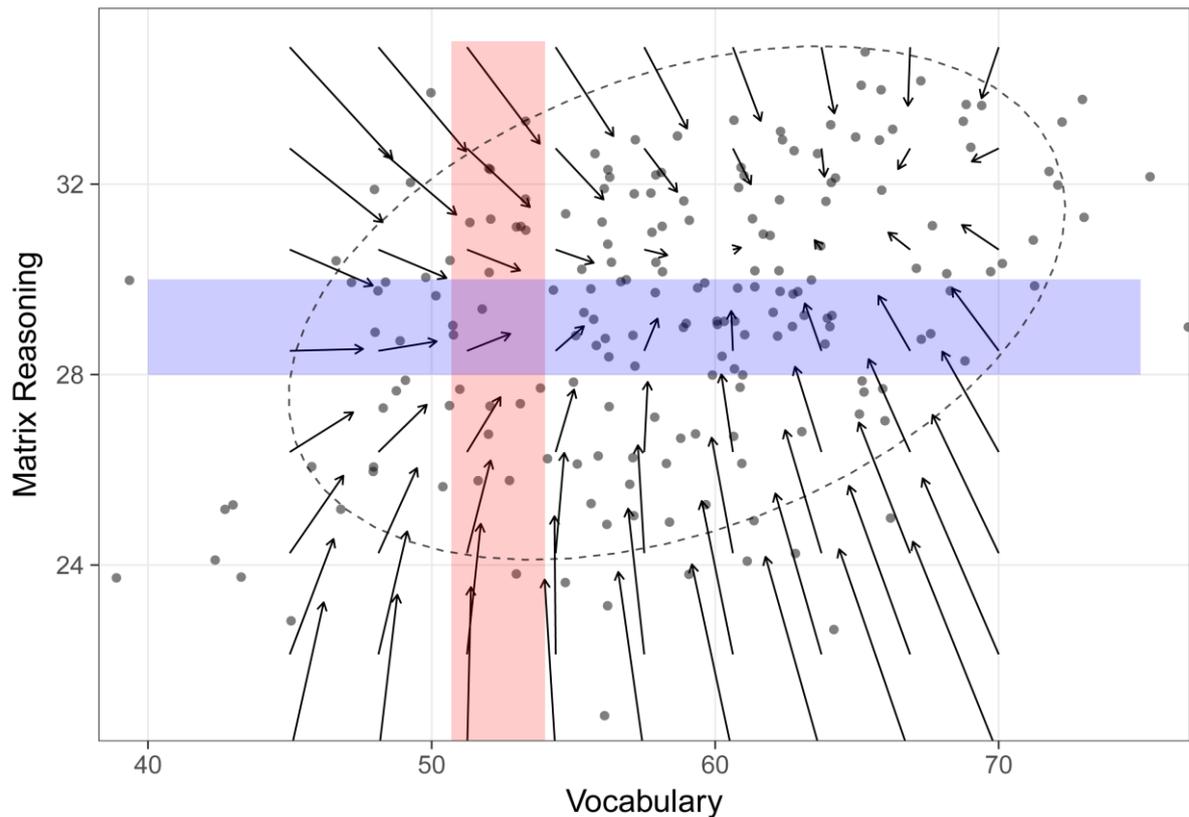
Figure 5: Dynamic change plot. Each dot represents a random subset of raw data at T1, and each arrow represents a model-implied change between T1 and T2. We highlight two sections illustrating mutualism. The blue rectangle (left to right) highlights the positive benefit of higher vocabulary scores on expected change in fluid abilities: The far left arrow is flat, indicating no expected improvement, whereas the far right arrow is pointed upwards, indicating an expected gain of approximately 2.5 points for those with high vocabulary starting scores. Similarly, the red rectangle (top to bottom) illustrates negligible expected vocabulary improvement for low matrix reasoning ability (bottom arrow), but considerable expected vocabulary improvement for those with high reasoning starting scores (top arrow). The ellipse shows the 90% confidence interval for the raw data – model-implied change is most reliable within this region.

Although analytic work (van der Maas et al., 2006) has demonstrated that a *g* factor may

arise through mutualism even in the complete absence of individual differences at the beginning of

development, we think it most likely that *g*-factor and mutualistic processes operate in tandem. For

example, it may be that children show (smaller or larger) consistent individual differences from very

early ages (e.g. Gignac, 2014), which are then amplified by developmental processes such as

mutualism. This is in line with previous suggestions of gene-environment interactions (Briley &

Tucker-Drob, 2013, p. 7) whereby initial differences lead to a 'reciprocal feedback loop between the

phenotype and the environment' that serve to amplify initial differences (Briley & Tucker-Drob,

2013; see also Dickens et al., 2001), a phenomenon observed even in genetically identical mice

(Freund et al., 2013). Such models can also reconcile the high heritability of higher cognitive abilities (Briley & Tucker-Drob, 2013) with considerable environmental impacts, and may serve to partially reconcile more puzzling facts about heritability and cultural load of cognitive tasks (Kan, Wicherts, Dolan, & van der Maas, 2013).

The role of age
In the model above we included age as a linear covariate predicting key scores to account for individual differences due to age at T1 (we discuss alternative parametrizations of age in the discussion). This reflects a hypothesis that age affects scores at T1, but that all aspects of development over time can be captured within the model. Allowing age to directly predict change scores did not improve model fit ($\chi^2\Delta$ = 0.13, $df\Delta$ =2, p = .93), in line with this hypothesis. Notably, this does not necessarily imply that cognitive development occurs at the same rate across development. The decelerating improvement in late adolescence age is captured by the negative self-feedback parameter in reasoning and vocabulary. A second analytic choice is to assume a linear effect of age on scores at T1. An age squared term as predictor of scores at T1 could be fixed to 0 without decrease in model fit ($\chi^2\Delta$ = 3.79, $df\Delta$ =2, p = .15) suggesting a linear term suffices. Third, above we include age as a predictor of the raw vocabulary and matrix reasoning scores at T1 for the mutualism and investment model, but only allow age to predict the g factor in the g model (under the assumption that this factor captures the 'true' shared variance). Although this is in line with the conceptualisation proposed here, we wanted to ensure this analytic choice did not (dis)favour the g model artificially. We therefore fit two additional versions of the g factor model, by including age either as a covariate of only the observed scores at T1 (alternative A), or as covariates of both the observed scores and the g factor (alternative B). The mutualism model was preferred to all three conceptualisations of the g model (ΔBIC: 28.94 (original g factor model), ΔBIC 46.17 (alternative A); ΔBIC 7.09 (alternative B). Together, these analyses suggest that a linear effect of age is sufficient within this sample, that differences in change scores are not affected by age beyond the indirect

15

effect, and that the mutualism model provides a compelling account of dynamic processes during cognitive development.

Discussion

In a large (*N* = 785) development cohort we compared three competing accounts that could explain age-related changes in key cognitive abilities. Using model comparison of data from a sample of adolescents and young adults, we show that mutualism outperforms alternative accounts based on *g* factor and investment theory. Specifically, we find evidence for bivariate coupling between matrix reasoning (as an index of fluid abilities) and vocabulary (as an index of crystallized abilities), such that higher starting points in one cognitive domain are associated with greater developmental gains in the other domain. Our findings refine our understanding of cognitive development in several ways. They suggest that covariance between cognitive abilities is, at least in part, a consequence of a developmental process rather than a single underlying causal entity *g*. Our data provide strong evidence that a model of intellectual development that omits coupling parameters is incomplete.

We can hypothesize several mechanisms to explain the coupling parameters, both direct and indirect. One direction pathway may be that a greater facility with vocabulary and verbal skills allows for swifter, more accurate decomposition into such rules, as well as lower working memory demands for maintenance, especially in younger adults. A more indirect pathway, in line with gene-environment correlations mentioned previously, is that greater vocabulary may be an easily detectable marker of high cognitive ability, which leads to environmental feedback effects in the form of more academically challenging classes or environments to support perceived ability in a manner that generalized to other domains. A final, intriguing possibility is that traditionally fluid tasks such as matrix reasoning may in fact reflect a hybrid of purely fluid abilities (or learning potential) with more strategic, verbal components akin to crystallized abilities (Kühn & Lindenberger, 2016). This would explain both the lifespan trajectories of fluid abilities and the considerable secular gains in fluid abilities in the 20th century (Flynn, 1987).

Our findings suggest a need for a shift away from a narrow focus on desirable cognitive end goals (i.e. adequate performance on abilities such as vocabulary or mathematics) and the incorporation of a simultaneous view across abilities that may have less intrinsic interest, but are essential in their capacity to *support* successful development. For example, skills such as processing speed or working memory may be less important in isolation, but may have coupling to other cognitive skills across the lifespan (Kail, 2007) which in turn may affect later life socioeconomic outcomes. In other words, to facilitate early detection and possibly even effective intervention, it may pay off to focus on abilities that have the strongest coupling strengths, rather than solely on desirable outcomes that are currently below some desirable threshold. For example, Quinn et al. (2015) used dynamic models to show that vocabulary was a leading indicator of gains in reading comprehension, but not vice versa. Such a finding offers insight into the causal pathways of children with reading difficulties, as well as informing appropriate interventions. Similarly, disruptions to typical development were observed by Ferrer et al. (2010) showing that within a subgroup with dyslexia (or 'persistently poor readers'), the coupling between IQ and reading ability observed in typical groups was absent. This suggests not only a possible mechanism for developmental disorders, but shows how multivariate longitudinal models can allow for early detection of developmental challenges that are likely to self-reinforce over time.

Although we compare various developmental models and quantify longitudinal coupling, our sample has certain limitations. First and foremost, we focus on two cognitive subtests alone, yielding a relatively simplistic *g* model. Although both are well validated, have highly similar reliabilities and represent broad cognitive domains, it will be desirable in future studies to represent cognitive abilities by more than one indicator variable, and to sample a wider range of cognitive abilities. Our sample was measured on two occasions, and undoubtedly measurement on more occasions would allow a more precise decomposition of kinetics and kinematics, such as the modelling of lead-lag relations using bivariate dual change score models (e.g. Ghisletta & Lindenberger, 2003). Here we show that baseline scores are positively associated with cross-domain rates of change. With three or

more waves it is possible to use the change scores at T themselves to predict the change scores at T1. Moreover, if age is sampled at sufficient frequency, it is possible to examine latent changes as a function of age itself (ΔY_{agei}), rather than as testing occasion (ΔY_{ti}), obviating the need for covariates by 'binning' individuals' scores into age bins and estimating models using methods that account for missingness (e.g. Grimm, An, McArdle, Zonderman, & Resnick, 2012; Voelkle & Oud, 2017).

An additional challenge with repeated measures data is improvement in test scores due to practice effects, which may inflate developmental gains or attenuate age-related decline (Lövdén, Ghisletta, & Lindenberger, 2004; Rabbitt, 2001; Salthouse & Tucker-Drob, 2008). Although in our sample practice effects may have led to greater increases between T1 and T2, it is unlikely that these effects impact our conclusions regarding mutualism. First, such practice effects will lead to an increase in test scores that are a combination of true (developmental) gains and increases due to practice effects (although see (Lövdén et al., 2004) on the interpretation of practice effects). Notably, if one interprets the gains between T1 and T2 as a combination of 'true' gains and practice effects, this would entail an underestimate of the mutualism effect (as the effect size reflects the prediction on the total gains rather than the non-practice related gains). In principle, a sufficiently large number of time points spaced at unequal retest intervals would allow for a decomposition of re-test effects, but both practical challenges as well as the inherent collinearity of re-test occasions with time intervals has proved methodologically challenging (Hoffman, Hofer, & Sliwinski, 2012).

Finally, we observe our effects in adolescents and young adults, limiting the generalizability to this developmental period alone. We hypothesize that the coupling effects we observe are likely to be stronger earlier in life, and the self-feedback parameters weaker, as developmental change in higher cognitive abilities is most rapid in during pre- and early adolescence. The other end of the lifespan provides for several intriguing questions. It is conceivable that mutualism only occurs during early development, with other processes and mechanisms taking over after initial peaks are reached. However, we suggest that studying later life decline from the perspective of mutualism might prove

a promising avenue of future work. If dynamic coupling is crucial for maintenance of cognitive abilities in later life, this may explain why declines are often strongly correlated (see Ghisletta & Lindenberger, 2003; Tucker-Drob, 2011 for further exploration of this hypothesis). Large longitudinal cohorts using similar tests across the entire lifespan will allow for the investigation of possible 'regime changes' within the same cohort.

Future work should study multi-wave, multi-domain cognitive data using principled model-selection methods to better fully capture the underlying dynamics of cognitive development. Data of high temporal resolution would allow us to move beyond group level dynamics of individual differences to the ultimate goal, namely that of estimating individual differences in intra-individual dynamics over time. The investigation of individual coupling parameters across domains, and across the lifespan, is likely to yield a wealth of information on cognitive development in health and disease. The recent convergence of novel modelling techniques, large scale data gathering facility via tools such as smartphones and the integration of behavioural datasets with data from neural and genetic sources of evidence together promise to provide new insight into some of the most elusive, yet fundamental, questions in cognitive psychology.

Author contributions

Conflict of interest statement:

E.T.B. is employed half-time by the University of Cambridge and half-time by GlaxoSmithKline; he holds stock in GlaxoSmithKline.

References

Aichele, S., Rabbitt, P., & Ghisletta, P. (2015). Life Span Decrements in Fluid Intelligence and

Processing Speed Predict Mortality Risk. *Psychology and Aging*, *30*(3), 598–612. Retrieved from

http://psycnet.apa.orgpsycarticles/2015-27315-001

Briley, D. A., & Tucker-Drob, E. M. (2013). Explaining the increasing heritability of cognitive ability

across development: a meta-analysis of longitudinal twin and adoption studies. *Psychological

Science*, *24*(9), 1704–13. http://doi.org/10.1177/0956797613478618

Cattell, R. B. (1971). *Abilities: their structure, growth, and action.* Boston: Houghton-Mifflin.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing

Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–

255. http://doi.org/10.1207/S15328007SEM0902_5

Christensen, H., Batterham, P. J., & Mackinnon, A. J. (2013). The Getting of Wisdom: Fluid

Intelligence Does Not Drive Knowledge Acquisition. *Journal of Cognition and Development*,

14(2), 321–331. http://doi.org/10.1080/15248372.2012.664590

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*, 453–482.

http://doi.org/10.1146/annurev-psych-120710-100353

Dickens, W. T., & Flynn, J. R. (2001). Heritability Estimates Versus Large Environmental Effects: The

IQ Paradox Resolved. *Psychological Review*, *108*(2), 346–369. http://doi.org/10.1037//0033-

295X

Ferrer, E., & McArdle, J. J. (2004). An Experimental Analysis of Dynamic Hypotheses About Cognitive

Abilities and Achievement From Childhood to Early Adulthood. *Developmental Psychology*,

40(6), 935–952.

Ferrer, E., Shaywitz, B. A., Holahan, J. M., Marchione, K., & Shaywitz, S. E. (2010). Uncoupling of

reading and IQ over time: empirical evidence for a definition of dyslexia. *Psychological Science*,

21(1), 93–101. http://doi.org/10.1177/0956797609354084

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological

Bulletin, *101*(2), 171–191. http://doi.org/10.1037/0033-2909.101.2.171

Freund, J., Brandmaier, A. M., Lewejohann, L., Kirste, I., Kritzler, M., Krüger, A., … Epstein, D. (2013). Emergence of individuality in genetically identical mice. *Science (New York, N.Y.)*, *340*(6133), 756–9. http://doi.org/10.1126/science.1235294

Fuhs, M. W., & Day, J. D. (2011). Verbal ability and executive functioning development in preschoolers at head start. *Developmental Psychology*, *47*(2), 404–416. Retrieved from http://psycnet.apa.orgjournals/dev/47/2/404

Ghisletta, P., & Lindenberger, U. (2003). Age-Based Structural Dynamics Between Perceptual Speed and Knowledge in the Berlin Aging Study: Direct Evidence for Ability Dedifferentiation in Old Age. *Psychology and Aging*, *18*(4), 696–713. http://doi.org/10.1037/0882-7974.18.4.696

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence*, *42*, 89–97. http://doi.org/10.1016/j.intell.2013.11.004

Gignac, G. E. (2016). Residual group-level factor associations: Possibly negative implications for the mutualism theory of general intelligence. *Intelligence*, *55*, 69–78. http://doi.org/10.1016/j.intell.2016.01.007

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. http://doi.org/10.1016/j.paid.2016.06.069

Gottfredson, L. S. (2002). Where and Why g Matters:Not a Mystery. *Human Performance*, *15*(1–2), 25–46. http://doi.org/10.1080/08959285.2002.9668082

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence Predicts Health and Longevity, but Why? *Current Directions in Psychological Science*, *13*(1), 1–4. http://doi.org/10.1111/j.0963-7214.2004.01301001.x

Grimm, K. J., An, Y., McArdle, J. J., Zonderman, A. B., & Resnick, S. M. (2012). Recent changes leading to subsequent changes: Extensions of multivariate latent difference score models. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(March 2015), 268–292. http://doi.org/10.1080/10705511.2012.659627

Hoffman, L., Hofer, S. M., & Sliwinski, M. J. (2012). On the Confounds among Retest Gains and Age-Cohort Differences in the Estimation of Within-Person Change in Longitudinal Studies: A Simulation Study. *Psychological Aging*, *29*(6), 997–1003. http://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Praeger.

Kail, R. V. (2007). Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. *Psychological Science*, *18*(4), 312–3. http://doi.org/10.1111/j.1467-9280.2007.01895.x

Kan, K.-J., Wicherts, J. M., Dolan, C. V, & van der Maas, H. L. J. (2013). On the nature and nurture of intelligence and specific cognitive abilities: the more heritable, the more culture dependent. *Psychological Science*, *24*(12), 2420–8. http://doi.org/10.1177/0956797613493292

Kievit, R., Brandmaier, A., Ziegler, G., Harmelen, A.-L. van, Mooij, S. de, Moutoussis, M., … Dolan, R. (2017). Developmental cognitive neuroscience using Latent Change Score models: A tutorial and applications. *bioRxiv*, 110429. http://doi.org/10.1101/110429

Kühn, S., & Lindenberger, U. (2016). Research on human plasticity in adulthood: A lifespan agenda. In K. W. Schaie & S. L. Willis (Eds.), *Handbook of the psychology of aging* (8th ed., pp. 105–123). Amsterdam: Academic Press.

Lövdén, M., Ghisletta, P., & Lindenberger, U. (2004). Cognition in the Berlin Aging Study (BASE): The First 10 Years. *Aging, Neuropsychology, and Cognition*, *11*(2–3), 104–133. http://doi.org/10.1080/13825580490510982

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*(October 2008), 577–605. http://doi.org/10.1146/annurev.psych.60.110707.163612

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. Retrieved from

http://psycnet.apa.orgjournals/dev/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. Decade*

of behavior. (pp. 139–175).

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. http://doi.org/10.1016/S1041-6080(00)00036-4

McCrimmon, A. W., & Smith, A. D. (2013). Test Review: Review of the Wechsler Abbreviated Scale of

Intelligence, Second Edition (WASI-II). *Journal of Psychoeducational Assessment*, *31*(3), 337–

341. Retrieved from http://eric.ed.gov/?id=EJ1011833

Penke, L., Maniega, S. M., Bastin, M. E., Valdés Hernández, M. C., Murray, C., Royle, N. A., … Deary, I.

J. (2012). Brain white matter tract integrity as a neural foundation for general intelligence.

Molecular Psychiatry, *17*(10), 1026–30. http://doi.org/10.1038/mp.2012.66

Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental relations between

vocabulary knowledge and reading comprehension: a latent change score modeling study.

Child Development, *86*(1), 159–75. http://doi.org/10.1111/cdev.12292

Rabbitt, P. M. (2001). Identifying and separating the effects of practice and of cognitive ageing

during a large longitudinal study of elderly community residents. *Neuropsychologia*, *39*(5), 532–

543. http://doi.org/10.1016/S0028-3932(00)00099-3

Rast, P., & Hofer, S. M. (2014). Longitudinal design considerations to optimize power to detect

variances and covariances among rates of change: simulation results based on actual

longitudinal studies. *Psychological Methods*, *19*(1), 133–54. http://doi.org/10.1037/a0034524

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical*

Software, *48*(2), 1–36.

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the

interpretation of longitudinal change. *Neuropsychology*, *22*(6), 800–11.

http://doi.org/10.1037/a0013091

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist*, *49*(4), 304–313. Retrieved from http://psycnet.apa.orgjournals/amp/49/4/304

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research - Online*, *8*(2), 23–74.

Schmidt, F. L., & Crano, W. D. (1974). A test of the theory of fluid and crystallized intelligence in middle- and low-socioeconomic-status children: A cross-lagged panel analysis. *Journal of Educational Psychology*, *66*(2), 255–261.

Spearman, C. (1927). *The abilities of man.* Oxford: Macmillan Publishers Limited.

Tucker-Drob, E. M. (2011). Global and domain-specific changes in cognition throughout adulthood. *Developmental Psychology*, *47*(2), 331–343. Retrieved from http://psycnet.apa.orgjournals/dev/47/2/331

Valentin Kvist, A., & Gustafsson, J.-E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment theory. *Intelligence*, *36*(5), 422–436. http://doi.org/10.1016/j.intell.2007.08.004

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. http://doi.org/10.1037/0033-295X.113.4.842

Voelkle, M. C., & Oud, J. H. L. (2017). Relating Latent Change Score and Continuous Time Models Relating Latent Change Score and Continuous Time Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 366–381. http://doi.org/10.1080/10705511.2014.935918

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. http://doi.org/10.3758/BF03206482

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence—Second Edition.* Minneapolis:

Pearson.

Weiland, C., Barata, M. C., & Yoshikawa, H. (2014). The Co-Occurring Development of Executive

Function Skills and Receptive Vocabulary in Preschool-Aged Children: A Look at the Direction of

the Developmental Pathways. *Infant and Child Development*, *23*(1), 4–21.

http://doi.org/10.1002/icd.1829

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. http://doi.org/10.1111/j.1750-8606.2009.00110.x

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for

Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational*

and Psychological Measurement, *76*(6), 913–934. http://doi.org/10.1177/0013164413495237

Regressions	Estimate	se	z-value	p-value	lower 95% CI	upper 95% CI
t1ssmatrix~Age	0.135	0.037	3.626	<0.001	0.062	0.207
t1ssvocab~Age	0.742	0.087	8.5	<0.001	0.571	0.913
etamat~T1Vocabulary	0.051	0.011	4.829	<0.001	0.031	0.072
etamat~T1Matrix	-0.374	0.028	-13.559	<0.001	-0.428	-0.32
etavoc~T1Matrix	0.163	0.048	3.396	0.001	0.069	0.257
etavoc~T1Vocabulary	-0.169	0.021	-7.92	<0.001	-0.211	-0.127
(Co)Variances						
etamat~~etamat	2.727	0.214	12.723	<0.001	2.307	3.147
T1Matrix~~T1Matrix	9.946	0.647	15.372	<0.001	8.678	11.214
etavoc	11.445	1.106	10.349	<0.001	9.278	13.613
T1Vocabulary~~T1Vocabulary	56.75	3.112	18.236	<0.001	50.651	62.85
T1Matrix~~T1Matrix	9.946	0.647	15.372	<0.001	8.678	11.214
Age~~Age	8.723	0.305	28.615	<0.001	8.126	9.321
etavoc~~etamat	0.561	0.236	2.378	0.017	0.099	1.023
T1ssmatrix ~~ T1Vocabulary	8.486	0.967	8.776	<0.001	6.591	10.381
Intercepts						
etamat	8.403	0.8	10.507	<0.001	6.835	9.97
t1ssmatrix	26.476	0.728	36.388	<0.001	25.05	27.902
etavoc	5.54	1.558	3.555	<0.001	2.486	8.595
age1	19.052	0.105	180.74	<0.001	18.846	19.259
t1ssvocab	44.428	1.645	27.009	<0.001	41.204	47.652

Supplementary Table 1. Raw parameter estimates and confidence intervals

Neuroscience in Psychiatry Network (NSPN) Consortium author list

Chief investigator:
Ian Goodyer (Ed Bullmore from 01/01/2017)

Principal investigators:
Edward Bullmore (Ian Goodyer from 01/01/2017)
Raymond Dolan
Peter Fonagy
Peter Jones

Associated faculty:
Michael Moutoussis
Pasco Fearon
Anne-Laura van Harmelen
Petra Vértes
Kirstie Whitaker
Rogier Kievit

Project managers:
Becky Inkster
Gita Prabhu

Data managers:
Cinly Ooi
Barry Widmer

Research assistants:
Ayesha Alrumaithi
Sarah Birt
Kalia Cleridou
Hina Dadabhoy
Sian Granville
Elizabeth Harding
Alexandra Hopkins
Daniel Isaacs
Janchai King
Danae Kokorikou
Harriet Mills
Ciara O'Donnell
Sara Pantaleone

Administration team:
Junaid Bhatti
Laura Villis