Running Head: PSYCHOMETRIC MODELING OF REDUCTIVE PSYCHOLOGY

Mind the gap: A psychometric approach to the reduction problem

Rogier A. Kievit
Department of Psychology, University of Amsterdam

Jan-Willem Romeijn
Department of Philosophy, Groningen University

Lourens J. Waldorp, Jelte M. Wicherts, H. Steven Scholte and Denny Borsboom
Department of Psychology, University of Amsterdam

Correspondence should be addressed to:

Rogier A. Kievit

University of Amsterdam, Department of Psychological Methods

Roetersstraat 15

1018WB Amsterdam, The Netherlands

(+31) 020-5256688

E-mail: r.a.kievit@uva.nl

Abstract

Cognitive neuroscience involves the simultaneous analysis of behavioral and neurological data. Common practice in cognitive neuroscience, however, is to limit analyses to the inspection of descriptive measures of association (e.g., correlation coefficients). This practice, often combined with little more than an implicit theoretical stance, fails to address the relationship between neurological and behavioral measures explicitly. This paper argues that the reduction problem, in essence, is a *measurement problem*. As such, it should be solved by using psychometric techniques and models. We show that two influential philosophical theories on this relationship, identity theory and supervenience theory, can be easily translated into psychometric models. Upon such translation, they make explicit hypotheses based on sound theoretical and statistical foundations, which renders them empirically testable. We examine these models, show how they can elucidate our conceptual framework and examine how they may be used to study foundational questions in cognitive neuroscience. We illustrate these principles by applying them to the relation between personality test scores, intelligence tests and neurological measures.

Keywords:

Cognitive neuroscience, Measurement theory, Philosophy of mind, Reductionism, Structural equation modeling, Psychometrics

Mind the gap: A psychometric approach to the reduction problem

"There is nothing more practical than a good theory"

(Lewin, 1951)

One of the hallmark neuroscientific findings on the 20th century is the discovery of the retinotopic representation of early visual areas (e.g. Hubel & Wiesel, 1968; Tootell, Switkes, Silverman & Hamilton, 1988). That is, activation patterns in the occipital lobe show striking structural similarity to visually presented geometric patterns. Such findings, originally only possible in animal research, have been replicated in humans in more indirect form. For instance, Miyawaki et al. (2008) show how basic visual stimuli (including letters) can be decoded from brain activity with high accuracy (>90%), based upon weighted linear combinations of voxel activation patterns. For such low-level perceptual processes, it seems plausible to consider the observation of activity patterns in early visual areas as a measurement of what particular stimulus is presented to a particular subject. However, the measurement theoretical relationship is not always so clear. Consider the following example: you are invited to a job interview for a high-status position. Shortly after being seated, the interviewer takes out a tape measure and starts measuring your skull. Upon enquiring what is going on, the interviewer tells you he just "measured your intelligence". In response to your protesting that such a procedure does nothing of the sort, the interviewer shows you a list of high profile journal articles that report a moderate but consistent correlation between brain volume and IQ (e.g. McDaniel, 2005; Posthuma et al., 2002). You may believe that such a procedure does not *measure* intelligence, but this appears to run counter to the view in cognitive neuroscience[1] that physiological measures may serve as measures of psychological attributes. We will later return to the empirical formalization of this question.

What is the essential difference between these two situations? Both take information about the brain to predict a certain (psychological) property, and both are based on statistically significant measures of association, but at the same time they seem quite distinct. It seems thoroughly unclear how to resolve this issue. This raises the question of how cognitive neuroscientists actually represent the relationship between the two classes of measures, and what presentation would justify the interpretation of neurological measures as representing psychological attributes.

The general practice in cognitive neuroscience is to limit statistical analyses to the study of descriptive measures of association (e.g., correlation coefficients). In fact, some authors have argued that cognitive neuroscience is by its very nature correlational (e.g. Jung & Haier, 2007, p. 148). However, this would leave open important questions: What is the precise relationship *between* these two classes of measurement? Does one measured property cause the other? Or is it the other way around? Do the different kinds of data really represent measurements of the *same thing*? Many papers implicitly embrace one of these options, possibly because there simply is no "value-free" way in which to describe the relationship between behavioral measurements and neurological measurements - unless, perhaps, if one is satisfied with the conclusion that "they both just happened". Certainly it is desirable (if not tempting) to attach some theoretical interpretation to the established empirical relationship between psychological-behavioral and neurological measures. However, the mere inspection of correlation coefficients provides no sound basis for deciding between different theoretical interpretations.

The suggestive nature of correlations between neurological and behavioral or psychological variables has thus led the literature to become densely populated with euphemisms, metaphors, and just-so stories regarding their precise relation.

Psychological processes and mental concepts can be "associated with" (Mobbs, Hagan, Azim, Menon, & Reiss, 2005, p. 16502), "recruit" (Morris & Mason, 2009, p. 59) "located in" (Hadjikhani, Liu, Dale, Cavanagh, & Tootell, 1998, p. 237), "instantiated in" (Davidson, 2004, p. 222), "subserved by" (Luna et al., 1998, p. 40), "related to" (McGregor, 2006, p. 304), "generated by" (DeYoung & Gray, 2009, p. 2) "served by" (Demetriou & Mouyi, 2007, p. 157), "implicated in" (Grossman & Blake, 2002, p. 1167), "correlated with" (Canli et al., 2001, p. 33), or "caused by" (Levine, 1999, p. 352) a dizzying array of cortical areas, process loops, frequency activation patterns, activation systems, structural differences and neurotransmitter levels. The conceptual elephant in the room is *how* such varied measures and concepts relate to each other, what they are indicators *of*, what the causal relationships between them are, and how we should structure our empirical studies so as to maximize the theoretical payoff of cognitive neuroscientific research.

Conceptual problems in reductive psychological science have not gone unnoticed. Several researchers have taken on theoretical, statistical and scientific issues concerning reductionism and reductive psychological science. For instance, Bennett and Hacker suggested that the vocabulary employed in neuroscientific studies is conceptually flawed. One of the issues they raised is the "mereological fallacy", or "assigning to a part what can only be assigned to a whole" (Bennett & Hacker, 2003, p. 68). They identified this fallacy in statements such as "the frontal lobe engages in executive functioning". They argue that this practice is philosophically misguided, and reflects a conceptual problem within reductive neurological science. Ross and Spurrett (2004) argued that functionalist cognitive psychology requires a solid metaphysical underpinning of its the conceptual and scientific foundations, if it is to function as an autonomous field of scientific inquiry. Other researchers (Fodor, 1974; Gold & Stoljar,

1999; Nagel, 1961) have examined the philosophical foundations of reductionism, and explicated the requirements necessary for reductionist claims. Recent efforts have examined whether the ontology of psychological categories is suitable for reductive analysis, and argued that an approach in terms of psychological primitives may be more appropriate (Feldman Barrett, 2009). Criticism of reductive studies has not been purely philosophical. In a controversial paper, Vul, Harris, Winkielman, and Pashler (2009) argued that a large number of claims in social neuroscience studies are overstated, and that overly liberal methodology has resulted in unrealistically high correlations between physiological and behavioral measurements (but see also associated comments to Vul et al., 2009).

These papers have focused largely on what conclusions that are *not* permissible, methodologies that should *not* be used and philosophical claims that can *not* be made. The aim of the current paper is to address the criticisms raised by the above authors by providing conceptual and statistical tools that may elucidate the type of claims that we *can* make in reductive science, and developing the requirements such claims should satisfy.

Cognitive neuroscience typically attempts to establish the relationships between at least two distinct explanatory levels, namely the neurological and psychological level (Oppenheim & Putnam, 1958). As such, it has drawn much attention from philosophers, who have articulated and analyzed many theoretical positions regarding the relations between the two levels of analysis (e.g. Churchland, 1981, 1985; Kim, 1984; Lewis, 1966; Putnam, 1973). Several recently developed perspectives on reduction, seeking to integrate certain developments in, for example, molecular neuroscience (e.g., the 'New Wave Reductionism' promoted by Bickle, 1998). It would seem that if such positions could be translated into statistical models that are

testable given the data that cognitive neuroscientists commonly have at their disposal, the theories articulated in the philosophy of mind could serve as a means to conceptually organize and guide the analysis of neurological and behavioral data. That is, if it were possible to find a statistical model representation of, say, the basic assumption that the property measured by means of fMRI recordings actually *is the same as* the property measured through a set of cognitive tasks or questionnaire items (i.e., identity theory; Lewis, 1966), then both the philosopher of mind and the empirical researcher in cognitive neuroscience would benefit: The philosopher of mind, because there would exist a means to empirically test theories that have hitherto been regarded as being speculative metaphysics at best. The empirical researcher, as this could provide statistical tests of interpretations of the data that go well beyond the speculative interpretations of correlations that currently pervade the literature.

How could statistical models be of help to the empirical researcher in cognitive neuroscience? Recall that, in this area of research, one typically aims to build connections between measures related to behavior, psychological attributes and processes on the one hand, and the (relative) activity and physiological characteristics of the brain on the other hand. In psychometrics, we can represent such diverging classes of measurement in a single measurement model. The central idea of this paper is that by varying the way in which a theoretical attribute relates to the observations, models can be built that allow for a more detailed investigation of the relation between neurological and psychological measurements than are in use to date. This paper proposes that modeling techniques suited to this purpose need not be developed for this purpose, because they already exist. These mathematically tractable models with known statistical properties, developed largely in the discipline of psychometrics, can map theoretical positions about the relationship between brain and behavioral

measurements as developed in the philosophy of mind in impressive detail. We argue that the statistical formalization of theoretical positions is both possible and desirable and we offer the empirical and conceptual tools to do so. Perhaps most importantly, such formalizations make clear that the reduction problem is, in essence, not just a substantive or philosophical problem, but a standard measurement problem that can be attacked by using standard measurement models of psychometrics. However, such models have been scarcely applied in cognitive neuroscience. From this perspective, therefore, it seems as if most empirical work has, instead of solving the measurement problem, largely circumvented it.

The structure of this paper is as follows. We first define the two classes of measurement under study. Subsequently, we examine two important theories from the philosophy of mind literature that explicitly treat the relationship between these higher and lower order properties, namely identity theory and supervenience theory. In addition, we introduce two psychometric models that may be used to represent these theoretical positions. Finally, we illustrate these ideas by applying both models to a dataset that examines the relationship between a personality dimension and physiological properties of the brain.

Two types of data

In the models we discuss below we distinguish between the two classes of data that feature in most cognitive neuroscientific studies. First, we refer to data that pertain to psychological attributes or mental processes as P-indicators. These include psychological measurements, such as “solving puzzle x”, “choosing answer c” or “the number of objects retained in working memory”. Second, we refer to data that pertain to neurological processes or characteristics as N-indicators. These may include data such as electrical measures of cortical activity (EEG), speed of processing

measurements, blood oxygenation level-dependent (BOLD)-signals, as well as physiological indicators such as gray matter density, brain volume, or neurotransmitter levels. The psychological indicators are indexed to denote either different questions on a test (P1 is one question, P2 another) or different types of measurement (P1 is an IQ-score, P2 a reaction time test). Neurological indicators are indexed to denote, for example, different regions of the brain (e.g. N1 is a BOLD measurement of the posterior parietal region, N2 of the amygdala), or different types of physiological variables (N1 is gray matter density, N2 is neural processing speed).

For example, in a cognitive neuroscientific study of *empathy,* psychological measurements of empathy could include P-indicators such as questionnaires, self-reports or behavioral assessments. In contrast, neurological measurements would include N-indicators such as the level of BOLD-activation in certain cortical regions in response to seeing another person suffer (see Decety and Jackson, 2004, for a review of the neurological study of empathy). It is clear that these two classes of data are qualitatively distinct (see also Barrett, 2009). Therefore, researchers require a conceptual foundation that informs data analytic techniques that can be used to test hypothesized relationships between two such sets of data. Two theories in the philosophy of mind provide a conceptualization of the relation between psychological and neurological properties: identity theory and supervenience theory. Identity theory states that psychological and neurological variables depend on the same underlying attribute, while supervenience theory states that neurological variables determine the psychological attributes. These theories are discussed briefly.

Philosophy of mind

Identity theory

The thesis of identity theory was proposed in several forms throughout the latter half of the 20th century. It has its roots in seminal papers such as those of Place (1956) and Smart (1959). In its most commonly accepted interpretation, as described in Lewis (1966), identity theory holds that psychological processes and attributes are *identical* to their neurological realizations.

The attractiveness of identity theory lies in the relatively non-problematic assignment of causal powers to mental events. Because a mental event or state is identical to a (particular) neural realization at any given time, it has the same causal powers as the neurological state that realizes it. This implies that in a cognitive neuroscientific study of a particular psychological attribute, one is essentially measuring the same attribute using two different measurements. The P and N indicators therefore have a common referent. This conceptualization paints a thoroughly realist picture of psychological attributes, in which the reality of these attributes is grounded in their physical realization.

Supervenience

Supervenience provides an alternative way of conceptualizing the relation between psychological and neurological measurements. Different interpretations of supervenience have been formulated in relation to a wide range of philosophical topics (Collier, 1988; Hare, 1952; Horgan, 1993). Historically, the concept arose from attempts to ground the properties of higher-level concepts such as beauty, morality and consciousness in their lower order realizations. The definition of supervenience is as follows: A property X can be said to supervene on lower order properties Y if *there cannot be X-differences without Y differences.* Thus, the presence of Y-differences is a necessary (but insufficient) condition for the presence of X-differences. This relation of necessity is a sufficient condition for calling the relation one of supervenience.

Consider, for example, the attribute of being morally good. Under supervenience theory, two people cannot differ in terms of morality (X) without being different on lower order Y attributes (e.g., behavioral ones; not stealing, cheating, donating money to charity etc). Equivalently, if there are no differences in the lower order attributes (Y, or behavioral attributes) then there are necessarily no differences in the higher order attribute (X, or morality). This is the sense in which morality supervenes on its lower order attributes. Properties such as morality and beauty are "along for the ride", so to speak: they supervene on lower order properties that *do not necessarily share* all the characteristics that relate to the supervenient property. The atoms that make up the Mona Lisa are not beautiful, and neurons are not neurotic: such higher order properties *supervene* on the lower order properties in a causally asymmetric manner.

The philosophical details of supervenience are still the subject of theoretical perspectives and debates. Its most vocal advocate in the realm of psychology has been Jaegwon Kim. His supervenience perspective on psychology (Kim, 1982, 1984, 1985) defines psychological attributes as supervenient on neurological realizations. That is, psychological attributes are completely determined by, or realized in, their neurological constituents. Supervenience has been the topic of various recent debates on specific alternative interpretations of the concept, varying in terms of modal strength and necessity (Horgan, 1993; Howell, 2009). Although these are of interest in and of themselves, a comprehensive discussion would lead us too far astray from our current aim. For sake of parsimony, we will adopt Kim's more traditional definition of strong supervenience[2]. Kim defines the supervenient status of higher and lower level properties A and B's, respectively, as follows: "Necessarily, for any x and y, if x and y share all properties in B, then x and y share all properties in A – that is, indiscernibility in B entails indiscernibility in A" (Kim, 1987, p. 315). The relationship of

supervenience is asymmetric, as neurological states or structures *can* differ while the higher order property remains the same (because lower order differences are necessary, but not sufficient, for higher order differences).

This implies that supervenience allows for multiple realizability (Putnam, 1980); several different combinations of N-realizations may lead to the same (value of the) psychological attribute. Because of this asymmetry, authors such as Kim give causal priority to the lower order realizations: the neurological indicators are considered to determine the causal properties of the system completely. Supervenience is consistent with a many-to-one mapping of the lower to the higher order properties, but not with an isomorphism (which would hold if all relations between instances of the lower order terms are preserved in the higher order relations), and therefore precludes identity.

To illustrate this, consider the following transaction. If John gives Jane five dollars (higher order process) then that means that John has either given Jane a five-dollar bill, has handed her the equivalent sum in coins, or has electronically transferred five dollars to Jane's bank account, etc. (lower order processes). Thus, the entire class of these lower order processes maps onto the same higher order process. If we know that John gave Jane five dollars, we can therefore infer that he performed one of the actions in the corresponding lower order class. However, we cannot determine which of these actions he performed (no isomorphism). It is evident that an identity theory perspective on such a monetary transaction is questionable: John giving Jane five dollars cannot simultaneously be identical to writing a cheque *and* to handing over a five dollar bill. We now show how such restrictions and theoretical considerations can be translated to and mapped on psychometric models. To do so, we must first examine the basic properties of the models that we shall consider.

Psychometrics

Psychometrics is concerned with the theoretical and technical development of measurement procedures and statistical inference techniques. One of the techniques, developed in tandem with psychometric theory, is structural equation modeling (SEM). SEM consists of both a graphical and a (equivalent) linear mathematical representation of the hypothesized causal directions and statistical associations between measured and latent variables. Such representations imply a specific covariance structure, which may be tested given appropriate data. Specifically, one can evaluate whether the observed covariance matrix is consistent with the covariance structure associated with the specified linear relationships. For a thorough introduction to structural equation modeling with latent variables, see Bollen (1989).

In SEM, there are two broad classes of model specification that we consider in detail, namely *formative* and *reflective* models (Bagozzi, 2007; Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). Both classes model relationships between observed variables and latent variables. Here "observed variables" refer to the variables as they appear in a data file, and "latent variables" refer to variables that are not directly observable, so that their values can only be estimated indirectly (Bollen, 2002; Borsboom, 2008). Many of the properties central in psychological science (e.g., intelligence, personality, working memory capacity) cannot be determined with certainty from the data, and are therefore properly conceived of as latent variables.

Formative and reflective models provide two ways of connecting a theoretical attribute, as targeted by a measurement procedure, to the observations. We discuss the conceptual difference between these two models in relation to the distinction between identity theory and supervenience theory. We present the models using standard SEM notation (Jöreskog & Sörbom, 1996). As mentioned, SEM permits the specification of

linear relations between the observed and latent variables as implied by theoretical considerations, and the evaluation of the degree to which the observed covariance structure is consistent with that implied by the theoretical relations. The models can either allow for tentative confirmation, in the sense that they fit the data, or rejection, in the sense that they can be overspecified or display poor fit. Thus, these models are amenable to empirical tests.

Reflective models

The most common measurement model in psychology is called the reflective model. Instances of the model include Item Response Theory models (Embretson & Reise, 2000) such as the models of Rasch (1960) and Birnbaum (1968) , and, most relevant to the present paper, the linear factor model (Lawley & Maxwell, 1963; Jöreskog, 1971; Mellenbergh, 1994). In reflective models, latent variables are seen as the underlying *cause* of variability on the measurable indicators (Bollen, 2002; Bollen & Lennox, 1991; Borsboom, Mellenbergh, & van Heerden, 2003; Edwards & Bagozzi, 2000). In other words, the hypothesized causal direction runs *from* the latent attribute *to* the measurable indicators. The various measurable indicators are seen as *reflecting* the underlying attributes. Perhaps the most common example of a reflectively measured attribute is intelligence. The conceptualization of intelligence posits a factor *g* that refers to the common cause of variability on intelligence test questions or subtests (Glymour, 1998).

A reflective model of *g* is given in Figure 1. In the figure, three indicators (for example IQ-test items or subtest scores) are conceptualized as measurements of a single underlying attribute (this is a simple, non-hierarchical model of *g,* chosen for illustrative purposes). Indicators of a reflectively measured latent variable should (after appropriate recoding) intercorrelate positively, capture the range of effects the latent variable can have, and be acceptably reliable (i.e., be characterized by acceptable levels of measurement error). In addition, in correctly specified reflective models, latent variables should be referentially stable. That is to say that the addition or deletion of an indicator may alter the accuracy by which the attribute is measured, but not the nature of the attribute (latent variable) itself. With regard to the measurement of *g*, Spearman called this characteristic *indifference of the indicators* (Spearman, 1927, p. 197-198, as cited in Jensen, 1998). Thus, the indicators are exchangeable in the sense that an exchange possibly affects measurement properties such as precision, but not the meaning of the attribute of interest. In a reflective model, observables are indicators of a common theoretical attribute, in the

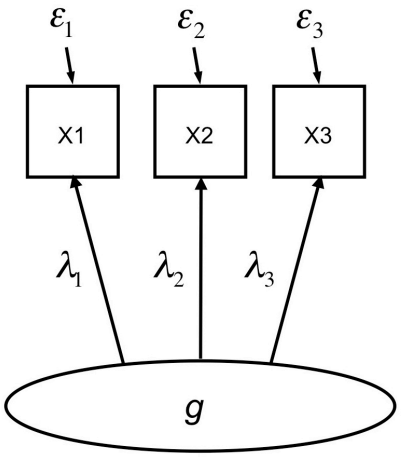

Figure 1. A reflective model of g. The latent attribute g is the underlying cause of the variability in the measured indicators. X's are observed variables, λ's factor loadings; ε's error terms.

same way that a set of differently constructed thermometers are indicators of a common attribute, namely temperature. Thus, it is assumed that the indicators measure the same thing. This implies that the latent variable or attributes exists independently of the model specification, at least with respect to the particular items used to measure it (Borsboom, Mellenbergh, & van Heerden, 2003). Of course, positing a reflective model does not guarantee the existence of purported latent variables: rather, the adoption of such a model generally carries with it a non-trivial ontological stance with regards to the latent variable.

Formative models

Formative models express the relationship between theoretical attributes and observations in terms of a regression function in which the theoretical attribute features as the dependent variable, and the observed variables as predictors. This is compatible with a conceptualization of the theoretical attribute (latent variable) as being in some way causally dependent on its indicators.

A common example of a formatively measured latent variable is socio-economic status (SES), where the SES-score for a given person is conceived of as a weighted sumscore of the measured variables, such as income and education level (Howell, Breivik, & Wilcox, 2007; Knesebeck, Lüschen, Cockerham, & Siegrist, 2003). Figure 2 depicts a path diagram of the formative model of SES. The three X indicators each contribute, with a certain weight, to the sumscore of the attribute SES. The X's in this example could be income, education, or other variables deemed relevant to the estimation of SES. The structure of the model is based on the idea that the indicators determine the latent attribute, rather than the other way around. With respect to SES, this seems to be a plausible model. For instance, you do not get a raise

because your SES level goes up; rather, your SES level goes up *because* you get a raise.

It is often argued that indicators in formative models should capture different aspects of the formative attribute, and should not be too strongly related (Bollen, 1984; Diamantopoulos & Siguaw, 2006). The latent attribute in such a model is represented as the weighted sum of different indicators that together predict a relevant phenomenon. An important theoretical characteristic of this model is that the latent attribute is defined by the choice of predictors. Thus, in contrast to the reflective model, a change of predictors implies a change in the nature of the attribute. In addition, in many circumstances the theoretical attribute is referentially unstable because the weights of the connections between the observations and latent variable are usually constructed to maximize the prediction of external criteria. That is to say, the value of the latent variable for a given person may change from one study to the next, if the predicted criterion changes (Bollen, 2002, 2007; Burt, 1976; Howell et al., 2007).

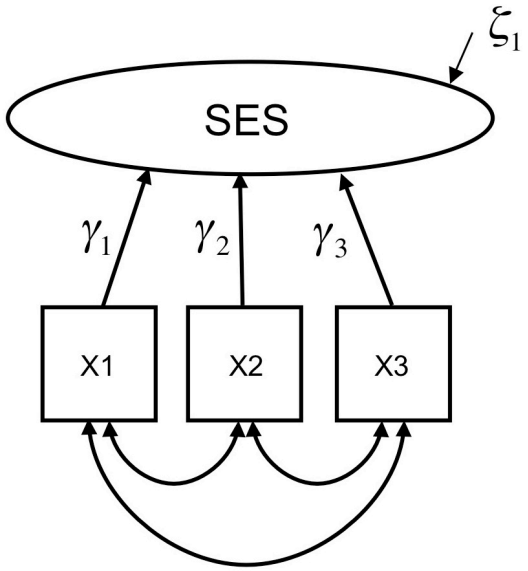

Figure 2. A formative model of SES (Socioeconomic Status). The attribute, SES, is determined by the measured indicators. The x's denote observed variables; zeta a residual term. Eta is a weighted sumscore; the weights denoted by lambda's.

Empirical testability of models

A crucial property of formative and reflective models is that they are testable, that is, they can be empirically corroborated or refuted, because the models impose restrictions on the joint probability distribution of the observations. Therefore, the support for a given specification of the underlying structure can be assessed by means of standard statistical tests and model fitting methods. Many fit indices have been developed for the evaluation of the fit of SEM models (Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003). Generally, fit indices are based on the discrepancy between the covariance structure implied by the specified model and the covariance structure, as observed in the data.

Commonly used fit-indices are the Chi-square for goodness-of-fit test, the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI). See Hu and Bentler (1999) and Schermelleh-Engel et al. (2003) for discussions of cutoff criteria for various fit indices for varying sample sizes and model complexity. A discussion of the details of model selection is beyond the scope of this paper. The main point is that such models can be fitted to empirical data and that this yields well-developed quantifications of the adequacy of the model. For detailed considerations of model specification and fitting procedures, an extensive and active area of psychometric literature focuses on the optimal manner in which to examine model fit and model selection (Howell, Breivik & Wilcox, 2007; Jarvis, Mackenzie, & Podsakoff, 2003; Myung & Pitt, 1997; Pitt, Myung, & Zhang, 2002, Waldorp, Grasman & Huizenga, 2006), parameter estimation (Diamantopoulos & Siguaw, 2006; Myung, 2006), stability over time (Hamaker, Nesselroade, & Molenaar, 2007; Van Buuren, 1997) and issues such as interpretational confounding (Bollen, 2007; Howell et al., 2007). Given that we have many tools to determine the (relative) adequacy of

our specified models, we now turn to the more relevant issue of how the theoretical positions discussed earlier may be mapped onto reflective and formative models.

Mapping of Psychometrics on Theory of Mind

We first examine identity theory, the theoretical position that at a given time, psychological and neurological properties of measurements reflect the same attribute. This implies that both P and N indicators have a common underlying cause, namely the true state of the latent variable. This is consistent with the reflective model, because that model views variability of the underlying attribute as the cause of variability in both P- and N-indicator values[3]. Therefore, when measuring brain activity and psychological behaviors related to a particular phenomenon such as intelligence, one is essentially *measuring the same thing.* Figure 3 shows how variation in the latent attribute (for example a subject's level of intelligence, or *g*) is the common cause of variation in both P-indicators (for example "giving the correct answer to a certain IQ-test question") and N-indicators (for example "increased activity in the dorsolateral prefrontal cortex").

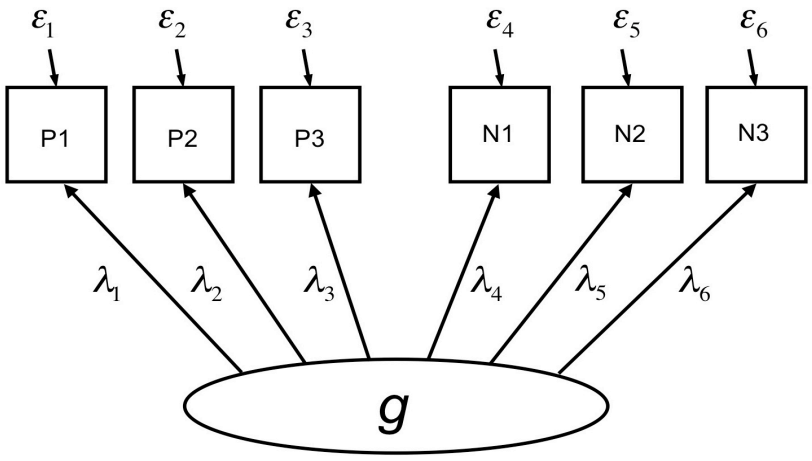

Figure 3. Reflective model of *g* and brain measurements. The latent attribute *g* is the underlying cause of variability in both the psychological and the neurological measurements. P1-P3 denote psychological measurements, N1-N3 denote neurological measurements.

If the reflective model of intelligence is correct, then the latent variable represents the actual value of *g*, which can be estimated in the same manner by both P and N indicators.

Therefore, P and N indicators can be said to be on equal empirical footing in that they are both assumed to be imperfect reflections of the true state of the underlying attribute. Identity theory is concordant with a realist perspective of psychological science, in the sense that it considers psychological attributes to be the underlying cause of variability of measurable indicators. The reflective model furnishes a psychometric implementation of identity theory: both the conceptual and the psychometric model assume a singular underlying cause that can be measured by two methods. The expected values of measurements within this model can be expressed as a function of the value of the latent attribute and the parameter that expresses the strength of the relationship between attribute and indicator. As such, it can be tested in the same way as psychometric models are usually tested. Thus, the reflective model can be used to *provide an empirical test of the identity hypothesis.*

The conceptual advantage of the reflective model is that it allows for a substantive interpretation of both classes of measurement by equating the psychometric status of neurological and psychological indicators. For example, some scientists argue that psychological concepts or processes are best measured by psychological measurements, while others maintain that neurological measures are more precise or insightful (e.g., the process or concept of consciousness, cf. Lamme, 2006). This dissension concerning the merits of neurological and psychological measurements in measuring a psychological attribute seems coherent only from an identity theoretical perspective. A debate on the relative merits of two methods of measurement requires that the object of measurement be the same. This allows one to gauge the relative

measurement precision of neurological and psychological indicators. At the same time, it allows for a comprehensible interpretation of both types of psychological research: A (non-neuroscientific) psychologist may acknowledge that corroborating evidence can be gained by the neurological approach (the same applies to the cognitive neuroscientist vis-à-vis psychometric data). Identity theory and reflective models view reductive psychological science as an integrated attempt to derive the best measure of the underlying attributes of interest. Such mutually insightful scientific interaction is in line with Heuristic Identity Theory (McCauley & Bechtel, 2001), which argues that simultaneous scientific study of two distinct explanatory levels from an identity theoretical perspective can be mutually beneficial.

Given its attractive theoretical properties, we conjecture that identity theory is implicitly assumed in most cognitive neuroscientific work. However, the conceptual benefits of this application of identity theory come at a price. For example, for both types of indicators to have the same underlying cause, the assumption of unidimensionality must be met. Unidimensionality has testable consequences such as local independence (Hambleton, Swaminathan, & Rogers, 1991) and vanishing tetrads (Bollen & Ting, 1993)[4]. If these tetrads are zero (by reasonable approximation), this is an indication that a unidimensional model may be appropriate, or at least, that it cannot be rejected. This suggests that the variability in both psychological and neurological indicators is attributable to a singular underlying cause. The criterion of unidimensionality is strict, and certainly need not be satisfied by purported behavioral and neurological measures of a given attribute. Thus, researchers should be clear on whether they believe that their neurological and psychological measurements are truly measuring the same attribute. To summarize, identity theory represents a strict theoretical and statistical position concerning the relationship between the two classes

of measurement. It posits that the variability found in both the P and N indicators has the same, unitary, underlying cause, and that the covariance between indicators can thus be fully explained by the underlying cause.

We now consider the integration of neurological and behavioral data from the perspective of supervenience theory. This theory is statistically less restrictive, conceptually distinct from identity theory, and may provide a more realistic alternative to the stringent requirements of identity theory. In a supervenience conceptualization of psychological processes, the higher order attributes are *realized in* their neurological properties. This is consistent with a specific implementation of the formative model, called the MIMIC (for Multiple Indicators, Multiple Causes) model (Jöreskog & Goldberger, 1975). To illustrate this, a path diagrammatic representation of the MIMIC model of *g* is displayed in Figure 4. In the MIMIC model, the variability of the determining indicators is a necessary but insufficient condition for variability at the level of the attribute. This is consistent with supervenience theory.

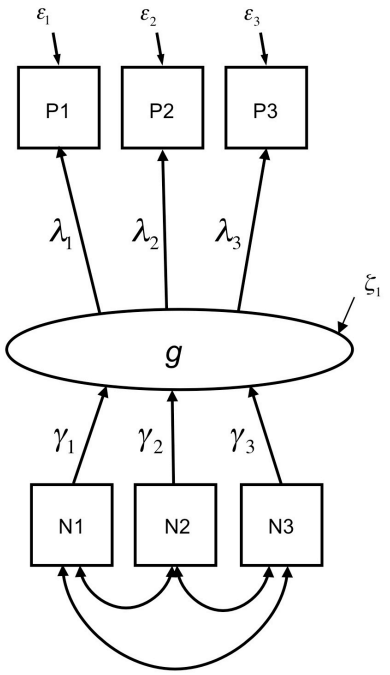

Figure 4. Formative (MIMIC) model of *g* and brain measurements. P1-P3 denote psychological measurements, N1-N3 denote neurological measurements. Psychological measurements are used to determine the parameter estimates of each of the *g* indicators. Variation in the N-indicators causally precedes variability in the latent attribute.

The essential aspect of this model is that there cannot be variation at the latent variable level if there is no variation in the indicators; therefore the theoretical attribute *supervenes* on its neurological constituents. Conversely, if two people have exactly the same lower order properties, that is, they have the same constellation of relevant neurological activation patterns, they necessarily have the same value on the attribute of interest. The restrictions and characteristics of the strong supervenience thesis and the formative model are identical in this sense. The insufficiency component implies that two people can have different indicator values but the same position at the latent attribute level. Therefore the position on the theoretical attribute is *multiply realizable*. Accordingly, the mapping of the observations to the theoretical attribute is many-to-one mapping, but no *isomorphism*, between the indicator values and the attribute value. Moreover, as is generally the case for supervenient properties (Kim, 1992), in the formative model any given position on the theoretical attribute corresponds to a *disjunction* of lower order properties. For example, a given level of SES, may correspond to either having a high salary and poor education, or having a low salary and high education, or having an average salary and average education, etc. Thus, the formative model is an instantiation of the supervenience hypothesis.

A formative approach seems a natural position to take in considering psychological effects of neurological deficits. Consider for example Korsakoff's syndrome. This condition is usually caused by alcohol abuse or malnutrition, which results in neuropathological symptoms, such as demyelination, neuronal loss, and small-scale hemorrhages (Kopelman, 1995). Psychological manifestations of Korsakoff's syndrome include impairment in the formation of new memories. In a reflective perspective on Korsakoff syndrome, the behavioral and neurological deficiencies would both be seen as *measurement of* the presence and severity of the

syndrome in a particular patient. This implies a causal direction that runs *from* the latent variable (a persons' value on a dimension representing the severity of Korsakoffs syndrome) *to* the neurological lesions. This seems counterintuitive. A more plausible conceptualization is provided by the formative, or MIMIC model. Under such a conceptualization, a person's Korsakoff "score" is *determined by* a weighted summation of the various lesions, by concurrently measuring and fitting a set of psychologically relevant predictors, such as memory tests. In this case, the lesions are the (partial) causes of Korsakoffs, not vice versa.

The theoretical status of the latent psychological attribute under supervenience theory is distinct from that under identity theory. A researcher who adheres to supervenience theory will represent the latent psychological attribute as being a formative attribute, i.e. as being *determined by* the constellation of neurological indicators. The relative influence of these neurological indicators is estimated on the basis of the predictive ability of the attribute in a network of psychologically relevant predictors.

The supervenience model, as displayed in Figure 4, has two components. The neurological indicators determine the latent psychological attribute. The parameter estimates, or the relative weights of the influence of the neurological measurements (Bollen, 2007), are estimated by predicting a psychologically relevant set of attributes or behaviors. The reflective component of a supervenience model is often required to be unidimensional. However, the formative part of the model is not so constrained: the indicators may even be uncorrelated (Bollen, 1984; Curtis & Jackson, 1962). This model is therefore less restrictive than a reflective, identity theoretical model. To summarize: An individual's position on a formative latent attribute, under the theory of supervenience, may estimated by fitting the model to a set of behaviorally predictive

psychological measurements. The identity of the attribute is determined by the neurological attributes included in the model that specifies the strength and direction of the neurological indicators. These indicators are assumed to determine variability in the latent attribute, which, in turn, determines variability at the psychological process.

The different empirical planes of the N-indicators and the P- indicators in a supervenience conceptualization, as opposed to identity theory, are important to neuroscience. The psychological indicators are scores derived from measurement instruments that are used to in the model specification. The parameter estimates, that relate variability in the latent attribute to variability on the N-indicators, depend on which P-indicators are chosen in the model. However, it is possible that the same set of N-indicators will fit models with different sets of P-indicators. Thus, the same N-indicators may realize different latent variables, as specified by different sets of P-indicators. This is a significant difference with the identity model, in which this is impossible.

This is important because it establishes that, given supervenience, the identification of attributes, even if they are neurologically grounded, depends on the psychological, not the neurological part of the model. This is consonant with the finding that certain neural structures are "implicated" in a wide range of different psychological concepts and processes. For example, the dorsolateral prefrontal cortex has been found to be differentially active in processes as psychologically diverse as response selection response selection (Hadland, Rushworth, Passingham, Jahanshahi, & Rothwell, 2001), pain modulation (Lorenz, Minoshima, & Casey, 2003), components of working memory (Ranganath, Johnson, & D'Esposito, 2003), voluntary willed action (Frith, Friston, Liddle, & Frackowiak, 1991), response inhibition (Ridderinkhof, Wildenberg, Segalowitz, & Carter, 2004), mastication (Takahashi,

Miyamoto, Terao, & Yokoyama, 2007), schizophrenia (Weinberger, Berman, & Zec, 1986) and intelligence (Jung & Haier, 2007). For an equally heterogeneous assessment of the functions of the anterior cingulate, see Vogt, Finch and Olson (1992) or Devinsky, Morrell and Vogt (1995). This functional heterogeneity should not be construed as a failure of cognitive neuroscience, but rather as an inherent property of brain function and organization. The point is that if certain cortical areas are associated with different cognitive functions, then it is unlikely that fMRI activity in such an area can be considered, for example, a "measurement of" working memory, as the assumption of unidimensionality will probably not be met.

We cannot think of an a priori reason to prefer either the identity or the supervenience model. Instead, we think that appropriateness of either model will depend on the attribute that is being studied and on theoretical considerations concerning that attribute. However, we note that precisely these theoretical considerations may be of great conceptual assistance to reductive psychological science, as they force researchers to consider the status of the attribute they are interested in, and the most appropriate manner to study it. Our argument here is that such choices are not esoteric statistical considerations: they concern *unavoidable assumptions implicit in any type of reductive research.* The goal of this approach is twofold: positions from philosophy of mind can be made empirical[5], and empirical neuroscience is provided with a method to get a grip on some of the more nebulous metaphors concerning the relation between psychological and neurological properties.

Doing so may yield several benefits, most notably the avoidance of ambiguous interpretations that may otherwise arise. If the issues mentioned above are not addressed explicitly, the questions being studied and the interpretations of the data may suffer. Consider, for instance, Jung and Haier (2007), who raised the question "Where

in the brain is intelligence?" (p. 135). Jung and Haier examined 37 methodologically heterogeneous studies that reported correlations between various measures of intelligence and the brain. Their model, called the PFIT (Parieto-Frontal Integration Theory) model, is built on the basis of what are, in the words of Norgate and Richardson (same paper, p. 162) "…correlations between those correlations", and describes what happens when an individual is involved in intelligent behavior (p. 138). Although the effort of combining insights from various studies is commendable, the conceptual ground for interpreting the correlations between intelligence and brain measures in this review is at times unclear, and findings are therefore hard to interpret.

Firstly, the question asked by Jung and Haier implies the possibility of the localization of intelligence. However, as intelligence is an inter-individual construct, this is akin to the question 'Where in the body is tallness?', a confusing question at best. Tallness is a property of the body; it does not reside in it. Similarly, intelligence is a property of the cognitive system, and does not reside in a particular part of the brain. Secondly, despite being based on inter-individual differences, the PFIT-model is in essence an *intra*-individual model of intelligent behavior. However, as Borsboom (2003) and Molenaar (2004) show, these two domains are quite distinct: results at the population level are not necessarily informative about the individuals that make up that population. The above illustrates that analyzing and interpreting relations between neurological and behavioral measurements can benefit from a sound conceptual basis. To illustrate how psychometric models may be able to provide more insight, we examine the application of the two previously discussed models to two empirical examples, focusing on neurological measurements with respect to personality characteristics and general intelligence.

Intelligence and brain volume

To illustrate the issues we discussed above, we return to the question we posed in the introduction: Can a measurement of the volume of a person's brain be considered a measurement of their intelligence? One of the more robust findings in the literature relating intelligence to physiological characteristics is the relationship between skull (or more recently brain) volume and estimates of general intelligence. Based on a meta-analysis, this correlation has been estimated at .33 (McDaniel, 2005). Given this relatively solid statistical association, can we consider measurements of brain volume to be measurements of intelligence, and therefore to conform to the identity theoretical perspective? That is, do measurements of brain properties and intellectual ability together fit a unidimensional reflective model?

Methods

To examine this question empirically, we consider behavioral measures (intelligence tests) and physiological (brain mass volume) measures. The sample consisted of physiological and behavioral data acquired from 80 healthy participants (21.1 years, sd = 2.55, 29 males, 51 males).The measures of intelligence are four domain scores of the commonly implemented Wechsler Adult Intelligence Scale (WAIS III). The domain score subscales used were Verbal Comprehension (M=117.16, SD=9.78), Perceptual Reasoning (M=112.10, SD=11.31), Working Memory (M=111.32, SD =13.11) and Processing Speed (M=116.38, SD=14.80). In addition to the behavioral measurements, all participants were scanned to estimate white matter, grey matter density and cerebrospinal fluid volume. Details of the scanning procedure and preprocessing steps are described in the appendix. To determine model fit, we examined the chi-square test of model fit, the Root Mean Square Error of Approximation (RMSEA, cut-off value 0.05), the comparative fit index, (CFI, cut-off value 0.95) the Akaike Information Criterion, or AIC (Akaike, 1974), and the Bayesian

Information Criterion, or BIC (Schwarz, 1978). For both models, the first reflective parameter was scaled to 1 to identify the reflective parameters. For discussions on the relative merits of these indicators, see Hu and Bentler (1999), or Schermelleh-Engel, Moosbrugger and Muller (2003).

We consider this experimental setup from the perspective of the two models that we discussed above. In fitting both models we use the same data, but impose distinct constraints consistent with the two models. In both models we view "intelligence" as an attribute that can be studied by psychological and physiological measurements, even though it cannot be observed directly. From the perspective of the reflective model, we consider both methods of measurement (i.e., VBM and the WAIS) as *measurements of* intelligence, in the same way that an electrical and a mercury thermometer may both measure temperature. This conceptualization has been represented previously in Figure 3: For this specific implementation, we would have 4 psychological measures and three neurological measures measuring the same property (*g*).

Conversely, one may view the neurological measurements as determining the latent psychological attribute. For instance, we may conjecture that the brain volume determines the level or degree of intelligence, in the same way that that we know that physiological damage can affect personality. In this case, we consider the MIMIC model to be appropriate. This is the model previously represented in Figure 4, in the MIMIC model of general intelligence and brain characteristics, the neurological indicators determine the value of the latent attribute (i.e., the *g* score). This in turn can be seen as the underlying cause of the variability of the scores at the WAIS level.

Model fit comparison

We used Mplus (Muthén & Muthén, 1998-2007) to fit the reflective and the formative (MIMIC) models for these seven indicators using maximum likelihood estimation. First, we examined the simple reflective model, in line with identity theory. The model was rejected by the chi-square test of model fit, χ^2 (14, N=80)=51.6, $p <$ .01. The other fit indices corroborate this poor fit (CFI=0.88, RMSEA=0.18, AIC=3706.39, BIC=3739.74). For this dataset therefore, Identity Theory is rejected, and we cannot consider measurements of brain volume to be measurements of intelligence. Next, we considered the MIMIC model, in line with supervenience theory. This model fits the data well. The model was not rejected by the chi-square test of model fit χ^2 (11, N=80)=11.20, $p >$.4. Other fit indices supported the good fit of the model (CFI=.996, RMSEA=0.015, AIC=3659.994, BIC=3686.196). Table 1 shows the parameter estimates for both models, which quantify the relative strength of the relationship between the indicators and the latent attribute "*g*".

Variable	Reflective model intelligence	MIMIC model intelligence
	Standardized factor loading	Standardized factor loading
WAIS1	0.273	0.601
WAIS2	0.236	0.714
WAIS3	0.301	0.534
WAIS4	0.246	0.5
Grey matter volume	0.983	0.883
White matter volume	0.967	-0.714
CSF	0.752	0.304

Table 1. Parameter estimates for Reflective and Formative (MIMIC) models of intelligence.

For this dataset therefore, a reflective model (identity theory) does not fit the data. The MIMIC (supervenience) model on the other hand fits the data quite well, and explains .25 of the variance in general intelligence, in line with previous analyses, clearly favoring this model for this dataset. However, the distinction in model fit will not be as clear-cut for all psychological constructs. Next, we will examine a dataset where the distinction is less pronounced.

Personality and the brain

Another type of construct traditionally of interest for scientific psychology is that of personality. One of the more famous models is the Big Five model of personality (McCrae & John, 1992), which describes variation in personality traits along five dimensions (Extraversion, Neuroticism, Conscientiousness, Openness and Agreeableness). Certain aspects of personality have been shown to correlate with differential brain activity and physiology (DeYoung & Gray, 2009; Wright et al., 2006). In fact, one of psychology's most famous case studies, i.e. the case of Phineas Gage, suggests that brain physiology may be of significance to researchers of personality (Damasio, Grabowski, Frank, Galaburda, & Damasio, 1994). We will examine the conceptual and statistical relationship between psychological data on a common personality subscale, conscientiousness, on the one hand, and a physiological measurements, in this case gray matter density, on the other hand.

Methods

In this study, physiological and behavioral data were acquired from 110 healthy participants (age M=21.4, SD=2.4, 27 males)[6]. The participants were tested on the abbreviated personality questionnaire NEO-PI (McCrae & Costa, 2004). This personality questionnaire comprises 60 items, with 12 items for every Big Five

personality dimension (i.e., extraversion, neuroticism, conscientiousness, openness and agreeableness). For the purpose of this illustration we focus on one subscale, conscientiousness. Additionally, we obtained of each subject two 3DT1 scans to study voxel based morphometry, or VBM. VBM is a voxel-wise comparison technique that uses high-resolution structural scans to estimate gray matter density values at the voxel level (Ashburner & Friston, 2000, 2001). Eight participants were excluded due to recording problems or the lack of a second scan, leaving 102 participants for subsequent analysis. We provide further preprocessing and scanning details in the appendix. As with the general intelligence data, we fit two models: a reflective model in line with identity theory, and a MIMIC model in line with supervenience theory.

Model fit comparison

We used Mplus (Muthén & Muthén, 1998-2007) to fit the reflective and the formative (MIMIC) models using maximum likelihood estimation. Using an iterative procedure that excluded parameters if model fit improved significantly by their removal, the final models included four brain regions (Left Supramarginal Gyrus, Right Middle Frontal Gyrus, Left Cerebellum, Right Cerebellum), and 11 of the original 12 conscientiousness questions.

First, we considered the reflective model, in line with identity theory. The reflective model was rejected by the chi-square test, χ^2 (90, N=105)= 120.49, $p < .05$. The other fit indices corroborated the poorer fit of the reflective model (RMSEA=0.06, CFI=.84, AIC=2129.21, BIC=2207.96). Secondly, we considered the MIMIC model. This was not rejected by the chi-square test of model fit, χ^2 (84, N=105)=100.65, $p > .10$. The other fit indices suggest reasonable fit (CFI=0.91), RMSEA (0.04), AIC (2101.37), BIC (2169.62). Table 2 shows the parameter estimates for both models,

which quantify the relative strength of the relationship between the indicators and the latent attribute "conscientiousness".

Variable	Reflective model	Formative (MIMIC) model
	Standardized factor loading	Standardized factor loading
C1	0.361	0.359
C2	0.231	0.23
C3	0.207	0.208
C4	0.336	0.337
C5	0.736	0.738
C6	0.397	0.398
C7	-0.204	-0.203
C8	0.313	0.315
C9	0.226	0.225
C10	0.802	0.797
C11	0.73	0.731
Left Supramarginal Gyrus	-0.3	-0.303
Right Middle Frontal Gyrus	0.29	0.311
Left Cerebellum	0.062	0.124
Right Cerebellum	-0.001	-0.062

Table 2. Standardized parameter estimates for Reflective and Formative (MIMIC) models of conscientiousness.

Because these two models are by their nature not nested, a chi-square test to compare them directly is not possible (Vuong & Wang, 1993). However, the formative (MIMIC) model shows better fit across the board than the unidimensional reflective model, with all fit indices outperforming those of the reflective model. Overall then, this suggests that the formative model provides a better fit to the data than the reflective model. The present study thus provides some support for a supervenience

interpretation of the relation between neurological and psychological variables with respect to conscientiousness.

Implications

As we show above, it is possible to fit such models to conventional neuroimaging data. There are several important aspects of the two illustrations. Firstly, the reflective, identity theoretical model was rejected for both datasets. Despite the adequate sample size and neurological variables known to correlate with the respective constructs, we cannot consider such measurements, for these datasets, to be measurements *of* the psychological constructs of interest. This points to an interesting conclusion that follows from the identity hypothesis: Researchers who view brain measurements as *measurements of* a latent psychological attribute (which may be plausible), must realize that this accords to brain measurements the same status as psychological measurements. Consequently, the brain measures may be rejected for the same reasons that poorly performing items in a questionnaire are rejected. This illustration shows what is required of neurological measurements if they are to figure as "measurements of" attributes in the same way that psychological measurements do. However, although the reflective model did not fit for the two examples we examined, this does not imply it will not fit for any dataset.

Firstly, the strength and nature of the relationship between psychological construct and neurological properties will vary depending on the construct, as it does in our two datasets. Given the variability of the psychological constructs that figure in scientific psychology, from early visual perception to complex dispositional constructs, it seems likely that the strength and nature of the relationship between neurological and behavioral measurements will be also different for such radically different behavioral phenomena. Secondly, we think it more likely that more restrictive models, such as the

reflective model, will fit for more 'basic', and less variable, psychological constructs and processes. For instance, whereas personality dimensions are at least partly culturally determined, other processes such as retinotopic mapping of early visual processing (mentioned in the introduction), depth perception and arousal may be easier to identity with unique, unidimensional neurological signatures. Such lower, more basic, less culturally dependent constructs, that display less variance across people, may be good candidates for identity theoretical models, although at this point this is largely speculation.

For both datasets, especially the brain volume data, the MIMIC model fit the data quite well. This implies that in these datasets, it is sensible to conclude that the neurological measurements statistically determine the variability in the psychological construct. Most importantly, the current findings show us that the relationship, especially for more complex psychological constructs such as intelligence and personality dimensions, is not likely to be simple. For this reason above all, we should be closely examining the nature of this relationship, and try to gain more insight by modeling hypotheses explicitly.

Summary

The results of our model fitting and speculations about other constructs brings to light an important aspect of the present reformulation of the reduction problem as a measurement problem: at the outset of any investigation, we should be *impartial* with respect to the status and quality of psychological and neuroscientific assessments *as measures.* For example, aspects of personality have been called "biologically based tendencies" (McCrae et al., 2000, p. 173). It remains to be seen whether certain empirical measurements behave in a way that allows for such an interpretation. Despite the popular view of neuroscientific measures as being "exact" or "hard", at least for

this dataset our analysis suggests that psychological measures may outperform neuroscientific measures. Insofar as such measures are interpreted as relevant to psychological attributes or processes, they should be evaluated on precisely the same basis as any other measure. This basis is psychometric in character. There is no way out of this issue, unless, perhaps, one comes up with an alternative to psychometrics, i.e., a practically workable theory of measurement that rests on a different basis. To the best of our knowledge, such a theory does not currently exist.

The above empirical illustrations serves as a proof of principle, in that it demonstrates that conceptual positions about the relationship between two classes of data can be constructed as statistical models and empirically tested. In this manner conceptual ideas about the relation between two levels of measurement can be translated into falsifiable models, and allow for theoretical interpretations of empirical data that can go beyond the simple observation that two measures are associated. In the next sections, we will discuss certain practical issues concerning SEM models, and the possibility of more exotic extensions.

Applying SEM in practice

Although structural equation models generally require larger sample sizes than more conventional analysis methods, this increase is by no means prohibitive. Sample sizes for SEM models are, as with other statistical analyses, related to model complexity. The models we discuss here are relatively simple, and sample sizes required are well within the reach of practicing neuroscientists. For instance, Marsh and Hau (1999) show that for models with 6 to 12 indicators per factor (as is the case for both our datasets), sample sizes of 50 may be adequate. Bentler (1995) recommends an N of at least 5 per free parameter, again within the limits of our empirical illustrations (cf. Schermelleh-Engel and Moosbrugger, 2003, for an

discussion on this topic). Although it is true that more simple or conventional data-analytic procedures such as ANOVA's and correlation tests yield results for very small sample sizes, over the long run, such analyses may represent pyrrhic victories over issues of inferential validity and replication.

Luckily, there are signs that the field of neuroimaging has increasingly moved towards sample sizes that are more than adequate for treatment with SEM. For instance, a quick inspection of the first 8 empirical papers in a recent issue of the journal Neuroimage (Volume 51, issue 1, the 'Anatomy and Physiology' section) that focus on structural anatomy (such as we examine in our paper) reveal sample sizes of 90, 55, 319, 185, 70, 40, 45 and 280 respectively, all of which would be amenable to SEM approaches given the guidelines above. At the same time, it is certainly true that lower sample sizes are a common occurrence in neuroimaging, especially in functional neuroimaging studies. However, to deal which the inherent complexity of the relationship between the brain and psychological constructs, more complex models, that require greater sample sizes, will need to be developed. The move towards larger, more versatile datasets may be part of a broader development in the field of cognitive neuroscience, taking inspiration from how neighboring fields deal with similar problems.

In the field of quantitative genetics, issues of replicability, power and interpretation have led to the realization that larger sample sizes are not a luxury but a necessity. This realization has lead to large collaborative projects such as the EAGLE and the GENEQOL consortium[7]. Such large-scale collaborative efforts combine the knowledge, resources and methodology from various research groups, can lead to increase collaboration and understanding, and therefore benefits the scientific community as a whole. It is such collaborations that would make the implementation

of more insightful models possible, and in our view would benefit the field as a whole. An additional advantage of the use of SEM in the context of collaborative projects is that there exist a statically and theoretically sound way to deal with group differences (i.e., measurement invariance; Meredith, 1993). A similar development in the fields of cognitive, affective and social neuroscience would be much welcomed.

To summarize, the sample sizes required to test conceptually guided SEM models are well within the reach of current empirical practice. To the extent that such datasets are not yet widely available, larger collaborations are desirable. Such collaborations are especially important if we want to tackle some of the most elusive and vexing phenomena: dynamic, reciprocal changes over time. In the next section, we will show how philosophy of mind and extensions of basic SEM models may help to get a grasp on such phenomena.

Top-down influences and temporal dynamics

The models that we have discussed represent two core philosophical positions, which have well-defined SEM counterparts. So far, we have focused on the most conventional method of analysis: the analysis of inter-individual differences in cross-sectional data. This method is dominant in contemporary psychological science. Although this method provided the basis of our proposed structuring of the relationship between behavioral and neurological data, other methodological approaches are possible. In fact, there are aspects of psychological and neurological phenomena that may be better studied by alternative means. In this section, we discuss some challenging problems for conceptual and statistical models in cognitive neuroscience. These concern dynamic, reciprocal changes of behavior and the brain structure and function through time. We note that SEM offers various possibilities to address these problems.

Conventional thought concerning the relationship between lower and higher order properties tends to consider (changes in) neurological properties as the source or cause of observable difference at the psychological/behavioral level. For instance, evidence shows that certain drugs influence cognitive abilities (Maylor & Rabbitt, 1993), that trauma may influence complex psychological traits, such as personality (Damasio, Grabowski, Frank, Galaburda, & Damasio, 1994), and that in Alzheimer patients amyloid peptide levels (constituents of amyloid plaques) and cortical activity are affected *prior* to observable cognitive symptoms (Buckner et al., 2005; Moonis et al., 2005). These findings all suggest that changes in cortical structure or functioning can, and do, affect psychological performance and functioning. However, there is also ample evidence for the reverse causal path. For instance, Maguire et al. (2000) showed that London taxi-drivers, following intensive training to learn the streets of London to the mandatory level of competence, showed structural changes to the hippocampus, and these changes were greater for taxi drivers who had served for a longer period of time. Also, intense juggling practice has been shown to affect both grey matter density (Draganski et al., 2004) and white matter integrity (Scholz, Klein, Behrens & Johansen-Berg, 2009). These findings suggest that persistent behavior may affect neurological structure and functioning. Finally, processes may even be *reciprocal* in nature, that is, simultaneous influences both from neurophysiological properties to behavior and vice versa. For instance, take the influence of hormone levels on psychology and behavior. Testosterone, when injected, can directly influence dominant or aggressive behavior, and is found to correlate positively with such behaviors (Mazur & Booth, 1998). However, Mazur and Booth illustrated that causal relationship may also be reversed: certain psychological behaviors may themselves lead to an increase in testosterone, and elevated testosterone levels affect behavior.

The above suggests that influences may run both from cognitive/psychological processes to neurological changes and vice versa. Modeling such dynamic interactions over time is a challenging problem, both conceptually and statistically. Philosophers have discussed such complex, dynamic systems in various terms, such as emergence, dynamic systems, and top-down causation. Emergence has a long philosophical tradition, going back to Mill and Broad in the late 19th and early 20th century (for an overview, see Kim, 1999). Recently, emergence and dynamic systems have enjoyed renewed interest as possible models for dynamic, neurocognitive changes through time. For instance, Jost, Bertschinger and Olbrich (2010) discussed the philosophical construct of emergence, and the description of a neuro-system as a (non-linear) dynamical reciprocal system. Similarly, Walmsley (2010) examined the concept of emergence, its relevance for complex systems, and possible manners in which law-like properties may emerge at higher (psychological) levels. Craver and Bechtel (2007) on the other hand focused on the concept of downward causation, and how this may be reconciled philosophically. They concluded that "When interlevel causes can be translated into mechanistically mediated effects, the posited relationship is intelligible and should raise no special philosophical objections" (p. 547). Furthermore, they stated that "There is a different sense in which a cause can be said to be at the top (or bottom) and a different sense in which its influence is propagated downward (or upward)" (p. 548).

Here we attempt to structure the distinction between such interlevel effects. Certain structural equation models offer the means to study such complex, interactive processes empirically. The origins of these models can be traced to the thirties and forties (e.g. Bartlett, 1946), but specific implementations in the behavioral/psychological sciences are relatively new. For instance, Hamaker,

Nesselroade, and Molenaar (2007) described a time series model in which two latent variables and their respective influences were modeled over time. We discuss how such a model may be used to model the time course of complex phenomena, such as discussed above. We distinguish two 'types' of situations: those in which psychological behavior affects, or at least precedes, variation in neurological properties (e.g. juggling), and those in which changes in neurophysiology affect (or precede), variation in psychological performance (e.g., Alzheimer's disease). In Figure 5, we present a model with two latent variables that evolve over time. Each latent variable has three observable indicators. The latent variable at the top represents a psychological construct, such as 'juggling ability' or 'cognitive ability', the latent variable at the bottom represents a neurological state of a person, such as 'neural density in motor cortex area x' or 'level of amyloid peptides'. We limit ourselves to just two latent variables for convenience; the model can be extended to include additional psychological or neurological latent variables, with varying numbers of indicators.

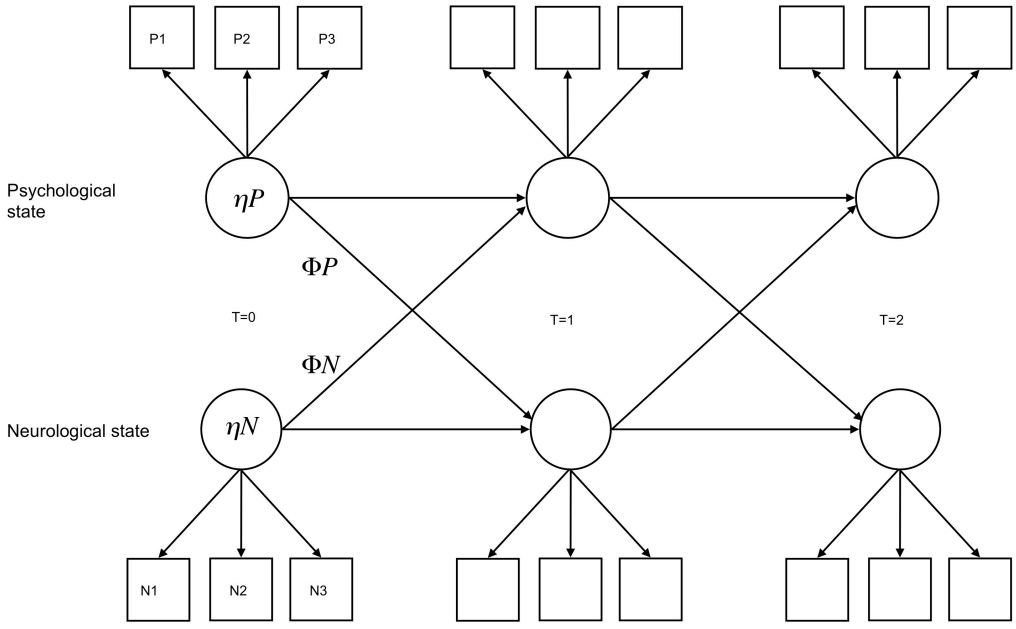

Figure 5. A possible representation of a time series model. Below are the latent variables of the brain states, that represent a persons neurological configuration at different time points. Above the traditional psychological latent variable, such as 'juggling ability' or 'reaction time'. Over time, it is possible to estimate the relative influences in both directions. Only the essential parameters are given.

If we measure the indicators of two (or more) latent variables in a given person repeatedly over time, we can relate the indicators to the latent variables, and we can model the time series of the latent variables. As mentioned above, we would like to differentiate between two scenarios: Those were variability in the psychological construct precedes neurological variability, and vice versa. If we were to assess a sample of people over time, either improving in juggling or deteriorating in cognitive performance, we could model this process by means of psychometric models such as the Integrated State-Trait model, described by Hamaker et al. (2007). To ease our presentation, we have included in Figure 5, only the most relevant parameters. The model parameter ϕP in Figure 5 represents the influence of the psychological latent variable at a given time point *on* the neurological latent variable at the next time point. This parameter should deviate from zero if changes in psychological abilities or behavior (e.g., practicing juggling) affect changes in neurological substrate (e.g., grey matter density in a motor cortex region). Conversely, the parameter ϕN reflects the influence of the neurological latent variable on the psychological latent variable. This parameter should deviate from zero if neural changes (e.g., amyloid peptide levels) affect changes at the psychological level (e.g., psychological performance).

So given appropriate time series measurements, we can test the hypothesis that the variation at the neurological precedes variation at the psychological level, or vice versa. In doing so, it is possible to empirically distinguish cases where influence should best be represented as 'bottom up' or 'top down'. Note that in this model we purposely estimate the relationship, and not assume it a priori: for this reason, a MIMIC model is not appropriate. Rather, we want to explore the time course of possibly reciprocal influences to gain insight into the nature of the underlying processes. Despite the

complex nature of such dynamic processes, SEM models allow researchers to, at least in principle, get a grip on the structure of the development over time and reciprocal interactions.

Discussion

The scientific future of reductive cognitive neuroscientific research rests both on advances in brain scanning technology and on the development of a comprehensive conceptual framework to link psychological-behavioral measures and neurological measures. To do so, a careful consideration of the status of neurological indicators in studies that measure both behavioral and neurological variables is required. We have shown a road forward in attacking this problem, by demonstrating that at least two theoretical stances on the reduction problem can be translated into well-understood formal psychometric models. To our knowledge, this is the first demonstration of how theoretical positions drawn from analytic philosophy can be translated to empirically testable models. Notably, our demonstration did not involve any rocket science; it merely used standard statistical models incorporated in widely available software packages. In this regard, the suggested models are ready for use, and there is little that stops the motivated researcher from utilizing their benefits.

Several of the issues raised by other authors we discussed in the introduction can be ameliorated, if not solved, by our proposed framework. Firstly, by explicitly framing the connection between neurological and behavioral variables as a measurement theoretical relationship, the mereological fallacy can be largely avoided. The models proposed above do not make claims about certain psychological processes being ‘in’ or ‘performed by’ a certain brain region any more than the answers of the questionnaire are the locus of personality traits. In fact, we would argue that the current

perspective offers a way to discuss brain-behavior relationships in a meaningful manner *without* running the risk of making mereological or neophrenological claims.

Furthermore, the correct application of SEM models (greatly) diminishes the problem of non-independence raised by Vul, Harris, Winkielman and Pashler (2009). The two-step procedure described in Vul et al. (2009) can be largely avoided by properly implementing formal measurement models. As the voxels are not treated as a large sets of independent statistical tests, but specified as part of a measurement model that implies certain covariance patterns, the multiple comparison issue is much less of a problem given the considerable size of datasets within neuroscience.

Finally, our approach can accommodate some of the ideas put forth by Feldman Barrett (2009). She argues that certain psychological processes may be more appropriately seen as a 'mix' (or 'recipe') of several classes or types of brain activity. That is, the categorical distinctions we make at the psychological level, e.g. between 'thinking', 'perceiving' and 'remembering', will probably not be found as categorically distinct processes or properties of the brain. For that reason, when studying neurological properties in relation to certain psychological properties, it may be more natural to think that different *combinations* of distributed activity in certain regions or systems can together be taken to represent distinct brain processes. According to Feldman Barrett, this more naturally accommodates the structure of the brain than old-fashioned perspectives such as positing a 'perception' region in contrast to a 'memory' region. Within the framework currently proposed, such hypotheses (that categorically distinct psychological concepts may be best seen as complex combinations of more basic processes) may be tested. This is best in line with the supervenience/MIMIC model. Given different psychological predictors (i.e. whether the reflective part of the model consists of personality items, Raven's matrices etc.), one would expect to find

different parameter estimates of the neurological measurements. The different weighting of neurological indicators is conceptually similar to the recipe metaphor proposed in Feldman Barrett. This underscores the flexibility of the current approach of implementing measurement models to test substantive hypotheses about brain-behavior relationships.

Hopefully, the current paper has served to convince the reader that the infamous reduction problem is at least partly a measurement problem. More specifically, one cannot hope to make true advances in solving the reduction problem without solving the associated measurement problems in parallel. This, we think, has substantial consequences for how we should evaluate reductionist claims, as well as what we can expect from reductionist research strategies. There are several reasons to pursue such a strategy.

There is a tendency, both in science and society, to view neuroscience as an exact area of research - closely related to physics, chemistry, and biology - while viewing psychology as a "soft" discipline (cf. Racine, Bar-Ilan, & Illes, 2005). However, the exact sciences are not exact because they use machines rather than questionnaires, but because they have successfully formalized theories. Such formalization is currently lacking at the interface of neuroscience and psychology. Thus, insofar as neuroscience has moved into the field of psychology, it has yet to earn the predicate of being a ‘hard’ science. Escape from this situation can only be realized by formalizing theories into mathematical models, which are likely to be statistical in nature; and insofar as these models concern measurement problems, they will likely be psychometric ones. For models to function properly, there should be no psychometric prejudice as to the quality of the measurements: from a measurement perspective,

neurological measurements do not have a privileged position over conventional psychological measurements.

To the researcher with expertise in the intricacies of psychometric modeling, the example illustration in this paper may be viewed as quite optimistic, and such an evaluation would not be entirely off the mark. For even though we think it is evident that psychometrics has much to offer to neuroscience, it should be noted that psychometric modeling can be quite complicated. For instance, as discussed before, successful modeling generally requires (slightly) larger sample sizes or extensive time series, attention to possible problems involving model identification and model equivalence (e.g., see Raykov & Penev, 1999), goodness-of-fit, and other general issues common to statistical modeling. However, we think that such issues, in general, do not pose greater problems for structural equation models than for other techniques, and should not detract from substantively guided model implementation. There really is no way around these problems; in particular, these issues will not be resolved by being ignored, and by proceeding as if one did not have a measurement problem to solve.

The models we have discussed in the present work are illustrative of how clean and simple identity and supervenience theories really are. As a result, it is likely that the models that we applied to personality measures may be too simplistic. This, however, is a benefit rather than a shortcoming of the psychometric representation of reductive theories: a psychometric representation makes the hypotheses proposed transparent and subject to informed criticism, and it does this to a degree that no verbal description could match. Moreover, *rejecting* these models brings with it the task of *inventing better ones*. And this, we think, is precisely the road to progress. In addition, it is likely that alternative models will lead to alternative philosophical views on the

relation between psychology and neuroscience. We have provided a proof of principle by fitting two models with a single latent variable measured at the inter-individual level. However, this approach is certainly not limited to such a design; one of the great benefits of structural equation models is their flexibility. Most psychological processes involve a complex interplay of more than one attribute. For example, complex cognitive processes such as problem solving almost surely involve the interaction of separate subsystems such as working memory, attention and intelligence. Structural equation models can be extended to include multiple latent variables, thereby testing hypotheses about the interactive, inhibitory or excitatory activity of several latent variables of psychological attributes within a measurement model. The theoretical approach discussed here is especially suited for the implementation of flexible models that may address a range of questions of substantive interest to both cognitive neuroscientists and philosophers.

Finally, there are at least two types of homogeneity within cognitive neuroscience that are often assumed rather than tested. For instance, one of the vexing and largely neglected issues within psychological science is the distinction between inter- and intraindividual explanation. This means that a result found at the group level is often taken to apply to the individual, despite the fact that this may not be true and is rarely tested (cf. Borsboom, Mellenbergh & Van Heerden, 2003; Molenaar, 2004; Molenaar, Huizenga & Nesselroade, 2002). This issue holds as much for cognitive neuroscience as it does for conventional psychological science. As described previously, the current approach can be extended to explicitly test the structure of intraindividual activation. For example, the time series model by Hamaker, Nesselroade and Molenaar (2007) showed how an intraindividual process can be modeled with repeated measurements of the same latent variable. This can be done in

much the same way within the current framework by including dynamic intraindividual measurements such as EEG or fMRI. In this manner, one can study the extent to which a latent variable at the inter-individual level is representative for the individuals that make up a population, by assessing the homogeneity of the latent variable at both levels. Inter-individual variability is commonly treated as measurement error, but by explicitly testing the tenability of this assumption, a more fine-grained understanding of psychological attributes may be possible. In fact, the extent to which this holds for certain psychological attributes but not for others is likely to yield valuable insights. Another largely neglected but potentially insightful area of cognitive neuroscience is the question of homogeneity across subpopulations. Within the current framework, it is relatively easy to test whether a latent variable representation of, say, working memory differs across age groups, gender, or other subpopulations (e.g., see Meredith, 1993). For example, Henrich, Hein and Norenzayan (2010) examined to what extent it is possible to generalize from the most commonly studied psychological subpopulation, namely young, white, highly educated people from industrialized nations, to other cultures and demographics. They showed that, even for the most 'basic' of cognitive phenomena such as the Mueller-Lyer illusion, such untested generalization is often unjustified. The assumption of generalization and homogeneity is, arguably, even more omnipresent within cognitive neuroscience than in conventional psychology. We would venture that the extent to which neuroscientific findings generalize across populations and cultures is an open empirical question, and that its premature acceptance may close off a considerable amount of potentially insightful empirical investigations.

This paper has served to illustrate both the necessity and the potential for conceptual and empirical progress that may be achieved by considering an integrated

psychometric perspective on reductive cognitive neuroscience. We have offered the conceptual and technical tools to do so, and hope that our efforts will be built on by others. The relationship between mind and body has fascinated generations of philosophers and scientists, and deserves closer methodological and psychometric scrutiny than it has so far enjoyed. If theories developed in the philosophy of mind are to escape from their current state of splendid metaphysical isolation, it is essential to translate these positions to empirical predictions. With recent advances in neuroscientific and psychometric techniques and methods, we finally have the opportunity to empirically address questions that were once restricted to the realm of speculative metaphysics. It would be a waste to forgo such opportunities.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*, 716-723.

Andersson, J.L.R., Jenkinson M., Smith, S.M. (2007). Non-linear registration, aka spatial normalisation. *FMRIB technical report* TR07JA2. Retrieved from: http://www.fmrib.ox.ac.uk/analysis/techrep/tr07ja2/tr07ja2.pdf

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry-The methods. *NeuroImage, 11*, 805-821.

Ashburner, J., & Friston, K. J. (2001). Why voxel-based morphometry should be used. *NeuroImage, 14*, 1238-1243.

Bagozzi, R. P. (2007). On the meaning of formative measurement and how it differs from reflective measurement: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods, 12*, 229-237.

Bartlett, M.S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society, 8,* 27-41.

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Malden, MA: Blackwell Publishers.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bickle, J. (1998). Psychoneural reduction: the new wave. Cambridge, MA: MIT Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bollen, K. A. (1984). Multiple indicators: internal consistency or no necessary relationship? *Quality and Quantity, 18*, 377-385.

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

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605-634.

Bollen, K. A. (2007). Interpretational confounding is due to misspecification, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods, 12*, 219-228.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305-314.

Bollen, K. A., & Ting, K. (1993). Confirmatory tetrad analysis. *Sociological Methodology, 23*, 147-175.

Borsboom, D. (2008). Latent variable theory. *Measurement, 6*, 25–53.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203-219.

Buckner, R.L., Snyder, A. Z., Shannon, B. J., LaRossa, G., Sachs, R., Fotenos, R. F., Sheline, Y. I., Klunk, W.E., Mathis, C.A., Morris, J.C., & Mintun, M. A. (2005). Molecular, structural, and functional characterization of Alzheimer's disease: Evidence for a relationship between default activity, amyloid, and memory. *The Journal of Neuroscience, 25,* 7709 –7717.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research, 5*, 3-52.

Canli, T., Zhao, Z., Desmond, J. E., Kang, E., Gross, J., & Gabrieli, J. D. E. (2001). An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience, 115*, 33-42.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78*, 67-90.

Churchland, P. (1985). Reduction, qualia, and the direct introspection of brain states. *The Journal of Philosophy, 82*, 8-28.

Collier, J. (1988). Supervenience and reduction in biological hierarchies. *Canadian Journal of Philosophy, 14*, 209-234.

Craver, C. F. & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22,* 547–563.

Curtis, R. F., & Jackson, E. F. (1962). Multiple indicators in survey research. *The American Journal of Sociology, 68*, 195-204.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damasio, A. R. (1994). The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science, 264*, 1102-1105.

Davidson, D. (1980). Mental events. In N. Block (Ed.), *Readings in philosophy of psychology* (pp. 107-119). London: Methuen.

Davidson, R. J. (2004). What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. *Biological Psychology, 67*, 219-233.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*, 71-100.

Demetriou, A., & Mouyi, A. (2007). A roadmap for integrating the brain with mind maps. *Behavioral and Brain Sciences, 30*, 156-158.

Devinsky, O., Morrell, M. J., & Vogt, B. A. (1995). Review article: Contributions of anterior cingulate cortex to behaviour. *Brain, 118*, 279.

DeYoung, C. G., & Gray, J. R. (2009). Personality neuroscience: Explaining individual differences in affect, behavior, and cognition. In P. J. Corr & G. Matthews

(Eds.), *Cambridge Handbook of Personality* (pp. 323-346). New York: Cambridge University Press.

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management, 17*, 263-282.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Neuroplasticity: Changes in gray matter induced by training. Nature, 427, 311-312.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*, 155-174.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

Feldman Barrett, L. (2009). The future of psychology: Connecting mind to brain. *Perspectives on Psychological Science, 4,* 326-339.

Fodor, J. A. (1974). Special sciences (Or the disunity of science as a working hypothesis). *Synthese, 28*, 97-115.

Frith, C. D., Friston, K., Liddle, P. F., & Frackowiak, R. S. J. (1991). Willed action and the prefrontal cortex in man: a study with PET. *Proceedings: Biological Sciences*, 241-246.

Glymour, C. (1998). What went wrong? Reflections on science by observation and The Bell Curve. *Philosophy of Science*, *65,* 1-32.

Gold, I., & Stoljar, D. (1999). A neuron doctrine in the philosophy of neuroscience. *Behavioral and Brain Sciences, 22*, 809-869.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron, 35*, 1167–1175.

Hadjikhani, N., Liu, A. K., Dale, A. M., Cavanagh, P., & Tootell, R. B. H. (1998). Retinotopy and color sensitivity in human visual cortical area V8. *Nature Neuroscience, 1*, 235-241.

Hadland, K. A., Rushworth, M. F. S., Passingham, R. E., Jahanshahi, M., & Rothwell, J. C. (2001). Interference with performance of a response selection task that has no working memory component: An rTMS comparison of the dorsolateral prefrontal and medial frontal cortex. *Journal of Cognitive Neuroscience, 13*, 1097-1108.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated trait-state model. *Journal of Research in Personality, 41*, 295-315.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hare, R. M. (1952). *The language of morals*. Oxford: Oxford University Press.

Henrich, J., Hein, S.J. & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences, 33,* 61-83.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81,* 945-960.

Horgan, T. (1993). From supervenience to superdupervenience: Meeting the demands of a material world. *Mind, 102*, 555-586.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12*, 205-218.

Howell, R. J. (2009). Emergentism and supervenience physicalism. *Australasian Journal of Philosophy, 87*, 83-98.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hubel, D.H. & Wiesel, T.N. (1968.) Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology, 195, 215-243

Jarvis, C. B., Mackenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*, 199-218.

Jensen, A. R. (1998). The g factor: The science of mental ability. Westport, CT: Praeger/Greenwood.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Jöreskog, K., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*, 631-639.

Jöreskog, K., & Sörbom, D. (1996). *LISREL8: User's Reference Guide*. Chicago: Scientific Software International.

Jost, J., Bertschinger, N. & Olbrich, E. (2010). Emergence. *New Ideas in Psychology, 28,* 265-273.

Jung, R. E., & Haier, R. J. (2007). The parieto-frontal integration theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences, 30*, 135-187.

Kim, J. (1982). Psychophysical supervenience. *Philosophical Studies, 41*, 51-70.

Kim, J. (1984). Concepts of supervenience. *Philosophy and Phenomenological Research, 45*, 153-176.

Kim, J. (1985). Supervenience, determination, and reduction. *The Journal of Philosophy, 82*, 616-618.

Kim, J. (1987). "Strong" and "global" supervenience revisited. *Philosophy and Phenomenological Research, 48*, 315-326.

Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research, 52*, 1-26.

Kim, J. (1999). Making sense of emergence. *Philosophical Studies, 95,* 3–36.

Knesebeck, O. v. d., Lüschen, G., Cockerham, W. C., & Siegrist, J. (2003). Socioeconomic status and health among the aged in the United States and Germany: A comparative cross-sectional study. *Social Science & Medicine, 57*, 1643–1652.

Kopelman, M. D. (1995). The Korsakoff syndrome. *The British Journal of Psychiatry, 166*, 154-173.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*, 494-501.

Lawley, D. N., & Maxwell, A. E. (1963). Factor analysis as a statistical method. London: Butterworth.

Levine, R. A. (1999). Somatic (Craniocervical) tinnitus and the dorsal cochlear nucleus hypothesis. *American Journal of Otolaryngology, 20*, 351-362.

Lewin, K. (1951). *Field Theory in Social Science; selected theoretical papers*. New York: Harper & Row.

Lewis, D. K. (1966). An argument for the identity theory. *The Journal of Philosophy, 63*, 17-25.

Lorenz, J., Minoshima, S., & Casey, K. L. (2003). Keeping pain out of mind: the role of the dorsolateral prefrontal cortex in pain modulation. *Brain, 126*, 1079-1091.

Luna, B., Thulborn, K. R., Strojwas, M. H., McCurtain, B. J., Berman, R. A., Genovese, C. R., et al. (1998). Dorsal cortical regions subserving visually guided saccades in humans: an fMRI study. *Cerebral Cortex, 8*, 40-47.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67,* 1-25.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, *2,* 245-264.

Maguire, E.A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J. & Frith, C.D. (2000). Navigation-Related Structural Change in the Hippocampi of Taxi Drivers. *Proceedings of the National Academy of Sciences, 97,* 4398-4403.

Marsh, H. W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical strategies for small sample size* (pp. 251–306). Thousand Oaks, CA: Sage.

Maylor, E. A., & Rabbitt, P. M. (1993). Alcohol, reaction time and memory: A meta-analysis. British Journal of Psychology, 84, 301–317.

Mazur, A. & Booth, A. (1998). Testosterone and dominance in men. *Behavioral and Brain Sciences, 21,* 353-397.

McCauley, R. N., & Bechtel, W. (2001). Explanatory pluralism and heuristic identity theory. *Theory and Psychology, 11*, 736-760.

McCrae, R. R., & Costa Jr, P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences, 36*, 587-596.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.

McCrae, R. R., Costa Jr, P. T., Hrebícková, M., Ostendorr, F., Angleitner, A., Avia, M. D., et al. (2000). Nature over nurture: Temperament, personality, and life span development. *Nature, 78*, 173-186.

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33*, 337-346.

McGregor, I. (2006). Offensive Defensiveness: Toward an Integrative Neuroscience of Compensatory Zeal After Mortality Salience, Personal Uncertainty, and Other Poignant Self-Threats. *Psychological Inquiry, 17,* 299-308.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

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

Miyawaki, Y, Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H.C., Sadato, N. & Kamitani, Y. (2008). Visual image reconstruction from human brain Activity using a combination of multiscale local image decoders. *Neuron, 60,* 915-929.

Mobbs, D., Hagan, C. C., Azim, E., Menon, V., & Reiss, A. L. (2005). Personality predicts activity in reward and emotional regions associated with humor. *Proceedings of the National Academy of Sciences, 102*, 16502-16506.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*, 201-218.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2002). The relationship between the structure of inter-individual and intraindividual variability: A

theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development* (339-360). Dordrecht, the Netherlands: Kluwer.

Moonis, M., Swearer, J. M., Dayaw, M. P. E., George-Hyslop, P. St., Rogaeva, E., Kawarai, T. & Pollen, D. A. (2005). Familial Alzheimer disease: Decreases in CSF Aß42 decline levels precede cognitive decline. *Neurology, 65,* 323-325.

Morris, M.W. & Mason, M.F. (2009). Intentionality in Intuitive Versus Analytic Processing: Insights From Social Cognitive Neuroscience. *Psychological Inquiry, 20,* 58-65.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus User's Guide* (Fifth ed.). Los Angeles, CA: Muthén & Muthén.

Myung, I.J. (2006). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47,* 90-100.

Myung, I.J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review, 4*, 79-95.

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation.* London: Routledge & Kegan Paul.

Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. In: *Minnesota Studies in the Philosophy of Science*, 3-36.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting models of Cognition. *Psychological Review, 109,* 472-491.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology, 47*, 44-50.

Posthuma, D., De Geus, E. J. C., Baaré, W. F. C., Pol, H. E. H., Kahn, R. S., & Boomsma, D. I. (2002). The association between brain volume and intelligence is of genetic origin. *Nature Neuroscience, 5*, 83-84.

Psillos, S. (2004). A glimpse of the secret connexion: Harmonizing mechanisms with counterfactuals. *Perspectives on Science, 12*, 288-319.

Putnam, H. (1973). Reductionism and the nature of psychology. *Cognition, 2*, 131-146.

Putnam, H. (1980). The nature of mental states. In N. Block (Ed.), *Readings in philosophy of psychology* (pp. 223-236). London: Methuen.

Quine, W.V.O. (1969). Epistemology Naturalized. In: S. Clough (Ed.), *Siblings under the skin: feminism, social justice, and analytic philosophy* (pp. 66-84). Aurora, CO: Davies Group.

Racine, E., Bar-Ilan, O., & Illes, J. (2005). fMRI in the public eye. *Nature Reviews Neuroscience, 6,* 159–164.

Ranganath, C., Johnson, M. K., & D'Esposito, M. (2003). Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia, 41*, 378-389.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.

Raykov, T., & Penev, S. (1999). On structural model equivalence. *Multivariate Behavioral Research, 34,* 199–244.

Ridderinkhof, K. R., Wildenberg, W. P. M. v. d., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition, 56*, 129-140.

Ross, D., & Spurrett, D. (2004). What to say to a skeptical metaphysician: A defense manual for cognitive and behavioral scientists. *Behavioral and Brain Sciences, 27*, 603-647.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961-962.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., & Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging, 18*, 712-721.

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,* 23 – 74.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6,* 461-464.

Scholz, J, Klein, M.C., Behrens, T.E. & Johansen-Berg, H. (2009). Training induces changes in white-matter architecture. Nature Neuroscience, 12, 1370–1371.

Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review, 68*, 141-156.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping, 17,* 143-155.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23*, 208-219.

Takahashi, T., Miyamoto, T., Terao, A., & Yokoyama, A. (2007). Cerebral activation related to the control of mastication during changes in food hardness. *Neuroscience, 145*, 791-794.

Tootell, R.B.H., Switkes, E., Silverman, M.S. & Hamilton, S.L. Funational anatomy of the macaque striate cortex: II. Retinotopic organization. *The Journal of Neuroscience, 8,* 1531-1568.

Van Buuren, S. (1997). Fitting ARMA time series by structural equation models. *Psychometrika, 62*, 215-236.

Vogt, B. A., Finch, D. M., & Olson, C. R. (1992). Functional heterogeneity in cingulate cortex: The anterior executive and posterior evaluative regions. *Cerebral Cortex, 2*, 435-443.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4,* 274-290.

Vuong, Q. H., & Wang, W. (1993). Minimum chi-square estimation and tests for model selection. *Journal of Econometrics, 56*, 141-168.

Waldorp, L.J., Grasman, R.P.P.P., & Huizenga, H.M. (2006). Goodness-of-fit and confidence intervals of approximate models. *Journal of Mathematical Psychology, 50,* 203-213.

Walmsley, J. (2010). Emergence and reduction in dynamical cognitive science. New Ideas in Psychology, 28, 274-282.

Wechsler, D. (2005). *WAIS-III Nederlandstalige bewerking.* Amsterdam: Harcourt Assesment.

Weinberger, D. R., Berman, K. F., & Zec, R. F. (1986). Physiologic dysfunction of dorsolateral prefrontal cortex in schizophrenia. I. Regional cerebral blood flow evidence. *Archives of General Psychiatry, 43*, 114-124.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science, 69*, 366-377.

Wright, C. I., Williams, D., Feczko, E., Barrett, L. F., Dickerson, B. C., Schwartz, C. E., et al. (2006). Neuroanatomical correlates of extraversion and neuroticism. *Cerebral Cortex, 16*, 1809-1819.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging, 20*, 45-57.

Appendix

Imaging and preprocessing

Participants were scanned on a 3-T Philips Intera scanner, and all data were analyzed using FSL (Smith et al., 2004), Matlab (Mathworks Inc.) and Mplus (Muthén & Muthén, 1998-2007). A structural MRI scan of each participants was acquired using a T1-weighted 3D sequence (Turbo Field Echo, TE 4.6 ms, TR 9.6 ms, FA 8°, 182 sagittal slices of 1.2 mm, FOV 250^2 mm, reconstruction matrix 256^2).

For the study on intelligence we first extracted the brains from the structural images (Smith, 2002) and subsequently segmented the white and gray matter and CSF using FAST4 (Zhang, 2001). The resulting volume counts on these compartments were directly used for the analysis.

For the study on personality we performed voxel based morphometry (VBM) carried out with FSL (Smith et al., 2004). For this study the structural images were brain-extracted (Smith, 2002). Next, tissue-type segmentation was carried out using FAST4 (Zhang, 2001). The so obtained gray-matter partial volumes were then aligned to MNI152 standard space using the affine registration. The resulting images were averaged to create a study-specific template, to which the native gray matter images were then non-linearly re-registered with a method that uses a b-spline representation of the registration warp field (Andersson, Jenkins & Smith, 2007, Rueckert et al. 1999). The registered partial volume images were modulated (to correct for local expansion or contraction) by dividing by the Jacobian of the warp field. The modulated segmentated images were smoothed with an isotropic Gaussian kernel with a sigma of 4 mm. The above procedure was applied to the first and second T1 scans separately, creating to independent datasets. The dataset was used to identify regions of interest that explained variance in the overall NEO-PI questionnaire (f-test over the demeaned

5 factors) using voxelwise permutation-based non-parametric testing. From this we obtained 12 regions of interest (ROI) that we used to extract values in these ROIs from the second (independent) dataset. ROIs were extracted if at least 200 connected voxels surpassed a threshold of $p<0.01$).

Author Note

This manuscript is original. None of these materials have been published elsewhere, nor is it under consideration for publication in any other journal. The participants were tested in accordance with the ethical guidelines of the American Psychological Association and this research was approved by the University of Amsterdam Ethical Committee.

Correspondence should be addressed to:

Rogier A. Kievit

University of Amsterdam, Department of Psychological Methods

Roetersstraat 15

1018WB Amsterdam, The Netherlands

(+31) 020-5256688

E-mail: r.a.kievit@uva.nl

Authors URL's

Rogier A. Kievit
R.A.Kievit@uva.nl
www.rogierkievit.com

Jan-Willem Romeijn
j.w.romeijn@rug.nl
http://www.philos.rug.nl/~romeyn/index.html

Lourens Waldorp
L.J.Waldorp@uva.nl
http://www.waldorp.nl/lourens/index.htm

H. Steven Scholte
H.S.Scholte@uva.nl
http://home.medewerker.uva.nl/h.s.scholte/

Denny Borsboom
D.Borsboom@uva.nl
http://borsboomdenny.googlepages.com/dennyborsboom

Footnotes

[1] We refer to the discpline here as *cognitive* neuroscience as it is the broadest and most common name for the concurrent study of psychological behavior and physiological properties. However, we do not aim to restrict our perspective to merely *cognitive* phenomena such as attention, memory or intelligence: The issues we raise are equally of interest for fields such as social neuroscience or affective neuroscience. Wherever we state cognitive neuroscience, we mean to encompass such more specific branches.

[2]A similar position can be found in Davidson (1980, p. 111)

[3]Given the exact formulation as a SEM, one should construe this to mean that variability in the underlying attribute causes variability in both the P- and the N-indicators.

[4]A tetrad is the difference of the products of the covariances of four measured indicators.

[5] The idea that epistemological speculation can gradually be replaced by empirical science, sometimes termed naturalism or naturalized epistemology, has a long history. See for instance Quine (1969) for a philosophical motivation.

[6]80 of the participants in the personality dataset were also analyzed in the intelligence dataset, albeit on different behavioral and neurological measurements.

[7]http://wiki.genepi.org.au/display/EAGLE/EAGLE and http://ideas.repec.org/p/rsw/rswwps/rswwps47.html