DIPFdocs
Institutionelles Open Access Repositorium

DIPF
Leibniz-Institut für Bildungsforschung
und Bildungsinformation

Voelkle, Manuel C.; Brose, Annette; Schmiedek, Florian; Lindenberger, Ulman

Toward a unified framework for the study of between-person and within-person structures. Building a bridge between two research paradigms

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Multivariate behavioral research 49 (2014) 3, S. 193-213, 10.1080/00273171.2014.889593

Bitte verwenden Sie in der Quellenangabe folgende URN oder DOI /
Please use the following URN or DOI for reference:
urn:nbn:de:0111-dipfdocs-178367
10.25657/02:17836

https://nbn-resolving.org/urn:nbn:de:0111-dipfdocs-178367
https://doi.org/10.25657/02:17836

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

DIPF | Leibniz-Institut für
Bildungsforschung und Bildungsinformation
Frankfurter Forschungsbibliothek
publikationen@dipf.de
www.dipfdocs.de

Mitglied der Leibniz-Gemeinschaft

This is an Accepted Manuscript of an article published by Taylor & Francis in Multivariate Behavioral Research on 02/06/2014, available online: http://www.tandfonline.com/10.1080/00273171.2014.889593.

Running head: BETWEEN AND WITHIN-PERSON STRUCTURAL EQUIVALENCE

Towards a Unified Framework for the Study of Between-Person and Within-Person Structures: Building a Bridge Between Two Research Paradigms

Manuel C. Voelkle[1]

Annette Brose[1]

Florian Schmiedek[1,2]

Ulman Lindenberger[1]

[1]Max Planck Institute for Human Development, Berlin, Germany

[2]German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

ACCEPTED FOR PUBLICATION IN

MULTIVARIATE BEHAVIORAL RESEARCH

Word count (exc. references, figures & tables): 11,118

*Requests for reprints should be addressed to Manuel C. Voelkle, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Email: voelkle@mpib-berlin.mpg.de, Voice: (+49)-30-82406-467

Author Note

Manuel C. Voelkle, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Email: voelkle@mpib-berlin.mpg.de, Voice: (+49)-30-82406-467

Annette Brose, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Email: brose@mpib-berlin.mpg.de, Voice: (+49)-30-82406-367

Florian Schmiedek, German Institute for International Educational Research (DIPF), Schlossstrasse 29, 60486, Frankfurt am Main, Germany. Email: schmiedek@dipf.de, Voice: (+49)-69-24708-820

Ulman Lindenberger, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Email: lindenberger@mpib-berlin.mpg.de, Voice: (+49)-30-82406-572

Acknowledgments

We thank Han Oud for his helpful comments on a previous draft of this paper, Kenneth Bollen for interesting discussions on the topic and for pointing us to earlier work in the econometric and sociological literature, Janek Berger and Michael Krause for their help in preparing the animated graphics, and Julia Delius for editorial assistance.

The COGITO Study was supported by the Max Planck Society, including a grant from Max Planck Society's innovation fund (M.FE.A.BILD0005); the Alexander von Humboldt Foundation's Sofja Kovalevskaja Award (to Martin Lövdén) donated by the German Federal Ministry for Education and Research (BMBF); the German Research Foundation (DFG; KFG 163); the Gottfried Wilhelm Leibniz award of the German Research Foundation (to Ulman Lindenberger) and the BMBF (CAI).

Abstract

The vast majority of empirical research in the behavioral sciences is based on the analysis of between-person variation. In contrast, much of applied psychology is concerned with the analysis of variation within individuals. Furthermore, the mechanisms specified by psychological theories generally operate within, rather than across, individuals. This disconnect between research practice, applied demands, and psychological theories constitutes a major threat to the conceptual integrity of the field. Following groundbreaking earlier work, we propose a conceptual framework that distinguishes within-person (WP) and between-person (BP) sources of variation in psychological constructs. By simultaneously considering both sources of variation, it will be shown how to identify possible reasons for nonequivalence of BP and WP structures, as well as establishing areas of convergence. For this purpose, we first introduce the concept of *conditional equivalence* as a way to study partial structural equivalence of BP and WP structures in the presence of unconditional nonequivalence. Second, we demonstrate the construction of *likelihood planes* to explore the causes of structural nonequivalence. Third, we examine four common causes for unconditional nonequivalence—auto-regression, subgroup differences, linear trends, cyclic trends—and demonstrate how to account for them. Fourth, we provide an empirical example on BP and WP differences in attentiveness.

Keywords: Within-Person; Between-Person; Ergodicity; Person-Oriented Research; State Space Modeling.

Towards a Unified Framework for the Study of Between-Person and Within-Person Structures: Building a Bridge Between Two Research Paradigms

If there is one thing that all scientific disciplines have in common, it is the goal to discover and analyze relationships. However, the most fundamental questions regarding the definition of a relationship, regarding the attributes or entities that are related, or regarding the observations based on which a relationship is established, may already lead to considerable disagreement among researchers. More than half a century ago, Raymond B. Cattell proposed the covariation chart (Cattell, 1946), and later its extension, the Basic Data Relation Matrix (Cattell, 1966), in an attempt to systematically set out the totality of relations in psychological research. The covariation chart is a 3-dimensional "data cube" with persons, variables, and time points (occasions) as Cartesian axes. This results in a total of six possible relational matrices: the relation between variables across people (e.g., R-technique factor analysis), the relation between variables across occasions (e.g., P-technique factor analysis), the relation between people across occasions (e.g., S-technique factor analysis), as well as the transposes of the three (e.g., Q-, O-, and T-technique; see Cattell, 1966, p. 70). It is certainly not an overstatement to claim that 90% of empirical research in psychology is based on the first of these six possible relational matrices. In contrast, there are only few studies focusing on the relations between variables across occasions. The amount of research based on the other four approaches[1] seems negligible, leaving us with two (unequally) common paradigms in present day psychological research: The analysis of interindividual variation (i.e., studying the relations between variables across people) and the analysis of intraindividual variation (i.e., studying the relations between variables across occasions).

Without doubt, studying the relations between variables across people is important for psychological research. For example, it may be important to know whether someone who scores high on a numerical ability test—as compared to other people—is also likely to do

better in college. Likewise it seems important to find out whether this association, if it exists, is limited to numerical ability, or whether similar associations can be observed for numerical, verbal, and figural ability, because all of these abilities are expressions of a single "mental energy" factor (Spearman, 1904). If such a mental energy factor exists, however, we would expect to find it not only when studying the relations between variables across people, but also when studying the relations between variables across occasions within a given person. That is, if we observe a single individual across many days, we would expect him or her to score high on numerical, verbal, and figural ability tests on days with high "mental energy", and low on all three tests on days with low "mental energy" (Cattell, 1966, p. 71).

Unfortunately, such a 1:1 relationship between the analysis of interindividual variation and intraindividual variation has often been implicitly assumed in psychological research, but hardly ever explicitly tested (cf. Borsboom & Dolan, 2006; 2007; Borsboom, Kievit, Cervone, & Hood, 2009; Molenaar & Campbell, 2009; see also Blalock, 1967; Kuh, 1959). As pointed out by Molenaar (2004), this is highly problematic because "the classical ergodic theorems for psychology and psychometrics invalidate [the] conjectured generalizability" from between-person (BP) to within-person (WP) variation (2004, p. 201). Hence, "only under very strict conditions—which are hardly obtained in real psychological processes—can a generalization be made from a structure of interindividual variation to the analogous structure of intraindividual variation" (2004, p. 201). Given that most of applied psychology is concerned with individuals (e.g., patients in clinical psychology), this lack of generalizability constitutes a major problem for the field.

An Illustrative Example

Before taking a closer look at the conditions under which BP and WP structures may or may not be equivalent, we start with a simplified example in order to develop an intuitive understanding of what it may mean to compare BP analyses (based on interindividual

variation) to WP analyses (based on intraindividual variation). For this purpose, let us assume we measured anxiety by three different variables: (self-) reported nervousness, (observed) trembling, and sweating assessed via the galvanic skin response (GSR). The model, along with some hypothetical parameter estimates, is shown in Figure 1.

Based on similar considerations as for the ”mental energy” factor, we could expect that the BP structure resulting from interindividual variability across $i = 1,\ldots, N$ individuals at one occasion (to the left of Figure 1) is equivalent to the WP structure, resulting from intraindividual variability of a single individual across $t = 1,\ldots, T$ occasions (e.g., days; to the right of Figure 1). That is, people with a high level of anxiety—as compared to other people—should also exhibit high levels of reported nervousness, trembling, and sweating. Likewise, an individual should show high levels of reported nervousness, trembling, and sweating in situations with high anxiety and low levels on all three indicators in situations with low anxiety. As shown in Figure 1A, the two structures should be identical in this case.

However, there are many reasons why this equivalence may not hold. For example, for a patient with Parkinson disease trembling would not be indicative of anxiety, so that the WP structure of this person would be different from the BP structure (i.e., nonequivalence due to a person; Figure 1B). Another reason could be that during hot summer days, sweating may not be a good indicator of anxiety. In this case, the two structures may be identical if the between variability was assessed in the same season of the year as the within variability, or they may differ if that is not the case (i.e., nonequivalence due to occasion; Figure 1C). Finally, there may be group differences. For example, men may exhibit greater variability in anxiety (both between and within) than women. When ignoring this factor, equivalence does not hold, because the average interindividual variability (e.g., 1.0) is neither equivalent to the WP variability of a woman (e.g., 0.5), nor of a man (e.g., 1.5; that is, nonequivalence due to group differences; Figure 1D). It is easy to come up with many more examples why the WP

and BP structure should differ, and we can only concur with Molenaar (2004) in his claim that it will be close to impossible to find empirical data that support the notion of equivalence between inter- and intraindividual variability for all possible combinations of persons and occasions.

In this paper, we take a reconciliatory stance that bridges the gap dividing BP and WP structures. Instead of asking *whether BP and WP structures are equivalent;* we ask *what are the specific reasons that prevent them from being equivalent?* The idea is that if we are able to identify and control for sources of nonequivalence, we may also be able to establish (conditional) equivalence. In terms of our example: If we account for the fact that the sample included a patient with Parkinson disease, consisted of men and women, and that measurements were taken during hot and cold days, the BP and WP structures are actually equivalent. Put more generally, the central theme of this work is to conceive of BP and WP analyses not as two independent, or even competing, research paradigms, but rather to identify their commonalities in conjunction with their differences.

On the Equivalence of Between-Person and Within-Person Structures

Before taking a closer look at the conditions and consequences for BP and WP structural equivalence, it is important to clarify what exactly is meant by "equivalence." In line with previous research (e.g., Molenaar, 2004) we thereby limit ourselves to the situation of multivariate normality, which allows us to focus on the first and second order moments of a distribution. That is, we assume a p-dimensional vector $\mathbf{Y}$ of multivariate normally distributed variables with mean vector $\boldsymbol{\mu}$ and $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. Under the null hypothesis of complete structural equivalence, any observation y_{it} of individual $i = 1,\ldots, N$ at time point (occasion) $t = 1,\ldots, T$ can be considered a random draw out of this distribution. For reasons of simplicity, we also limit ourselves to discrete measurement occasions with

equidistant time intervals. Finally, we assume that *T* occasions and *N* individuals are drawn from a finite population of T_{pop} occasions and N_{pop} individuals (with large *T* and *N*).

Structural Equivalence: A Working Definition

The *BP* mean vector and covariance matrix at any occasion *t* are defined as[2]:

$$\boldsymbol{\mu}_t = N_{pop}^{-1} \sum_{i=1}^{N_{pop}} \mathbf{y}_{it} \qquad \boldsymbol{\Sigma}_t = N_{pop}^{-1} \sum_{i=1}^{N_{pop}} (\mathbf{y}_{it} - \boldsymbol{\mu}_t)(\mathbf{y}_{it} - \boldsymbol{\mu}_t)^{\mathrm{T}}. \quad (1)$$

Likewise, the *WP* mean vector and covariance matrix for any individual *i* are defined as:

$$\boldsymbol{\mu}_i = T_{pop}^{-1} \sum_{t=1}^{T_{pop}} \mathbf{y}_{it} \qquad \boldsymbol{\Sigma}_i = T_{pop}^{-1} \sum_{t=1}^{T_{pop}} (\mathbf{y}_{it} - \boldsymbol{\mu}_i)(\mathbf{y}_{it} - \boldsymbol{\mu}_i)^{\mathrm{T}}. \quad (2)$$

Obviously, if all observations $\mathbf{y}_{it}$ are drawn independently from the same underlying distribution (the same generating model), the BP (at any fixed *t*) and WP (for any individual *i*) structures are asymptotically equivalent. In the case of multivariate normality, we can thus define structural equivalence as the equivalence of the BP and WP mean vector and covariance matrix:

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_t \quad (3)$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_t \quad (4)$$

If this applies to all possible combinations of *i* and *t*, we speak of *unconditional equivalence*.

The statement "applies to all possible combinations of *i* and *t*" implies that individuals are independent and that the same generating model applies to all individuals. In addition, all individual processes are stationary (i.e., strongly stationary in case of Gaussian processes as discussed in the present paper) and contain no systematic (cyclic) trends. This corresponds to the two conditions of *ergodicity* set forth by Molenaar (i.e., homogeneity and stationarity; 2004; pp. 206-207), who used the term ergodicity to describe a "process in which the structures of IEV [interindividual variation] and IAV [intraindividual variation] are (asymptotically) equivalent" (Molenaar, 2004, p. 206).

Factors Affecting Structural Equivalence

Based on the definitions in the previous section and in line with Molenaar (2004; Molenaar & Campbell, 2009; Molenaar, Huizenga, & Nesselroade, 2003), there are two conditions for unconditional equivalence of BP and WP structures: (1) *Homogeneity* of individuals, meaning that the same generating model underlies all persons (*i*). This implies that all *i* are interchangeable. For example, individuals must not be grouped or nested in any meaningful way. (2) *Stationarity*, meaning for Gaussian processes that no mean, variance, or covariance changes over time are permitted and that all *t* are interchangeable. Taken separately, both conditions are necessary, but only in combination are they sufficient for unconditional structural equivalence. If both conditions are met, this entails that all *i* are interchangeable with all *t*.

Conversely, the two requirements for unconditional equivalence imply that BP-WP-equivalence is affected by (1) *persons* and (2) *time*. As illustrated in the introductory example, structural equivalence may be violated *due to persons* if the generating model differs across people, for example because individuals are grouped in a meaningful way (e.g., men vs. women; healthy vs. patient with Parkinson disease). Likewise, structural equivalence may be violated *due to time* if there are mean trends or cyclic trends (e.g., sweating increases during hot summer days).

Given that almost all psychological constructs exhibit some systematic change over time and given that (groups of) individuals tend to differ from each other in meaningful ways, the requirements for unconditional structural equivalence as outlined above are almost never met. It is important to note that this does *not* imply that the BP and WP structures are completely independent of each other and that nothing can be learned from one about the other. Indeed, it may be useful to posit that these structures consist of multiple sources of variation, some of which are equivalent. In order to learn something from one about the other,

it is therefore crucial to control for factors which are known to affect only one of the two structures. Hence, we introduce the notion of *conditional equivalence* as a way to study partial structural equivalence in the presence of unconditional nonequivalence.

Conditional Equivalence

Let $\boldsymbol{\theta}$ denote a vector of all parameters to be estimated in any given statistical model. In case of a saturated model, $\boldsymbol{\theta}$ corresponds simply to the mean vector and covariance matrix: $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Thus, structural equivalence is given if

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_t \tag{5}$$

for all possible combinations of *i* and *t*.

In case of a saturated model, the definition in Equation 5 is identical to the definition provided above (Equation 3 and 4). However, Equation 5 is more general and may refer to any model that is over- or just-identified (saturated). This opens up the possibility to control for factors that are known to affect only one of the two structures (e.g., to control for trends in the WP structure or to control for group differences in the BP structure). We therefore propose to define structural equivalence as conditional equivalence of model parameters:

$$(\boldsymbol{\theta}_{ic}|\boldsymbol{\theta}_{iu}) = (\boldsymbol{\theta}_{tc}|\boldsymbol{\theta}_{tu}). \tag{6}$$

The subscript *c* denotes parameters that are common to both models (BP and WP), while the subscript *u* denotes parameters that are unique in each of the two models. Conditional equivalence is met if $\boldsymbol{\theta}_{ic} = \boldsymbol{\theta}_{tc}$ after controlling for $\boldsymbol{\theta}_{iu}$, respectively $\boldsymbol{\theta}_{tu}$.

For any identified model and any combination of *i* and *t*, maximum likelihood (ML) parameter estimates of $\boldsymbol{\theta}_{ic}, \boldsymbol{\theta}_{iu}, \boldsymbol{\theta}_{tc}, \boldsymbol{\theta}_{tu}$ can be obtained by maximizing the natural logarithm of the likelihood (*LL*):

$$LL(\boldsymbol{\theta}_{ic}, \boldsymbol{\theta}_{iu}, \boldsymbol{\theta}_{tc}, \boldsymbol{\theta}_{tu}) = LL_i(\boldsymbol{\theta}_{ic}, \boldsymbol{\theta}_{iu}) + LL_t(\boldsymbol{\theta}_{tc}, \boldsymbol{\theta}_{tu}) \tag{7}$$

with $LL_i(\boldsymbol{\theta}_{ic}, \boldsymbol{\theta}_{iu})$ defined as

$$LL_i(\boldsymbol{\theta}_{ic}, \boldsymbol{\theta}_{iu}) = -\frac{1}{2}[T \cdot p \cdot \log(2\pi) + \log|\boldsymbol{\Upsilon}_i| + (\mathbf{y}_i - \boldsymbol{\nu}_i)^{\mathrm{T}}\boldsymbol{\Upsilon}_i^{-1}(\mathbf{y}_i - \boldsymbol{\nu}_i)], \quad (8)$$

and $LL_t(\boldsymbol{\theta}_{tc}, \boldsymbol{\theta}_{tu})$ defined as

$$LL_t(\boldsymbol{\theta}_{tc}, \boldsymbol{\theta}_{tu}) = -\frac{1}{2}\left[N \cdot p \cdot \log(2\pi) + N \cdot \log|\boldsymbol{\Upsilon}_t| + \sum_{i=1}^{N}(\mathbf{y}_{t,i} - \boldsymbol{\nu}_t)^{\mathrm{T}}\boldsymbol{\Upsilon}_t^{-1}(\mathbf{y}_{t,i} - \boldsymbol{\nu}_t)\right]. \quad (9)$$

There are several things worth mentioning: First, Equation 9 corresponds to the standard maximum likelihood function (e.g., Bollen, 1989), with $\boldsymbol{\Upsilon}_t$ denoting the $p \times p$ model-implied covariance matrix, and $\boldsymbol{\nu}_t$ denoting the p-variate model-implied mean vector[3]. For a sample of $i = 1,\ldots, N$ independent individuals, the total log-likelihood is simply the sum of the individual log-likelihoods. By adding the subscript t, we emphasize that this refers to the BP structure at time point t, which in the context of the present paper is always cross-sectional. Second, Equation 8 gives the log-likelihood of a single individual across T time points, resulting in a $(p \cdot T) \times (p \cdot T)$ model-implied covariance matrix $\boldsymbol{\Upsilon}_i$ and $(p \cdot T)$-dimensional mean vector $\boldsymbol{\nu}_i = (\boldsymbol{\nu}_{i,1,1} \quad \cdots \quad \boldsymbol{\nu}_{i,p,1} \quad \cdots \quad \boldsymbol{\nu}_{i,1,T} \quad \cdots \quad \boldsymbol{\nu}_{i,p,T})^{\mathrm{T}}$. Note, that $\mathbf{y}_i$ in Equation 8 represents a $(p \cdot T)$-dimensional vector $\mathbf{y}_i = (\mathbf{y}_{i,1,1} \quad \cdots \quad \mathbf{y}_{i,p,1} \quad \cdots \quad \mathbf{y}_{i,1,T} \quad \cdots \quad \mathbf{y}_{i,p,T})^{\mathrm{T}}$ for a single individual i, whereas $\mathbf{y}_{t,i}$ is a p-dimensional vector of observed values for individual i at a single time point t. As demonstrated by Voelkle, Oud, von Oertzen, and Lindenberger (2012) this allows not only to obtain true maximum likelihood parameter estimates of individual time series by means of structural equation modeling (e.g., dynamic factor models; Hamaker, Dolan, & Molenaar, 2003; Molenaar, 1985; Nesselroade, McArdle, Aggen, & Meyers, 2002), but also to integrate the analysis of WP and BP structures as shown in Equation 7. The equivalence of the two structures can now be explicitly tested by subtracting two times the log-likelihood of the model under the null hypothesis of structural equivalence $[(\boldsymbol{\theta}_{ic}|\boldsymbol{\theta}_{iu}) = (\boldsymbol{\theta}_{tc}|\boldsymbol{\theta}_{tu})]$ from two times the log-likelihood of the model under the alternative hypothesis $[(\boldsymbol{\theta}_{ic}|\boldsymbol{\theta}_{iu}) \neq (\boldsymbol{\theta}_{tc}|\boldsymbol{\theta}_{tu})]$, that is, a standard likelihood ratio test (Bollen, 1989; Voelkle et al., 2012). Third,

even though $\boldsymbol{\theta}_{iu}$ and $\boldsymbol{\theta}_{tu}$ refer, by definition, to parameters that are unique to either the within *or* between structure, it does not mean that they are independent of the parameters of the other group. For example, a linear WP trend (ρ) is by definition unique to the within structure $\boldsymbol{\theta}_{iu} = \{\rho\}$ because it is defined over occasions. Nevertheless, if the WP trend applies to all individuals, it also affects the parameters of the between structure: If everyone shows an intraindividual increase of ρ units from occasion to occasion, then the BP mean α_{between} at occasion *t* will also be ρ units higher than at $t - 1$. The crucial point, however, is that this relationship is known and can be expressed algebraically. In our example the BP mean ($\boldsymbol{\theta}_{tc} = \{\alpha_{\text{between}}\}$) at occasion $t = 1,\ldots, T$ would be $\alpha_{\text{between}}(t) = \alpha_{\text{between}} + \rho \cdot (t - 1)$. Thus, when the goal is to compare $\boldsymbol{\theta}_{tc} = \{\alpha_{\text{between}}\}$ to $\boldsymbol{\theta}_{ic} = \{\alpha_{\text{within}}\}$ after controlling for a WP trend $\boldsymbol{\theta}_{iu} = \{\rho\}$, it is important not to compare α_{between} directly to α_{within}, but to compare α_{within} to $\alpha_{\text{between}} + \rho \cdot (t - 1)$ at any given time point *t*. Of course other constraints are necessary for other factors we may want to control for (i.e., $\boldsymbol{\theta}_{iu}$; $\boldsymbol{\theta}_{tu}$).

Another complication—but also an opportunity—arises from the fact that across individuals, the parameters that are unique to the WP structure ($\boldsymbol{\theta}_{iu}$), may be correlated with parameters that are unique to the BP structure ($\boldsymbol{\theta}_{tu}$). For example, this may be a correlation between the position of an individual relative to other individuals at a given point in time and the WP change over time (e.g., a fanspread effect; Campbell & Erlebacher, 1970; Kenny, 1974). Obviously, it takes a full data cube (multiple individuals at multiple occasions) to estimate such a correlation, so it is of little use if only a single WP structure and a single BP structure are available. If the relationship is known, however, it can be used to predict WP parameters from BP parameters and vice versa. We will return to this issue below where we will also discuss an alternative approach to the use of nonlinear constraints among WP and BP parameters, which is less explicit but simpler to implement in practice.

State Space Modeling

The approach of comparing a WP structure to a BP structure by means of a likelihood ratio test as described above has originally been proposed within a structural equation modeling (SEM) framework. While this approach offers great flexibility in terms of model specification, with an increasing number of time points and variables it comes at the price of a large model-implied covariance matrix and greatly increased computation time (Voelkle et al., 2012)[4]. For a large number of model comparisons, this poses great computational difficulties. In the present paper, we therefore use the Kalman filter (Kalman, 1960), a different estimation approach within the state space modeling framework. In contrast to the simultaneous estimation approach in SEM, the Kalman filter operates recursively on the time series, making it a practically more suitable approach, especially when the number of occasions is large. Here, we only briefly review the basics of state space modeling; for a more detailed discussion, see Commandeur and Koopman (2007), Durbin and Koopman (2001), and Harvey (2001). For a concise introduction and a comparison to SEM, we refer the reader to Chow, Ho, Hamaker, and Dolan (2010; see also Oud & Singer, 2008; Oud, van den Bercken, & Essers, 1990; Zhang, Hamaker, & Nesselroade, 2008). Most importantly, Chow et al. (2010, Appendix; Otter, 1986) demonstrated that for $T = 1$ as in Equation 9, all elements in SEM and the state space framework are identical, resulting in identical (maximum likelihood) estimates. Likewise, for $N = 1$ as in Equation 8 many—but not all—SEMs can be reparameterized as state space models. Whenever such a reformulation is possible, the resulting parameter estimates by means of the Kalman filter are maximum likelihood estimates—given that the usual assumptions hold true (cf. Shumway & Stoffer, 2004). Thus, our decision to use a state space modeling framework instead of an SEM framework was primarily motivated by pragmatic reasons (computation time). Our arguments, however, are more general and not bound to either of the two frameworks. For this reason we focus

primarily on the commonalities between state space modeling and SEM, as the differences between the two frameworks are largely irrelevant for the models discussed in this article.

Similar to the measurement model and structural model in SEM (Bollen, 1989), one distinguishes between an *observation equation* and a *transition equation* in state space modeling. Using conventional SEM notation, the observation equation is given in Equation 10:

$$\mathbf{y}_{it} = \boldsymbol{\tau}_{it} + \boldsymbol{\Lambda}_{it}\boldsymbol{\eta}_{it} + \boldsymbol{\epsilon}_{it} \tag{10}$$

$$\mathbf{y}_{it} = \boldsymbol{\tau}_{i} + \boldsymbol{\Lambda}_{i}\boldsymbol{\eta}_{it} + \boldsymbol{\epsilon}_{it} \tag{10a}$$

$$\mathbf{y}_{it} = \boldsymbol{\tau}_{t} + \boldsymbol{\Lambda}_{t}\boldsymbol{\eta}_{it} + \boldsymbol{\epsilon}_{it} \tag{10b}$$

Matrix $\mathbf{y}_{it}$ denotes a *p*-dimensional vector of manifest variables with observations of individual *i* at time point *t*. The manifest variables are regressed on (caused by) *q* latent variables $\boldsymbol{\eta}_{it}$ with factor loading matrix $\boldsymbol{\Lambda}_{it} \in \mathbb{R}^{p \times q}$, intercept vector $\boldsymbol{\tau}_{it} \in \mathbb{R}^{p}$ and *p*-dimensional measurement error vector $\boldsymbol{\epsilon}_{it}$. Note that Equation 10 defines a generic model where all parameters are allowed to differ across individuals *i* and time points *t*. If fitted to a single individual *i*, parameters $\boldsymbol{\tau}_i$ and $\boldsymbol{\Lambda}_i$ are assumed to be time-invariant (WP Equation 10a), whereas parameters $\boldsymbol{\tau}_t$ and $\boldsymbol{\Lambda}_t$ are assumed to be invariant across individuals when estimating a BP structure at a single time point *t* (Equation 10b). As discussed before, if unconditional BP-WP equivalence of the factor loading structure holds, then $\boldsymbol{\Lambda}_t = \boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}$ for all *i* and *t*.

The transition equation is defined as

$$\boldsymbol{\eta}_{it} = \boldsymbol{\alpha}_{it} + \mathbf{B}_i\boldsymbol{\eta}_{i(t-1)} + \boldsymbol{\zeta}_{it} \tag{11}$$

$$\boldsymbol{\eta}_{it} = \boldsymbol{\alpha}_{i} + \mathbf{B}_i\boldsymbol{\eta}_{i(t-1)} + \boldsymbol{\zeta}_{it} \tag{11a}$$

$$\boldsymbol{\eta}_{it} = \boldsymbol{\alpha}_t + \boldsymbol{\zeta}_{it} \tag{11b}$$

with $\mathbf{B}_i \in \mathbb{R}^{q \times q}$ denoting the transition matrix which relates $\boldsymbol{\eta}_{i(t-1)}$ at $t-1$ to $\boldsymbol{\eta}_{it}$ at time point t, $\boldsymbol{\alpha}_{it} \in \mathbb{R}^q$ denoting the intercept vector of the transition equation, and $\boldsymbol{\zeta}_{it}$ the q-dimensional vector of dynamic errors. Equation 11 defines the generic transition equation, Equation 11a the WP transition equation, and Equation 11b the BP equation. Note, that transition matrix $\mathbf{B}_i$ is only present at the WP level, while no transition from one individual to the next is assumed at a given time point t. Throughout, we assume that WP parameter estimates are time-invariant and that the dynamic errors as well as measurement errors are uncorrelated and normally distributed with $\boldsymbol{\zeta}_{it} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}_i)$, respectively $\boldsymbol{\epsilon}_{it} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Theta}_i)$ at the WP level, and $\boldsymbol{\zeta}_{it} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}_t)$, respectively $\boldsymbol{\epsilon}_{it} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Theta}_t)$ at the BP level.

In contrast to SEM, parameter estimation is not carried out by minimizing the difference between the full $T \cdot p \times T \cdot p$ model-implied covariance matrix $\boldsymbol{\Sigma}$, respectively mean vector $\boldsymbol{\mu}$, and the observed covariance matrix $\mathbf{S}$, respectively mean vector $\bar{\mathbf{x}}$, but rather by minimizing the so-called one-step-ahead prediction error (Commandeur & Koopman, 2007). In principle, this is done in two steps: In a first step, a person's state at time point t is *predicted* based on the information of the previous time point ($\boldsymbol{\eta}_{i(t|t-1)} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_{i(t-1|t-1)}$). The same applies to the associated covariance matrix ($\mathbf{P}_{t|t-1} = \mathbf{B}\mathbf{P}_{t-1|t-1}\mathbf{B}^{\mathrm{T}} + \boldsymbol{\Psi}$). Given $\boldsymbol{\eta}_{i(t|t-1)}$, it is easy to compute the predicted observations $\mathbf{y}_{i(t|t-1)}$ at time point t (based on information at time point $t-1$) as shown in Equation 11. Now it is also possible to compare $\mathbf{y}_{i(t|t-1)}$ to the actually observed $\mathbf{y}_{it}$ resulting in the one-step-ahead prediction error $\mathbf{e}_{i(t|t-1)} = \mathbf{y}_{it} - \mathbf{y}_{i(t|t-1)}$. In a second step the predictions (i.e., $\boldsymbol{\eta}_{i(t|t-1)}$ and $\mathbf{P}_{t|t-1}$) are *updated* in light of the new observations at time point t, resulting in $\boldsymbol{\eta}_{i(t|t)}$ and $\mathbf{P}_{t|t}$. Thereby the so-called Kalman gain function weights the relative importance of the newly obtained information (i.e., the certainty of the measurement) against the certainty of the prediction.

The process is then repeated for all time points. Before it can be started, however, the initial state value $\boldsymbol{\eta}_{i(0|0)}$ and covariance matrix $\mathbf{P}_{0|0}$ need to be specified. If no prior information is available, we may simply fix it to any noninformative values. For more detailed information, we refer the reader to the above-mentioned literature (e.g., Commandeur & Koopman, 2007; Durbin & Koopman, 2001).

The Likelihood Plane

Having provided a working definition of BP and WP structural equivalence, and having sketched out the statistical approach that allows us to test this equivalence, we can now turn to the question of how this may be done in practice. Obviously, the problem is that for T time points and N individuals, $T \cdot N$ combinations of a BP and WP structure exist, and for every single BP-WP combination the structures (models) may or may not be equivalent. Previous studies have approached this problem by selecting some individuals and comparing their WP structure to the BP structure at a selected occasion, for example the BP structure at the first occasion, or the average BP structure across all occasions (Molenaar & Campbell, 2009; see also Hamaker, Dolan, & Molenaar, 2005; Lebo & Nesselroade, 1978, who focus primarily on a comparison among individuals; for an alternative approach to accomodate for sample heterogeneity in recovering effective connectivity maps, see Gates & Molenaar, 2012). Based on the strict definition that equivalence is only met if the BP and WP structures are equivalent for *all* possible combinations of *i* and *t*, this approach is reasonable because a single instance in which the two are not equivalent suffices to reject the assumption of (unconditional) structural equivalence. What goes undetected, however, is the *proportion* of equivalent BP-WP-structures given all possible combinations. Arguably, it makes a big difference whether the between structure reveals absolutely nothing about any given individual, whether there is some minor relationship between the two (e.g., 10% of the WP and BP structures are identical), or whether the two are the same in a large number (e.g.,

90%) of BP-WP combinations. We propose to address this issue through the construction of likelihood planes.

A One-Factor Model

To illustrate our arguments we use data that were generated by a 1-factor model with three indicators, similar to our introductory example in Figure 1. Later on the model will be modified in a stepwise fashion to examine the reasons and consequences of structural nonequivalence. For now, let the true model parameters according to Equations 10 and 11 be $\boldsymbol{\tau}^{\mathrm{T}} = (0 \quad 0 \quad 0)$, $\boldsymbol{\Lambda}^{\mathrm{T}} = (1.0 \quad 0.8 \quad 0.8)$, $\boldsymbol{\alpha} = (0)$, $\mathbf{B} = (0)$, with $\boldsymbol{\epsilon}_{it} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(0.2))$ and $\boldsymbol{\zeta}_{it} \sim \mathrm{N}(0,1.0)$. We refer to this specification as the baseline condition. Data for $N = 100$ independent individuals and $T = 80$ time points were generated according to this model. Note that, because $\mathbf{B} = (0)$, there is no temporal order, thus not only are all *i* interchangeable, but also all *t*. Likewise, because data for each person and time-point combination were generated according to the same model, all *i* are interchangeable with all *t*. In the baseline condition the BP and WP structures are therefore equivalent.

Constructing and Interpreting Likelihood Planes

In the baseline condition all parameters ($\boldsymbol{\theta}$) are common to both models (BP and WP). There are no parameters that refer to only one of the two structures (i.e., no $\boldsymbol{\theta}_{iu}$ or $\boldsymbol{\theta}_{tu}$). Thus, the question of structural equivalence reduces to comparing the null hypothesis $\boldsymbol{\theta}_i = \boldsymbol{\theta}_t$ against the alternative hypothesis $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_t$. In line with common conventions, let us fix the first factor loading to 1.0. We further constrain the intercepts of the manifest variables to zero, resulting in a total of seven free parameters under the null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_i = \boldsymbol{\theta}_t = \{var(\varepsilon_1);\ var(\varepsilon_2);\ var(\varepsilon_3);\ \lambda_2;\ \lambda_3;\ \alpha;\ var(\zeta)\}$. Note, that in this basis form (i.e., without additional constraints), the covariance structure is saturated. Although this simple model constitutes a special case in this regard, this does not affect the generality of our arguments.

Under the alternative hypothesis (H_1) all parameters are allowed to differ, resulting in a total of 14 free parameters.

For each person and occasion combination, we may now formulate two models in terms of Equation 10 and 11. In one model the parameters of the BP and WP structure are constrained to equality, in the other they are allowed to differ. In a next step, maximum likelihood parameter estimates and the log-likelihood of the data given the model are obtained. Finally, because the two models are nested, the log-likelihoods can be compared by means of a log-likelihood ratio test as described above. For $N = 100$ individuals and $T = 80$ time points this results in $2 \cdot T \cdot N = 16{,}000$ models to be estimated and 8,000 associated likelihood ratio tests. Using R version 2.15.0 (R Core Team, 2012) for data generation and mkfm6 (Dolan, 2010) for the state space analyses, the computation of the entire likelihood plane took about 300 min on a standard 2.4GHz processor personal computer. (Dolan, 2010)

The likelihood ratios may now be plotted for each person and time-point combination, resulting in a 3-dimensional likelihood plane as shown in Figure 2. For each likelihood ratio test the color indicates whether it is significant (red) or not (green) at an alpha level of 1%. Thus, given that the null hypothesis of structural equivalence is correct, we should see many "green valleys" and few "red mountain tops." As a matter of fact, in this example (baseline condition), 99.61% of all likelihood ratio tests were nonsignificant, which is close to the expected 99% of nonsignificant tests (given an alpha level of 1%).

The likelihood plane is a powerful tool to visually inspect the reasons for structural nonequivalence: As discussed above, nonequivalence may be due to persons, occasions, or combinations thereof. For example, if a single occasion would be "unusual" (e.g., a public holiday), this would be indicated by a red "mountain range" at this occasion across all individuals. Likewise, if a person "sticks out" (i.e., his or her within-structure is different from other WP and BP structures), this should be reflected in a red "mountain range" of this

person across all occasions. Finally, structural equivalence may be violated for specific person × occasion interactions, which would show up as single mountain tops. Having identified "odd" occasions, "odd" persons, or "odd" occasion-person combinations, we can proceed with a closer investigation of the specific reasons for nonequivalence. This is of primary concern in the remainder of this article.

There are two more things worth noting with respect to the construction and interpretation of likelihood planes: First, because any individual *i* is also part of the between structure at time point *t*, there is a small overlap of *p* data points in the BP-and WP-data (in this example $p = 3$). This is comparable to the problem of an item-test correlation in psychometrics, where the item is also part of the whole test and thus biases the "item-total" correlation. Just like in psychometrics, the scores of person *i* at occasion *t* should hence be removed from the between structure before comparing the two. This is done for all analyses in the present paper. Second, although the data ($\mathbf{y}_{it}$) for each person and occasion combination (each "cell") were independently generated in the baseline condition (according to the true model) the results of the $T \cdot N$ likelihood ratio tests in Figure 2 are *not* independent of each other. This is because we do not compare single cells. Instead, each person-occasion combination in Figure 2 represents the comparison of an *entire* time series (person) to an *entire* sample of individuals at a given occasion *t*. Thus, as mentioned above, if the WP structure of a single person is somehow "deviant" this is likely to be the case for all occasions under consideration, resulting not in a single "mountain top" but rather in an entire "mountain range." This is nicely illustrated by person ID60 in Figure 2. This dependency, however, does not affect the alpha level under the null hypothesis because the number of expected type I errors increases in the same fashion as the number of occasions, respectively persons, increases.

Examining Four Common Reasons for Structural Nonequivalence

A likelihood plane in line with the baseline condition of complete structural equivalence as illustrated in Figure 2 is unlikely to be found in most real-world situations. Rather, basic mechanisms common to psychological phenomena will result in systematic changes over time or group differences (see Figure 1 for examples), thereby producing a pattern of (possibly complete) nonequivalence. In the following, we explore four such factors. We thereby first "destroy" *unconditional* structural equivalence and then discuss the model changes necessary in order to (re)establish *conditional* structural equivalence, that is, to recover the hidden equivalence between two structures. To facilitate visual inspection, we use 2D likelihood planes in the following, which correspond to the top view of the more detailed 3D planes introduced above. We provide the 3-dimensional likelihood planes as supplementary material online. In contrast to the printed version, these images are interactive, that is, they may be rotated and inspected from different angles.

Autoregression

Maybe the most unrealistic assumption of the baseline model is the assumption that there is no temporal order among observations within individuals over time (i.e., the assumption that all *t* are interchangeable). Notably, the original P-technique factor analysis proposed by Cattell et al. (1947) is based on this assumption. P-factor analysis has been repeatedly criticized for this reason, and a class of models generally referred to as dynamic factor models—a special variant of state space models—has been proposed as a better alternative (Browne & Nesselroade, 2005; Molenaar, 1985; Molenaar & Nesselroade, 2009). In contrast to P-technique, dynamic factor analysis explicitly accounts for the lagged structure of the data. We presume that this is of equal importance when investigating structural equivalence.

To illustrate this point, we first generated data in line with the baseline model but added a first-order autoregressive effect to the WP structure by setting $\mathbf{B} = \beta = 0.8$ and

$var(\boldsymbol{\zeta}) = \psi = 1 - 0.8^2 = 0.36$. An overview of all simulation conditions is given in Table 1. Second, for each person and time-point combination, we fitted two models to the data *without* accounting for the newly added autoregressive effect. In the first model all BP-WP-parameters were constrained to equality (H_0: $\boldsymbol{\theta}_i = \boldsymbol{\theta}_t$), in the second model all parameters were freely estimated (H_1: $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_t$). As before, the two models were compared by means of a likelihood ratio test. The resulting likelihood planes is given on the left side of Figure 3A. In a third step, we accounted for the autoregressive effect by explicitly including β as a parameter to be estimated in the within structure, but not in the between structure. Again two models were estimated, one in which all BP-WP-parameters were constrained to equality and one in which all parameters were allowed to differ. That is, we tested for conditional structural equivalence by testing H_1: $(\boldsymbol{\theta}_{ic}|\boldsymbol{\theta}_{iu}) \neq \boldsymbol{\theta}_{tc}$ against H_0: $(\boldsymbol{\theta}_{ic}|\boldsymbol{\theta}_{iu}) = \boldsymbol{\theta}_{tc}$ for all *i* and *t*, resulting in a total of $2 \cdot 2 \cdot 80 \cdot 100 = 32{,}000$ state space models to be estimated.

As pointed out before, even though the autoregressive parameter is by definition a WP parameter ($\boldsymbol{\theta}_{iu} = \{\beta\}$), it also affects BP parameters. More specifically, because $\beta > 0$ the dynamic error variance $\psi_{\text{within}} = var(\eta_{\text{within}})$ is reduced. Thus, there is no point in comparing ψ_{within} directly to the BP variance $\phi_{\text{between}} = var(\eta_{\text{between}})$. Instead, ψ_{within} must be compared with $\phi_{\text{between}}(1 - \beta^2)$. This relationship, however, is known and can be implemented as a nonlinear constraint during parameter estimation[5]. The resulting likelihood planes is shown on the right side of Figure 3A. Average parameter estimates of the constrained condition (under the H_0) are presented in Table 2. For reasons of space, parameter estimates of the other conditions are not given, but were of equal quality (i.e., similarly close to the true parameters).

As apparent from Figure 3A (left side), the BP and WP structures are no longer equivalent in the presence of a lagged effect and the number of nonsignificant likelihood ratios reduces to 85.23% in this example. Furthermore, because nonequivalence is due to an

autoregressive effect at the person level, we see a pattern of vertical red lines ("mountain ranges") of individuals across occasions. The nonequivalence is due to the reduced variance of the dynamic error term. Because β is 0.8 at the within level, but is nonexistent ($\beta = 0$) at the between level, the dynamic error variance differs between the two models. It is important to note that the factorial structures remain identical apart from this parameter. Regardless of this, according to the definition of unconditional equivalence outlined above, we would need to conclude that the BP and WP structures are nonequivalent. In contrast, when testing for conditional equivalence—that is, when controlling for the autoregressive effect (β) at the within level—the number of significant likelihood ratio tests reduces to 0.76% (at an alpha level of 1%) and results in a likelihood plane (right side of Figure 3A) that is similar to the likelihood plane in the baseline condition (Figure 2).

Group Differences

Another reason for nonequivalence of BP and WP structures may be that individuals are grouped in a meaningful way. For example, if a sample is comprised of men and women, it may be that the two groups differ in the variance of the construct in question. In that case, the BP variance of the entire sample (men *and* women) may be quite different from the WP variance of either a man or a woman (see Figure 1D for an example).

To mimic this situation, we generated data in line with the baseline model but for two different groups of 50 persons each. In the first group (say men) the variance of the latent factor was set to $var(\eta_{g1}) = \psi_{g1} = \phi_{g1} = 4$, in the second group (say women) it was set to $var(\eta_{g2}) = \psi_{g2} = \phi_{g2} = 1$. For each person and time-point combination we first tested for unconditional equivalence of the between and within structures (H_0: $\boldsymbol{\theta}_i = \boldsymbol{\theta}_t$ vs. H_1: $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_t$) and then for conditional equivalence after controlling for group membership (H_0: $\boldsymbol{\theta}_{ic} = (\boldsymbol{\theta}_{tc}|\boldsymbol{\theta}_{tu})$ vs. H_1: $\boldsymbol{\theta}_{ic} \neq (\boldsymbol{\theta}_{tc}|\boldsymbol{\theta}_{tu})$, with $\boldsymbol{\theta}_{tu} = \{\phi_{g1}, \phi_{g2}\}$).

Average parameter estimates of the constrained condition (under the H_0) are presented in Table 2, and the resulting likelihood planes are shown in Figure 3B. Before discussing the results, however, a word on the practical implementation of this test is in order. Other than in the previous example, each state space model is now comprised of three (sub)models: two BP models and one WP model. This allows us to let the variance of η (or any other parameter for that matter) differ between men and women. When doing so it is important to compare the WP model to the appropriate BP model (i.e., the WP variance of a man must be constrained to the BP variance of the correct group [of men]) and to remove the p overlapping data points of person i from the appropriate group.

Apart from the slightly more complicated model setup, the pattern of results when testing for conditional equivalence after controlling for BP factors (i.e., nonequivalence due to persons) is the same as when controlling for time-related factors (i.e., nonequivalence due to occasion). As apparent from the left side of Figure 3B, the hypothesis of unconditional structural equivalence is readily rejected. Only 62.66% of all likelihood ratio tests were nonsignificant. This is particularly true for the group of women (person 51 to 100 in Figure 3B), because the average BP variance of the entire sample (2.44; see Table 2) is closer to the WP variance of men than women (i.e., an average BP variance of (4 + 1) / 2 = 2.5 is an increase by factor 2.5 / 1 = 2.5 for women, but only a decrease by factor 4 / 2.5 = 1.6 for men). If the difference in variance between the two groups were larger, it could very well be that all likelihood ratio tests would turn out significant. In contrast, when testing for conditional equivalence—that is, when controlling for the difference in variance at the between level—the number of significant likelihood ratio tests reduces to 1.08% (alpha level: 1%), and results in a likelihood plane (right side of Figure 3B) that is similar to the likelihood plane in the baseline condition (Figure 2).

Mean Trends

Many psychological constructs change in a systematic manner over time and this also affects the equivalence of the BP and WP structures, as will be demonstrated in this section. As long as the model is identified, we can control for any parameter $\boldsymbol{\theta}_{tu}$ and/or $\boldsymbol{\theta}_{iu}$ when testing for conditional BP and WP structural equivalence. At present, however, there are certain software limitations. For example, in the present version of mkfm6 it is neither possible to impose more complicated parameter constraints, nor is it possible to estimate time-varying parameters. While this is possible with current SEM software, for $T \geq 80$ and with several thousand models to be estimated, SEM is no viable alternative for reasons outlined above. This is an active field of research and we are optimistic that future software and/or estimation procedures will overcome these limitations (e.g., see Molenaar, 1994; Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009; S.-M. Chow, Zu, Shifren, & Zhang, 2011 for work on time-varying parameters). For the time being and for the purpose of the present paper, however, a workaround is needed. Instead of simultaneously estimating all parameters, we propose to detrend all measures prior to testing for (conditional) equivalence. This two-step procedure is not optimal, but seems to work well in practice. This is true for simple mean trends (as will be shown in this section) as well as more complicated oscillating trends (next section). In addition, detrending (e.g., by means of local linear regression smoothing) avoids the need of introducing parameter constraints between the WP and BP structures, which may turn out to be nonlinear and hard to implement in practice.

To investigate the effects of trends on structural equivalence, we generated data in line with the baseline model but included a simple mean trend at the latent level by setting $\alpha_t = 0.05 \cdot (t - 1)$, with $t = 1,\ldots, T = 80$. This results in a mean increase in α by four units from the beginning to the end of each individual time series. As in the previous conditions, we then tested for unconditional equivalence of the BP and WP structures, followed by a test of conditional equivalence. For the latter, the mean trend was removed by local regression

(loess) smoothing with a span of .90 (Cleveland, Grosse, & Shyu, 1993). Due to the linear trend over time, the long smoothing span seems reasonable. As will be shown in the next example, more complicated changes over time may require shorter smoothing spans.

Average parameter estimates of the constrained condition (under the H_0) are presented in Table 2 and the resulting likelihood planes are shown in Figure 3C. As apparent from the left side of Figure 3C 90.23% of all likelihood ratio tests were significant. Thus, we would conclude that the BP and WP structure are not equivalent. What cannot be seen from the 2D-plane is that due to the linear increase in α, the resulting likelihood plane is no longer flat, but has a concave shape, with a little "green valley" in the middle (the green band in Figure 3C). This is due to the fact that the intraindividual mean (4 / 2 = 2) is compared to the sample mean at each occasion, which is only close to 2 in the middle of the time series. In contrast, when testing for conditional equivalence—that is, when controlling for the linear trend through loess smoothing—the number of significant likelihood ratio tests reduces to 0.93% (alpha level: 1%), and results in a likelihood plane (right side of Figure 3C) that is again similar to the likelihood plane in the baseline condition (Figure 2).

Cyclic Trends

The latter example can be easily extended to more complicated (e.g., cyclic) trends. For this purpose, we generated data in the same way as before, but let α oscillate with an angular frequency of $\omega = \frac{2\pi}{30}$, that is a period length of 30 days [i.e., $\frac{\mathrm{d}^2\alpha(t)}{\mathrm{d}t^2} = -\omega\alpha(t)$] and an amplitude of 0.5. We first tested for unconditional equivalence of the BP and WP structures, followed by a test of conditional equivalence. For the latter, the mean trend was removed by loess smoothing with a shorter span of 0.5.

Average parameter estimates of the constrained condition (under the H_0) are presented in Table 2 and the resulting likelihood planes are shown in Figure 3D. As apparent from the left side of Figure 3D, 85.30% of all likelihood ratio tests were nonsignificant. We can also

see the oscillating pattern, which is clearly reflected by the horizontal bands in the likelihood plane. With a period length of 30 days, and $T = 80$, we observe a little more than 5 such bands ("mountain ranges") across individuals ($80 / 30 \cdot 2 = 5.33$). In contrast, when testing for conditional equivalence—that is, when controlling for the oscillating trend through loess smoothing—the number of nonsignificant likelihood ratio tests increased to 98.34%[6] (alpha level: 1%), and results in a likelihood plane (right side of Figure 3D) that is again similar to the likelihood plane in the baseline condition (Figure 2).

An Empirical Example: Between- and Within-Person Differences in Attentiveness

Working with simulated data is a good way to illustrate an idea, demonstrate how it can be implemented in practice, and provide evidence that it works. However, the ultimate question is whether there are psychological constructs that exhibit (conditional) equivalence of BP and WP structures, and if so, to what degree. In our opinion, this is primarily an empirical question that depends on the construct, the sample of individuals, and the time period under consideration. Theoretical considerations on causes of variation should be informative on whether equivalence is more or less likely, yet the only way to find out about the degree of equivalence is to conduct (more) empirical studies that are capable of addressing this issue. Even though a full-fledged analysis is beyond the scope of this article, in the remainder of the paper we apply the ideas outlined above to the data of one such study: the COGITO study (Schmiedek, Lövden, & Lindenberger, 2010).

Procedure, Participants, and Measures

In the COGITO study, 101 younger adults (51 women; age: 20-31, $M = 25.6$, $SD = 2.7$)[7] practiced different tests of perceptual speed, working memory, or episodic memory over 100 daily sessions of about 1 hour each. In addition to the cognitive tests, various measures of affect, stress, and health were assessed. For a detailed description of the procedure, sample, and measurement instruments, we refer the reader to Schmiedek, Lövden, and Lindenberger

(2010). For the purpose of the present paper, we focus on one of the original content categories of the PANAS (Watson, Clark, & Tellegen, 1988; Zevon & Tellegen, 1982) that represents the factor *attentiveness*. It was assessed at the beginning of each session via the items "attentive", "interested", and "alert" on an 8-point rating scale from 0 (does not apply at all) to 7 (applies very well; cf. Brose, Lindenberger, & Schmiedek, 2013, April). A graphical illustration of the change and amount of fluctuations in individual and average attentiveness for men and women is given in Figure 4.

Analysis

A 1-factor model of attentiveness was specified with the factor loading matrix $\mathbf{\Lambda}^{\mathrm{T}} = (1.0 \quad \lambda_2 \quad \lambda_3)$, $\mathbf{\Phi} = (\phi)$, $\boldsymbol{\epsilon}^{\mathrm{T}} = (\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3)$, and intercepts $\boldsymbol{\alpha} = (0)$, $\boldsymbol{\tau}^{\mathrm{T}} = (\tau_1 \quad \tau_2 \quad \tau_3)$. Following the procedure outlined above, we first tested the (unconditional) between- and within-person structural equivalence of this model by comparing H_0: $\boldsymbol{\theta}_i = \boldsymbol{\theta}_t$ against H_1: $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_t$ with $\boldsymbol{\theta}_i = \{\lambda_{2,i},\ \lambda_{3,i},\ \phi_i, var(\varepsilon_{1,i}),\ var(\varepsilon_{2,i}),\ var(\varepsilon_{3,i}),\ \tau_{1,i},\ \tau_{2,i},\ \tau_{3,i}\}$ and $\boldsymbol{\theta}_t = \{\lambda_{2,t},\ \lambda_{3,t},\ \phi_t,\ \varepsilon_{1,t},\ \varepsilon_{2,t},\ \varepsilon_{3,t},\ \tau_{1,t},\ \tau_{2,t},\ \tau_{3,t}\}$ for each combination of *i* and *t*. That is, we conducted 10,100 likelihood ratio tests, each with 9 degrees of freedom. Note that the model is saturated under H_1.

Second, for each individual we removed the within-person mean, prior to carrying out the same test of (conditional) structural equivalence. As compared to loess smoothing this is a more conservative approach, because it does not account for trends, but only for the WP mean. However, just like loess smoothing at the within level, it affects not only the WP structure, but also the BP structure, by removing stable interindividual differences. In this sense the approach is one way to account for unobserved heterogeneity as discussed in the econometric and sociological literature (cf. Arellano, 2003; Halaby, 2004).

Third, in addition to demeaning we added a first-order autoregressive parameter to the within-structure model ($\boldsymbol{\theta}_{iu} = \{\beta\}$) in order to account for possible lagged effects, followed by a test of conditional structural equivalence.

Fourth, the model of step 3 was fitted separately for men and women in order to account for possible gender differences that may affect structural equivalence. In each group a test of conditional equivalence was carried out. As before, we used R version 2.15.0 (R Core Team, 2012) and mkfm6 (Dolan, 2010) for all statistical analyses.

Results

Median parameter estimates across all person-by-occasion combinations in each condition are presented in Table 3, and the resulting likelihood planes are shown in Figure 5. With a total of $2 \cdot 5 \cdot 101 \cdot 100 = 101{,}000$ state space models fitted to empirical data, inadmissible solutions and/or parameters at the boundary of the admissible parameter space are unavoidable. While we excluded inadmissible solutions[8], we did not remove parameters at the boundary. However, because the few "outliers" may have a strong and undue effect on the mean, we report the median of the parameter estimates instead.

As apparent from the 3-dimensional likelihood plane in Figure 5, with two exceptions (out of 10,100), not a single WP structure is unconditionally equivalent to a BP structure at any occasion *t*. This is exactly what we would expect for most psychological constructs and is further empirical evidence that a simple generalization from a BP structure to a WP structure and vice versa is not justified (Borsboom, Mellenbergh, & van Heerden, 2003; Molenaar, 2004; Molenaar & Campbell, 2009). Closer investigation of the likelihood plane in Figure 5 reveals that there is a systematic structure due to persons ("mountain ranges" parallel to the time axis), but not one due to occasions. This is because individuals were not assessed on the same (calendar) days in the COGITO study so that only the order of measurement occasions was considered.

When testing for *conditional* structural equivalence after removing the WP mean, 31.04% of likelihood ratio tests were nonsignificant (see Figure 6A). Nevertheless, the systematic pattern of vertical lines ("mountain ranges") in the likelihood plane shown in Figure 6A suggests the presence of additional person effects, such as lagged effects. This suspicion is confirmed by an increase of over 10% to a total of 42.24% nonsignificant likelihood ratios when an autoregressive process of order 1 was added to the WP model (Figure 6B). Interestingly, as the pattern of equivalent and nonequivalent structures becomes clearer, it also becomes obvious that the first occasion seems to differ in a systematic way from the other occasions (horizontal lines at the bottom of each panel in Figure 6). Follow-up analyses revealed that this nonequivalence is due to a heightened level of attentiveness indicated by most individuals during the first session of the study, which quickly dissipated in the following sessions.

In a last step, we conducted separate analyses for men and women (Figure 6C and D), resulting in another increase of about 10% to a total of 53.46% nonsignificant likelihood ratios for men (Figure 6C) and 54.29% nonsignificant likelihood ratios for women (Figure 6D).

Discussion of the Empirical Example

Controlling for some very basic factors like mean differences, autoregression, and gender differences, the BP factorial structure of attentiveness turned out to be indistinguishable from the WP structure in about 50% of all possible comparisons. This finding leaves us with two possible interpretations: First, given that we know the BP structure of attentiveness, we can be confident that it will be identical to the WP structure for a large portion of individuals, and vice versa. Second, given that we know the BP structure of attentiveness, we can be confident that it will *not* be identical to the WP structure for a large portion of individuals, and vice versa. As a matter of fact, this is the typical "the glass is half

full versus the glass is half empty" situation. For example, woman ID25 (marked by the first black arrow in Figure 6D) shows a factorial structure that is highly idiosyncratic. There is not a single occasion on which her WP structure of attentiveness is identical to the BP structure. This is also illustrated in Figure 7, which shows the density distribution of likelihood ratios for woman ID25. At an alpha level of 1%, not a single likelihood ratio is below the critical value. The estimated factor structure of ID25 is depicted at the top of Figure 7. Comparing the parameter estimates to the average (BP-WP) parameter estimates of women reported in Table 3, we find that the error variances of the first two indicators in particular are much higher for woman ID25. In contrast, with few exceptions, the factorial structure of woman ID48 (marked by the second black arrow in Figure 6D) is indistinguishable from the BP structure on almost all occasions and thereby close to the density distribution under the null hypothesis of perfect structural equivalence (black curve in Figure 7). Even though this is just a single empirical example—and the situation may be quite different for other constructs—it is prototypical for our general expectations: Under realistic conditions, we cannot expect the BP structure to be identical to the WP structure for all possible combinations of persons and occasions. That being said, for many psychological constructs, we would not expect the BP and WP structure to be independent either. Rather, some very basic factors (such as group differences or serial dependencies) may obscure the commonalities of BP and WP structures, and it is up to the researcher to identify and control for these factors.

Overall Discussion and Conclusions

In a world without constraints on time and money, the topic of the present paper would be inconsequential. In such a world, any individual could be assessed at any occasion, including the past and the future. If one were interested in a particular person, one would simply analyze his or her WP structure, if one were interested in relationships between variables across individuals, one would analyze the BP structure. Reality is different. In our

world, researchers are forced to trade the number of individuals against the number of time points. Either a few individuals are extensively assessed over time, or many individuals are observed at one (or few) occasion. This raises the question whether a structure observed by means of BP analysis generalizes to the WP structure and vice versa. Usually, such equivalence—which we referred to as unconditional equivalence—is violated, for reasons that are at the core of the phenomenon of interest (Blalock, 1967; Borsboom et al., 2009; Kuh, 1959; Molenaar, 2004; Molenaar & Campbell, 2009).

Unconditional nonequivalence, however, should not be confused with *independence*. Unconditional nonequivalence of WP and BP structures does not imply that these structures are orthogonal and that nothing can be learned from one about the other. As discussed before, lack of structural equivalence is either due to persons (lack of homogeneity), due to time (lack of stationarity), or due to combinations of both. The factors that affect homogeneity or stationarity, however, may often be known (e.g., lack of stationarity due to mean trends caused by a learning process, or lack of homogeneity due to gender differences) and can be controlled for. If the BP and WP structures are identical after controlling for factors that are unique to either structure, we speak of *conditional equivalence*.

In the present paper we argued that instead of blindly assuming BP and WP structural equivalence (which will hardly ever be met in practice), or focusing either on the analysis of the BP *or* the WP structure in the presumed absence of a relationship between the two, it may be worthwhile to explore their commonalities in conjunction with their differences. To this end we have proposed the construction of 2 and 3-dimensional likelihood planes as a tool for inspecting BP-WP structural equivalence and deviations thereof. This provides insights into (1) the amount of equivalence (red to green ratio) and (2) possible reasons for nonequivalence (i.e., systematic patterns due to persons, occasions, or combinations thereof). Using simulated data, we examined four common reasons for structural nonequivalence and demonstrated

how to restore conditional equivalence. All analyses were carried out within a state space modeling framework using the Kalman filter. Finally, an empirical example was provided based on a recent study in which 101 individuals were repeatedly assessed over 100 days. As expected, for almost all possible combinations of persons and occasions, the BP and WP structures of attentiveness differed significantly from each other when testing for unconditional equivalence. However, after demeaning, controlling for autoregression and for gender differences, the number of conditionally nonequivalent BP and WP structures reduced to less than 50%.

Traits Versus States

In most of the existing literature, the discussion on BP versus WP analyses is inherently confounded with the discussion on traits and states. We propose to disentangle this discussion by defining a trait as the variance that is *unique* to the BP structure ($\boldsymbol{\theta}_{tu} = \{\sigma^2_{between\_trait}\}$ for all *t*). This implies that the variance must be *caused* by differences between people and not merely *reflect* differences between people. This distinction is subtle but important. Variability (e.g., $\boldsymbol{\theta}_t = \{\sigma^2_{between}\}$) at any time point *t* is comprised of two different sources of variance: trait variance and state variance ($\sigma^2_{between} = \sigma^2_{between\_trait} + \sigma^2_{between\_state}$)[9]. Both reflect differences between people. The more trait-like a construct, the larger the proportion of $\sigma^2_{between\_trait}$ in $\sigma^2_{between}$. By comparing $\boldsymbol{\theta}_t = \{\sigma^2_{between}\}$ to $\boldsymbol{\theta}_i = \{\sigma^2_{within}\}$ in a test of unconditional structural equivalence, we are comparing two sources of variance that—ceteris paribus—are by definition increasingly caused by different factors, the more trait-like the construct. Thus, it is not surprising that the most stable (trait-like) constructs in differential psychology (intelligence, personality factors) are also the ones least likely to exhibit unconditional structural equivalence. It is important to note, however, that this is circular reasoning: If we define a trait-like construct as a construct with a high proportion of $\sigma^2_{between\_trait}$ to $\sigma^2_{between\_state}$ and if we define unconditional structural

equivalence as the equivalence of $\sigma^2_{between}$ and σ^2_{within}, by definition a trait-like construct then results in structural nonequivalence.

Instead, it seems more insightful to attempt to disentangle $\boldsymbol{\theta}_{tc} = \{\sigma^2_{between\_state}\}$ and $\boldsymbol{\theta}_{tu} = \{\sigma^2_{between\_trait}\}$ by accounting for factors that are known to affect only the BP structure, but not the WP structure before testing for *conditional* equivalence. The advantage of this approach is that (1) its outcome is not determined by its definition, and (2) it forces us to identify the factors that *cause* differences between persons. To this end, theory is of utmost importance because it determines whether a factor affects only the BP structure but not the WP structure (for example, gender could be one such factor, as it generally does not change within persons). Unfortunately, this approach is no panacea. While it seems possible to theoretically derive factors that affect the BP structure but not the WP structure, it is less clear that this is possible when it comes to explaining BP state versus trait variance (i.e., $\sigma^2_{between\_state}$ vs. $\sigma^2_{between\_trait}$). For example, the approach of controlling for factors affecting the BP structure is reduced to absurdity if we attempt to control for BP differences at any occasion *t* by the individual genetic make-up. Given that all people are genetically distinct, controlling for individual differences in genes would "explain" the entire BP variance ($\sigma^2_{between}$) *at this occasion*, without allowing for a more detailed distinction between $\sigma^2_{between\_state}$ vs. $\sigma^2_{between\_trait}$ and would thus render any further tests of (conditional and unconditional) structural equivalence meaningless. This is tantamount to saying that BP variance exists because people differ, which is true but not a very illuminating perspective on the issue at hand. In contrast, if prior theory allows the selection of (few) candidate genes that are known to produce stable BP differences, this may help to control for these differences prior to testing for conditional structural equivalence.

Note that the same considerations apply to a trait factor that has been derived by some sort of weighted averaging over time. Here the only difference is that the ratio of

$\sigma^2_{between\_trait}$ to $\sigma^2_{between}$ is close to 1 from the very beginning. In a test of unconditional equivalence, σ^2_{within} and $\sigma^2_{between}$ may or may not be the same, with the chances that they are the same not being high for reasons outlined before. The meaning of such a comparison, however, remains unclear. This is readily apparent if we accept the definition of a trait as the variance that is unique to the BP structure. In this case, if $\frac{\sigma^2_{between_{trait}}}{\sigma^2_{between}} = 1$ there is no variance left for a comparison to the within-structure, so that a test of conditional equivalence is rendered impossible.

Finally, if we conceive of a trait as something that is stable within but varies between individuals, the analog at the WP level is variation across time but constancy across individuals. Following our definition of a trait, we propose to conceive of such a "time-trait" as the variance that is *unique* to the within structure ($\boldsymbol{\theta}_{tu} = \{\sigma^2_{within\_trait}\}$ for all *i*). All other considerations regarding the "person-trait" apply equally to the "time-trait."

Limitations, Future Directions, and Practical Implications

At some level the present paper may have raised more questions, for both methodologists and substantive researchers, than it has answered. However, rather than offering some conclusive answers, a major goal of our work was to outline a unified framework for the study of BP and WP structures within which these questions can be pursued in future research.

At the methodological level, an obvious limitation of the approach to testing for (conditional) structural equivalence is its implementation by using currently available software. Ideally, all parameters that ought to be controlled for ($\boldsymbol{\theta}_{iu}, \boldsymbol{\theta}_{tu}$) should also be part of the actual model and should be estimated simultaneously with all other parameters ($\boldsymbol{\theta}_{ic}, \boldsymbol{\theta}_{tc}$). This, however, requires more complicated parameter constraints across groups (other than simple equality constraints), which is not possible in mkfm6 (Dolan, 2010), which

was used in the present paper (see also Footnote 5). While SEMs offer more flexibility in this regard, they are limited in the number of time points. Hence, we resorted to some workarounds in this article, most importantly the two-step procedure of controlling for individual trends by means of loess smoothing in the simulation study prior to the actual test of conditional equivalence. Unfortunately, this introduces additional problems like the choice of optimal span parameters to avoid under- or oversmoothing. Thus, from a theoretical perspective, this workaround is suboptimal. However, it seems to work well in practice and appears to be a good intermediate solution until more efficient software, optimizers, or both have been developed.

Another technical detail that deserves closer attention in future research is the choice of the initial covariance matrix $\mathbf{P}_{0|0}$ and its effect on parameter estimates—for example as compared to SEM—for short time series. In the spirit of the present article, we suggest that it may be promising to use information from between-subject analyses to initialize the estimation of a within-person process. We are currently exploring this option in greater detail.

Furthermore, in the present paper we only considered some basic mechanisms that may lead to structural nonequivalence and how to account for them. In particular, we did not discuss situations of multiple, possibly unknown, group memberships and how recursive partitioning and/or finite mixture modeling may help to establish conditional equivalence under these more complex conditions. The approach also offers the possibility to remove deviant individuals and/or occasions in an iterative fashion in order to maximize (conditional) equivalence in an exploratory (non-theory driven) way. In contrast to previous studies the removal would not be limited to only individuals or only occasions, but could be an arbitrary mixture between the two, depending on the relative contribution to the overall likelihood plane. We consider these to be important and promising directions for future research.

At a more substantive level, it seems possible that—at least theoretically—the causal factors underlying the BP structure are *different* from the causal factors underlying the WP structure, but that the BP and WP structures are nevertheless equivalent. That is, nonequivalence of causes does not necessarily imply nonequivalence of effects (i.e., the resulting BP and WP structures). In that case, a test of conditional equivalence may result in a higher amount of significantly different WP-BP comparisons than a test of unconditional equivalence.

Furthermore, in the present paper we did not consider longitudinal BP structures, that is, BP structures at more than one time point. We also did not consider models with time-varying parameters (e.g., time-varying factor loadings). Likewise, the possible causes for structural nonequivalence studied in the present paper are limited. For example, we did not discuss subject specific structural models in which the number of factors may differ from individual to individual. While the likelihood plane may help to identify individuals whose WP structure differs markedly from the BP structure, there may be a multitude of possible reasons for such differences, only few of which were discussed in this work.

Although speculative at this early state of affairs, we ultimately see three general ways in which this work may have practical implications for methodologists and substantive researchers: (1) new insights into psychological mechanisms and improvements in measurement and construct validity; (2) improvements in diagnostics and intervention; (3) methodological improvements. Regarding the first aspect (1), at present little is known about the degree to which some of the most well-established constructs in psychology generalize to the individual. By being able to explicitly test the BP-WP equivalence of established constructs, we can test for construct validity at different levels and improve the design of person-oriented measurement instruments. The approach may also offer new insights into psychological mechanisms by studying why (groups of) individuals exhibit high or low

structural equivalence and how this depends on the time period under consideration. For example, Brose, Voelkle, Lövdén, Lindenberger, and Schmiedek (submitted) demonstrated that the degree of divergence between WP and BP structure of affect could be reliably predicted by contextual variables such as certain aspects of well-being. In addition (2) the approach may have direct implications for diagnostics and intervention. If the relationship between variables at the BP level is known and we find the same at the WP level, then we can use variations observed at the BP level to predict variations at the WP level. For example, if we find that 10 out of 100 patients show improvements in momentary physical health when taking a certain drug, it may seem reasonable to prescribe this drug A to any individual *i*. If, however, we have evidence that the BP relationship ($r_{\text{drug, improvements in physical health}}$) does not hold, or is reversed, for a given person (indicated by a red mountain range in the likelihood surface), then the recommendation to use this drug would be ill advised. Needless to say that in modern health care provision it makes a tremendous difference whether WP-BP equivalence holds for 99%, 50%, or 1% of individuals. Finally (3), we hope the present paper spurs more research on the development of better methods for the integration of WP time series (*T* large, *N* small) analysis with BP statistics (*N* large, *T* small).

Equivalence of Between- and Within-Person Structures as an Empirical and Continuous Phenomenon

The dream of integrating intra- and interindividual differences research is an old one that has existed for many years and comes in different varieties (cf. Cronbach, 1957, 1975). However, as recently suggested by Borsboom et al. (2009) "a dreamed route of progress […may be] really a dead end street" (p. 94), and the gap between the two research paradigms "may very well be here to stay" (Borsboom et al., 2009, p. 92). Based on our findings on unconditional nonequivalence of BP and WP structures we are inclined to agree with this "gloomy conclusion" (Borsboom et al., 2009, p. 94). However, in the present article we have

demonstrated that even though unconditional equivalence of a BP and WP structure is unlikely for most psychological constructs, controlling for some simple factors that are known to affect either the BP or WP structure may result in a considerable degree of conditional equivalence. This is not (yet) the salutary integration of two different research paradigms, but it is a bridge over the gap that separates the two. At present the bridge is quite shaky, but it allows researchers on BP differences to explore the territory of WP research and vice versa.

Given that more than one century of psychological research has focused almost exclusively on the analysis of BP structures, even a small degree of conditional equivalence would be good news, as it would suggest that the findings of previous studies may offer some guidance in the development of person-oriented research, interventions, and theory. Clearly, the behavioral sciences need to develop a better understanding of the differences and commonalities between the two research paradigms and this article is one potential initial step in this direction. Ultimately, however, it is up to future research to study the *degree* of structural equivalence which may vary considerably from construct to construct. To this end, complete independence and unconditional equivalence of BP and WP structures may be viewed as two endpoints on a continuum, or as two continents separated by a large rift. Whether crossing the rift pays off and eventually transforms the shaky bridge into a solid highway or whether the dangers and efforts of crossing the rift outweigh possible gains, so that the bridge remains shaky and unused, is an empirical question. The future will tell.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Arellano, M. (2003). *Panel data econometrics*: Oxford University Press.

Blalock, H. M. (1967). Causal inferences, closed populations, and measures of association. *American political science review, 61*(1), 130-136.

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

Borsboom, D., & Dolan, C. V. (2006). Why g is not an adaptation: A comment on Kanazawa (2004). *Psychological Review, 113*(2), 433-437. doi: 10.1037/0033-295x.113.2.433

Borsboom, D., & Dolan, C. V. (2007). Commentary: Theoretical equivalence, measurement invariance, and the idiographic filter. *Measurement, 5*, 236-243. doi: 10.1080/15366360701765020

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 67-97). New York: Springer.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203-219.

Brose, A., Lindenberger, U., & Schmiedek, F. (2013, April). Affective States Contribute to Trait Reports of Affective Well-Being. *Emotion*, Advance online publication. doi: 10.1037/a0032401

Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (submitted). Positive and negative affect: Between-person associations do not generalize to the individual.

Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of autoregressive moving average time series models. In A. Maydeu & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 415-452). Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate* (Vol. 3). New York: Brunner/Mazel.

Cattell, R. B. (1946). *Description and measurement of personality*. Yonkers-on-Hudson, New York: World Book Co.

Cattell, R. B. (Ed.). (1966). *Handbook of multivariate experimental psychology*. Chicago: Rand McNally.

Cattell, R. B., Cattell, A. K. S., & Rhymer, R. M. (1947). P-technique demonstrated in determining psychophysical source traits in a normal individual. *Psychometrika, 12*, 267-288. doi: 10.1007/BF02288941

Chow, S.-M., Zu, J., Shifren, K., & Zhang, G. (2011). Dynamic Factor Analysis Models With Time-Varying Parameters. *Multivariate Behavioral Research, 46*(2), 303-339. doi: 10.1080/00273171.2011.563697

Chow, S. M., Ho, M.-h., Hamaker, E. L., & Dolan, C. V. (2010). Equivalence and differences between structural equation modeling and state-space modeling techniques. *Structural Equation Modeling, 17*, 303-332. doi: 10.1080/10705511003661553

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1993). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 309-376). New York: Chapman & Hall.

Commandeur, J. J. F., & Koopman, S. J. (2007). *An introduction to state space time series analysis*. New York: Oxford University Press.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L. J. (1975). Beyond the two disciplines of psychology. *American Psychologist, 30*, 116-127.

Dolan, C. V. (2010). *MKFM6 Multi-group, multi-subject stationary time series modeling based on the Kalman filter*.

Durbin, J., & Koopman, S. J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 128-141.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage, 63*(1), 310-319. doi: http://dx.doi.org/10.1016/j.neuroimage.2012.06.026

Gill, P. E., Murray, W., Saunders, M. A., & Wright, M. H. (1998). User's guide for NPSOL 5.0: A Fortran package for nonlinear programming (Technical Report SOL 86-2, Revised July 30, 1998). Stanford, CA: Systems Optimization Laboratory, Department of Operations Research, Stanford University.

Halaby, C., N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*(30), 507-544.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2003). ARMA-based SEM when the number of time points T exceeds the number of cases N: Raw data maximum likelihood. *Structural Equation Modeling, 10*(3), 352-379.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate Behavioral Research, 40*(2), 207-233.

Harvey, A. C. (2001). *Forecasting, structural time series models and the Kalman filter*. Cambridge, UK: Cambridge University Press.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering, 82 (Series D)*, 35-45.

Kenny, D. A. (1974). A quasi-experimental approach to assessing treatment effects in nonequivalent control group designs. *Psychological Bulletin, 82*, 345-362.

Kuh, E. (1959). The validity of cross-sectionally estimated behavior equations in time series applications. *Econometrica, 27*(2), 197-214.

Lebo, M. A., & Nesselroade, J. R. (1978). Intraindividual differences dimensions of mood change during pregnancy identified in five P-technique factor analyses. *Journal of Research in Personality, 12*(2), 205-224. doi: 10.1016/0092-6566(78)90098-3

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*(2), 181-202. doi: 10.1007/BF02294246

Molenaar, P. C. M. (1994). Dynamic latent variable models in developmental psychology. In A. von Eye & C. C. Clogg (Eds.), *Analysis of latent variables in developmental research* (pp. 155-180). Newbury Park, CA: Sage.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201-218.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychology, 18*(2), 112-117. doi: 10.1111/j.1467-8721.2009.01619.x

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of Developmental Systems Theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development: Dialogues with life-span psychology* (pp. 339-360). Dordrecht, NL: Kluwer.

Molenaar, P. C. M., & Nesselroade, J. R. (2009). The recoverability of P-technique factor analysis. *Multivariate Behavioral Research, 44*(1), 130-141. doi: 10.1080/00273170802620204

Molenaar, P. C. M., Sinclair, K. O., Rovine, M. J., Ram, N., & Corneal, S. E. (2009). Analyzing Developmental Processes on an Individual Level Using Nonstationary Time Series Modeling. *Developmental Psychology, 45*(1), 260-271.

Nesselroade, J. R., McArdle, J. J., Aggen, S. H., & Meyers, J. M. (2002). Dynamic factor analysis models for representing process in multivariate time-series. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 235-265). Mahwah, NJ: Lawrence Erlbaum Associates.

Otter, P. W. (1986). Dynamic structural systems under indirect observation: Identifiability and estimation aspects from a system theoretic perspective. *Psychometrika, 51*(3), 415-428. doi: 10.1007/BF02294064

Oud, J. H. L., & Singer, H. (2008). Continuous time modeling of panel data: SEM versus filter techniques. *Statistica Neerlandica, 62*(1), 4-28.

Oud, J. H. L., van den Bercken, J. H., & Essers, R. J. (1990). Longitudinal factor score estimation using the Kalman filter. *Applied Psychological Measurement, 14*, 395-418.

R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Schmiedek, F., Lövden, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience, 2*(27), 1-10. doi: 10.3389/fnagi.2010.00027

Shumway, R. H., & Stoffer, D. S. (2004). *Time series analysis and its applications*. New York: Springer.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Voelkle, M. C., Oud, J. H. L., von Oertzen, T., & Lindenberger, U. (2012). Maximum likelihood dynamic factor modeling for arbitrary N and T using SEM. *Structural Equation Modeling, 19*(3), 329-350. doi: 10.1080/10705511.2012.687656

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology, 43*(1), 111-122. doi: 10.1037/0022-3514.43.1.111

Zhang, Z., Hamaker, E. L., & Nesselroade, J. R. (2008). Comparisons of four methods for estimating a dynamic factor model. *Structural Equation Modeling, 15*(3), 377-402.

Footnotes

[1]Let alone the thousands of possible combinations (and associated research designs) that result from the comprehensive 10-dimensional coordinates of the basic data relation matrix.

[2]T indicates the transpose of a vector or matrix.

[3]To keep the notation as simple as possible, we ignore the fact that, due to missings, the number of manifest variables (and as a consequence the size of the model-implied covariance matrix and mean vector) may differ across individuals (Arbuckle, 1996; Enders, 2001, 2010). Likewise, we ignore that the number of time points may differ across individuals (i.e., T_i). Both, however, are only ignored in notation but not in the actual estimation.

[4]For example, the estimation of a joint BP ($N = 99$; $T = 1$) and WP structure ($N = 1$; $T = 80$) by means of the Kalman filter (mkfm6) took about 1.3 seconds on a standard 2.4GHz processor personal computer (not including the time for data generation or plotting). In contrast, the estimation by means of SEM (a model as described by Voelkle et al., 2012) took about 16.6 seconds, which is more than 12 times as long. For $100 \cdot 80 = 8{,}000$ comparisons this amounts to a difference in computation time of more than a day.

[5]Unfortunately, apart from equality constraints, mkfm6 does not (yet) allow more complicated constraints between groups. Thus, it was not possible to implement the true relationship $\psi_{\text{within}} = \phi_{\text{between}}(1 - \beta^2)$ when testing for conditional equivalence. Instead, even under the null hypothesis, ψ (WP) and ϕ (BP) were freely estimated. This reduces the degrees of freedom of the resulting likelihood ratio test to $df_{\text{Diff}} = 6$ instead of 7 (to 8 instead of 9 in Table 3, respectively). Note, however, that this is solely a limitation of the software used and results in a test that is somewhat less powerful than it could be. For the purpose of the present paper this seems negligible.

[6]As discussed in the text, the use of loess smoothing is suboptimal from a theoretical point of view because the number of (non)significant LR tests, as well as the parameter estimates, are affected by the degree of smoothing. Also, while in this example the trend was induced at the latent level, the smoothing took place at the level of the manifest variables. Even if H_0 is true, there is no theoretical basis to expect 99% nonsignificant LR tests after having accounted for an unknown trend by means of loess smoothing. From a practical point of view, however, loess is easy to implement, easy to interpret, and seems to work well as demonstrated in this example. As with any smoothing procedure, however, it is critical to find the right trade-off between fitting noise by over-smoothing or failing to account for systematic trends. In the present example the reduction in error variance suggests that the smoothing parameter (span = 0.50) may have been too low, while visual inspection suggested an even lower span.

[7]In addition, 103 older adults (49.5% women; age: 65-80, $M = 71.3$, $SD = 4.1$), participated in the COGITO study. They will not be considered in this article.

[8]We took a conservative approach to identifying and removing invalid solutions by excluding any model for which the optimizer (NPSOL) employed by mkfm6 returned an error message. This includes inform = 1 error messages, which were comparatively frequent but are usually unproblematic. For details, the reader is referred to Gill, Murray, Saunders, and Wright (1998; p. 36). In particular, 1.01% of models were excluded in the baseline condition (unconditional structural equivalence), 0.04% in condition A (demeaning), 2.08% in condition B (autoregression), 3.47% for men (C), and 0.67% for women (D). Invalid solutions are represented by blanks ("holes") in the 2- and 3-dimensional likelihood planes.

[9]For reasons of simplicity we ignore measurement error variance.

Table 1

Simulation Conditions

Parameter	**Population Values in Condition**								
	Baseline	Autoregression		Group differences		Mean trend		Cyclic trend	
		BP	*WP*	*BP*	*WP*	*BP*	*WP*	*BP*	*WP*
β	•	--	0.8	•	•	•	•	•	•
$Var(\eta_{g1})$	•	•	•	1.0	--	•	•	•	•
$Var(\eta_{g2})$	•	•	•	4.0	--	•	•	•	•
Linear slope	•	•	•	•	•	--	0.05	•	•
ω	•	•	•	•	•	•	•	--	$2\pi/30$
Df_{Diff}	7	6[a]		7		7		7	

Note. ● = parameter neither simulated nor estimated; -- = parameter not available in this group, but simulated in the corresponding between/within group. BP = between-person; WP = within-person; Df_{Diff} = degrees of freedom of the likelihood ratio test that compares the between structure to the within structure.

[a] See text and Footnote 5 for details.

Table 2

Average Parameter Estimates Across 8,000 Person Time-Point Combinations Under the Assumption (H_0) of Conditional Equivalence

Parameter	**Average Parameter Estimate**								
	Baseline	Autoregression		Group differences		Mean trend		Cyclic trend	
		BP	*WP*	*BP*	*WP*	*BP*	*WP*	*BP*	*WP*
λ_2	0.804	0.804		0.799		0.797		0.803	
λ_3	0.801	0.810		0.796		0.802		0.805	
$Var(\varepsilon_1)$	0.196	0.190		0.201		0.181		0.090	
$Var(\varepsilon_2)$	0.205	0.199		0.208		0.187		0.089	
$Var(\varepsilon_3)$	0.196	0.213		0.202		0.188		0.095	
$\phi \mid\mid \psi$	0.996	0.959	0.356	2.443[a]		0.979		0.919	
$M(\eta) = \alpha$	–0.010	–0.002		0.017		–0.005		–0.021	
β	•	--	0.758	•	•	•	•	•	•
$Var(\eta_{g1})$	•	•	•	0.986	--	•	•	•	•
$Var(\eta_{g2})$	•	•	•	3.901	--	•	•	•	•
Linear slope	•	•	•	•	•	--	loess	•	•
ω	•	•	•	•	•	•	•	--	loess

Note. • = parameter neither simulated nor estimated; -- = parameter not available in this group, but simulated and estimated—corrected for—in the corresponding between/within group; BP = between-person; WP = within-person; loess = local regression smoothing.

[a] Estimated variance constrained to equality across groups. See text for details.

Table 3

Median Parameter Estimates Across 10,100 Person Time-Point Combinations Under the Assumption (H_0) of Conditional Equivalence.[a]

Parameter	**Average Parameter Estimate**									
	Baseline	Mean removed		Autoregression		Men (N = 49)		Women (N = 51)	
		BP	*WP*	*WP*	*WP*	*WP*	*WP*	*BP*	*WP*
λ_2	0.799	0.918		0.938		0.938		0.944	
λ_3	0.943	0.842		0.932		0.833		0.957	
$Var(\varepsilon_1)$	0.273	0.377		0.342		0.342		0.374	
$Var(\varepsilon_2)$	0.835	0.652		0.603		0.530		0.631	
$Var(\varepsilon_3)$	0.972	0.585		0.532		0.540		0.512	
τ_1	3.828	–0.019		–0.006		–0.007		–0.005	
τ_2	3.875	–0.018		–0.004		–0.004		–0.005	
τ_3	3.668	–0.032		–0.006		–0.006		–0.006	
β	•	•	•	--	0.397	--	0.398	--	0.367
$\phi \mid\mid \psi$	1.734	0.520		0.505	0.307[a)]	0.466	0.225[a)]	0.608	0.321[a)]
Df_{Diff}	9	9		8[b]		8[b]		8[b]	

Note. • = parameter not estimated; -- = parameter not available in this group, but estimated—corrected for—in the corresponding within group; BP = between-person; WP = within-person.

[a] The five conditions (five columns) are cumulative, beginning with a test of unconditional equivalence (baseline),followed by a test of conditional equivalence controlling for mean differences, controlling for mean differences and autoregression, and finally controlling for mean differences, autoregression, and age group differences by fitting two separate models for men and women.

[b] See text and Footnote 5 for details.

Figure Captions

Figure 1. Schematic illustration of a between-person factor model for $i = 1,\ldots, N$ persons at a single occasion (left side) and a within-person factor model of a single person at $t = 1,\ldots, T$ occasions (right side). A: All parameters of the two models are equivalent. B: Nonequivalence due to a single person. For example, the "trembling" of a patient with Parkinson disease may not be indicative of anxiety. C: Possible nonequivalence due to time of assessment. For example, on hot summer days, sweating may not be indicative of anxiety. The two structures may only be equivalent if the occasion at which the BP structure was assessed is representative for the time period during which the WP structure was assessed. D: When ignoring a grouping factor (here: gender differences in variance), it is possible that not a single WP structure (variance either 0.5 or 1.5) is equivalent to the average BP structure (variance 1.0).

Figure 2. Three-dimensional likelihood plane. For each person $i = 1,\ldots, N$ (x-axis) and occasion $t = 1,\ldots, T$ (z-axis) combination, a likelihood ratio test is conducted by subtracting $2LL_{H0}$ of the model under the null hypothesis of BP-WP equivalence from $2LL_{H1}$ of the model under the alternative hypothesis of no equivalence. The resulting likelihood ratios $[2LL_{H1} - 2LL_{H0}]$ are given on the y-axis. For each likelihood ratio test the color indicates whether it is significant (red) or not (green) at an alpha level of 1%.

Figure 3. Four common reasons for structural nonequivalence illustrated by eight 2-dimensional likelihood planes. Significant likelihood ratio values are red, nonsignificant values green (alpha 1%.). Left: Test for unconditional equivalence.

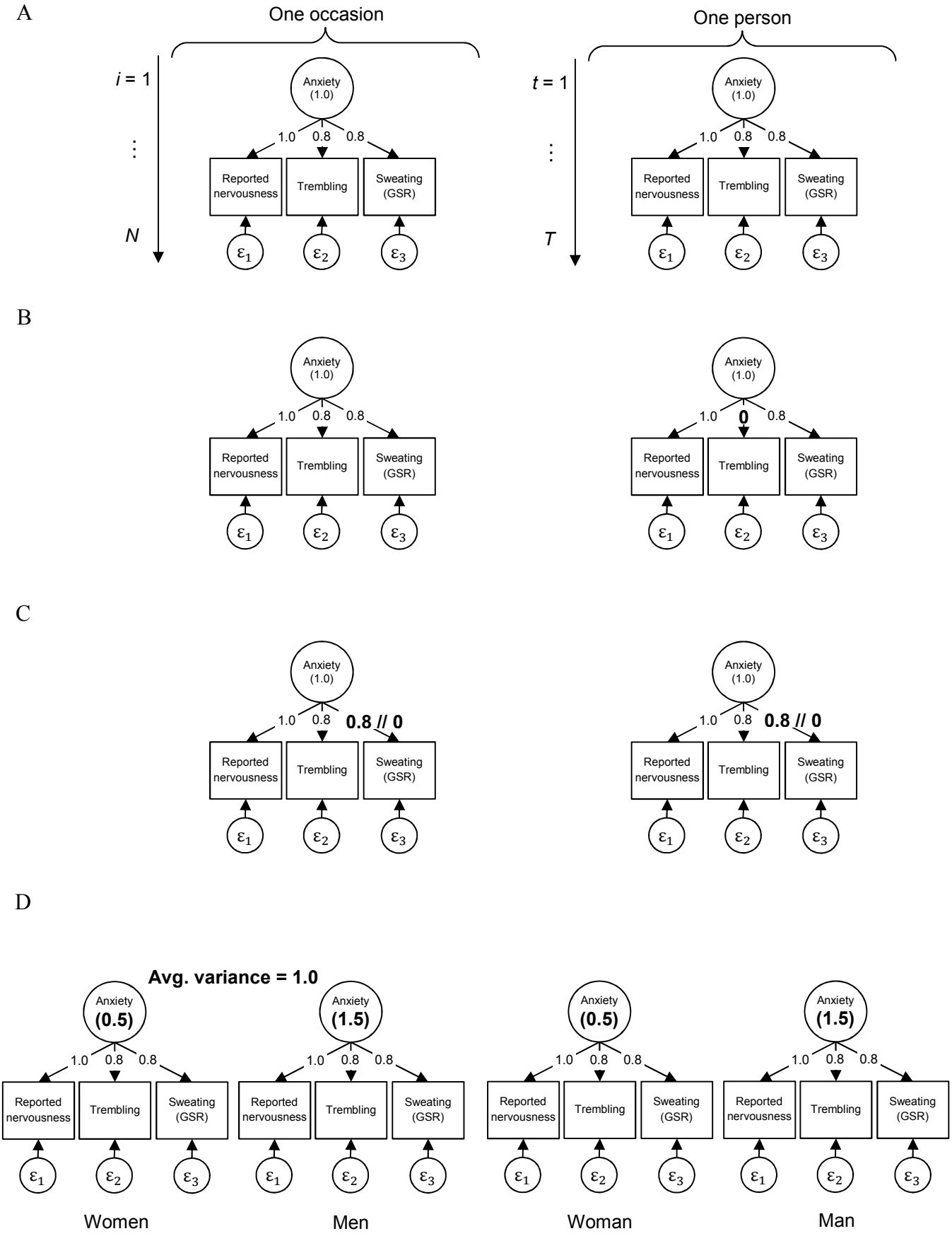
Right: Test for conditional equivalence. A: Autoregression, B: Group difference, C: Linear trend, D: Cyclic trend.

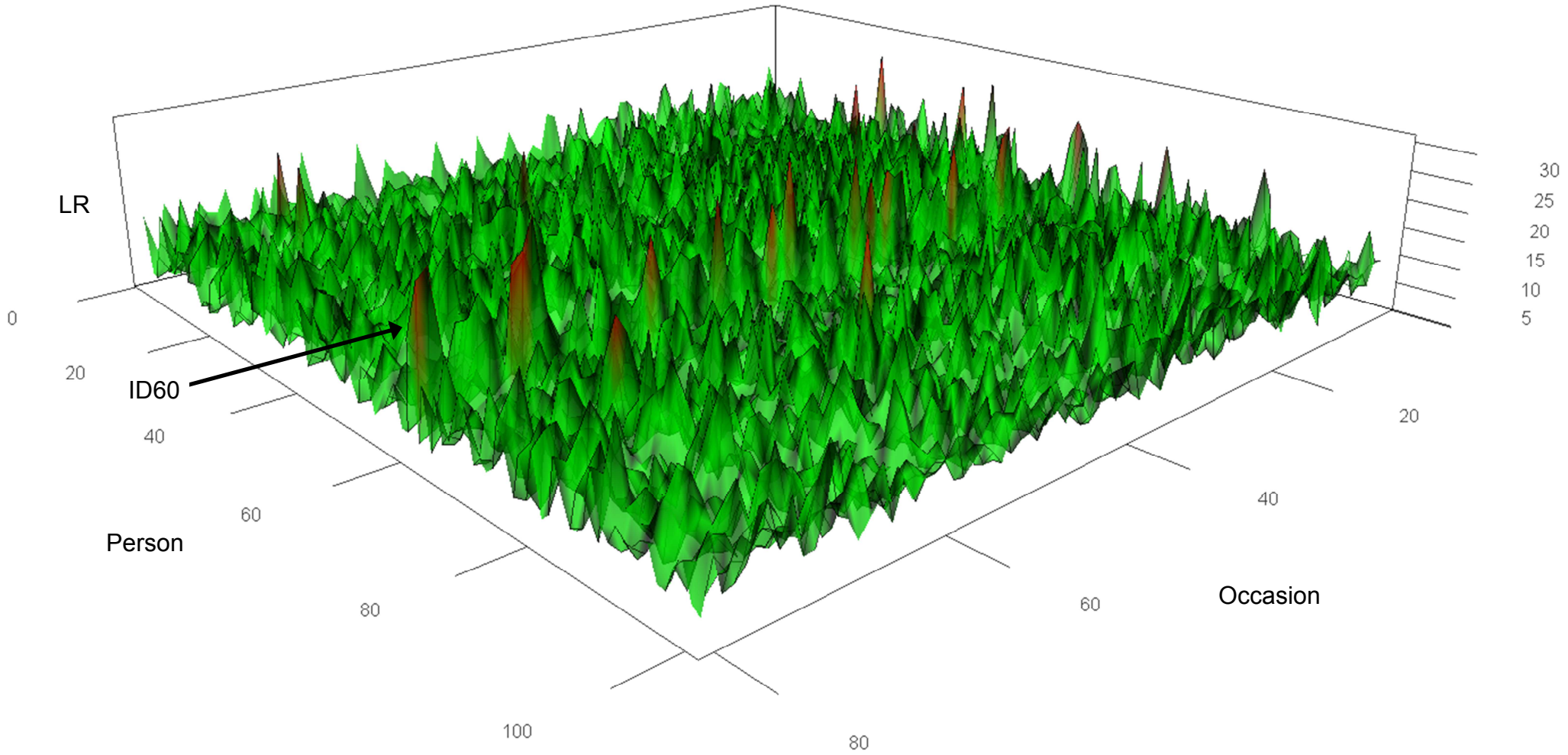
Figure 4 Average attentiveness across 100 days for men and women. Individual trajectories are depicted in grey.

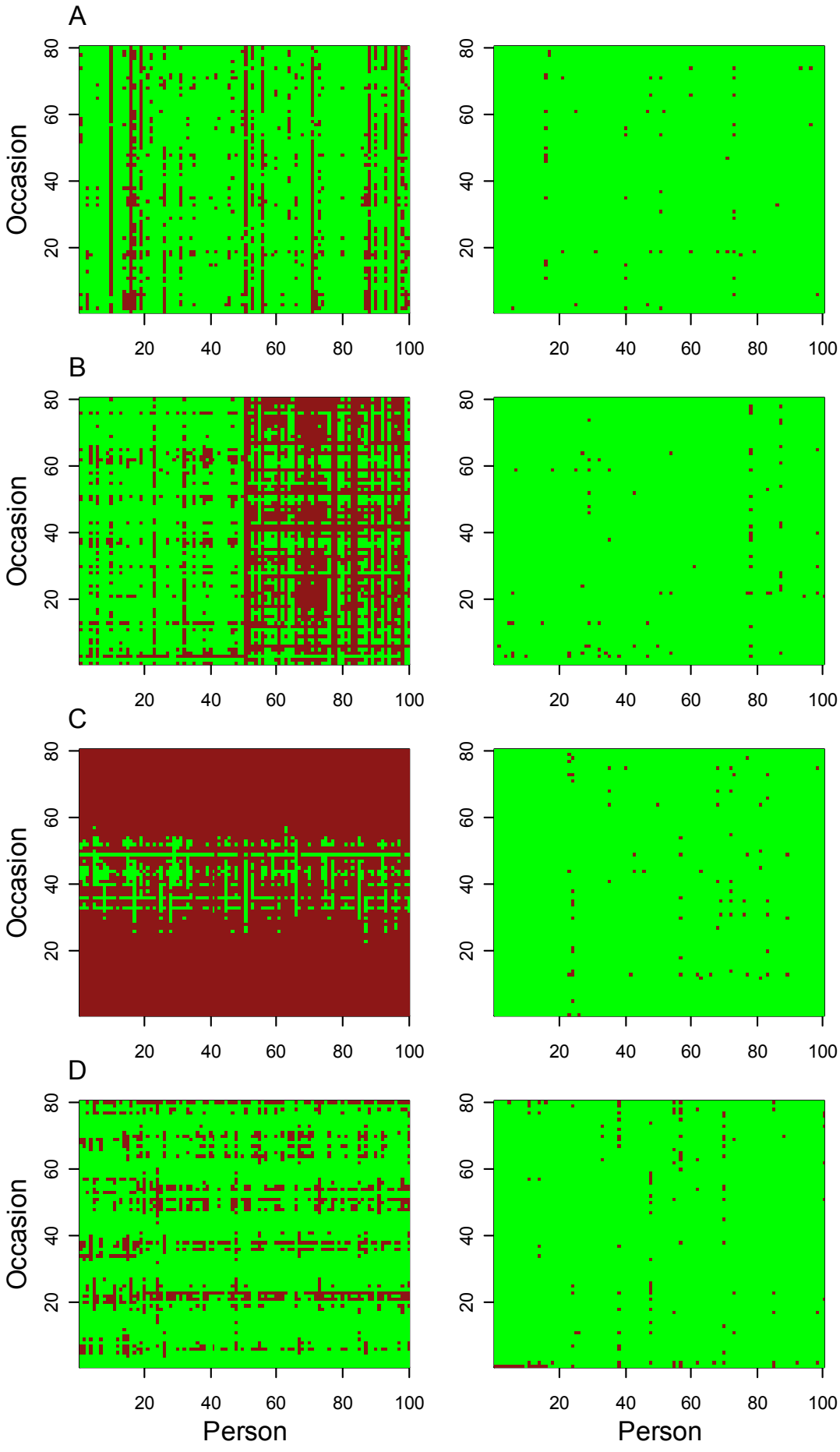
Figure 5. Three-dimensional likelihood plane resulting from tests for unconditional equivalence of the BP and WP structures in attentiveness. Only 2 out of (T = 100) × (N = 101) = 10,100 likelihood ratio tests are nonsignificant (green dots), demonstrating the clear difference of BP and WP structures (alpha 1%).

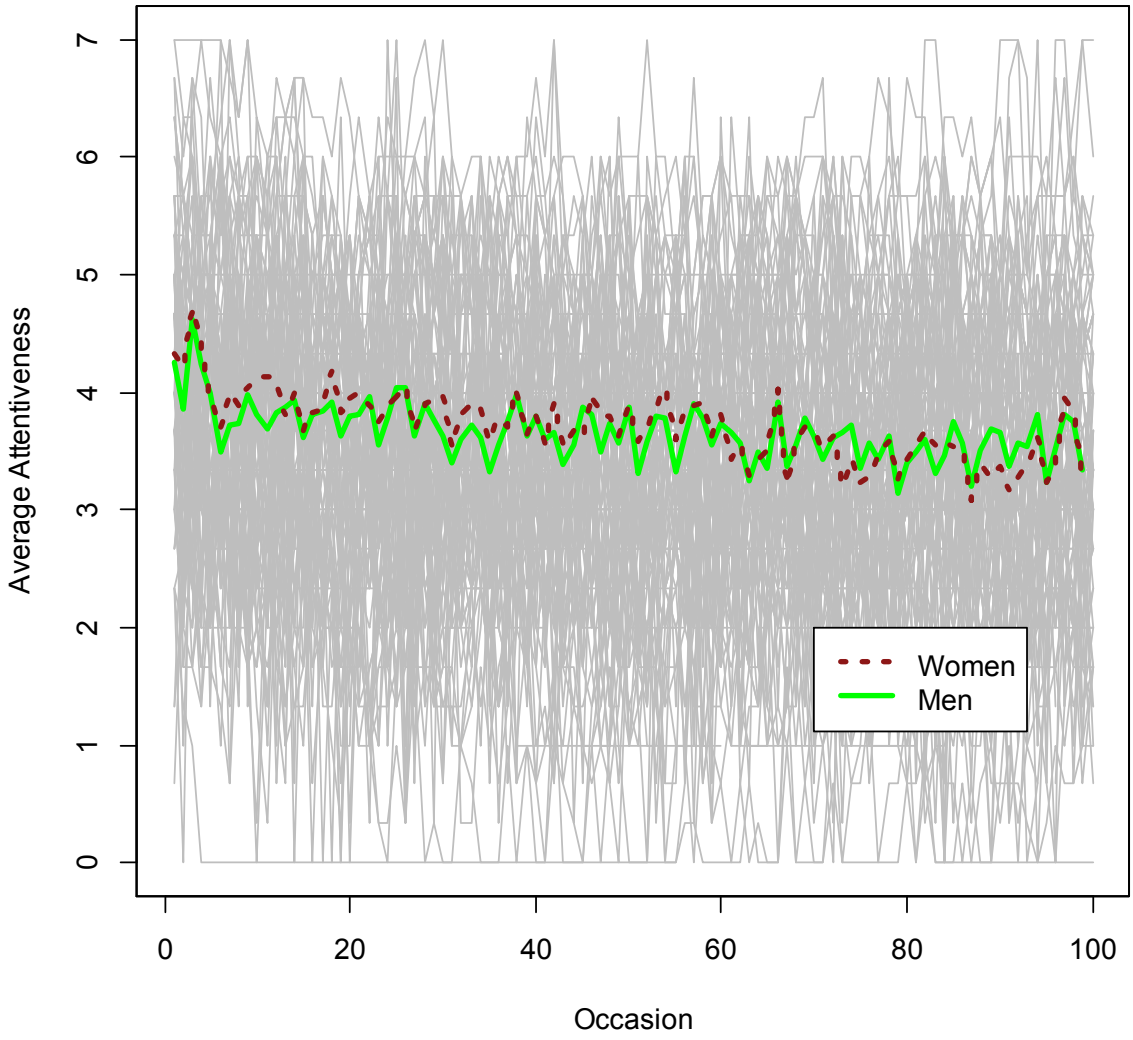
Figure 6. Two-dimensional likelihood planes resulting from tests for conditional equivalence of the BP and WP structures in attentiveness. Significant likelihood ratio values are red, nonsignificant values green (alpha 1%.). A: Controlling for mean differences. B: Controlling for mean differences and autoregression. C: Controlling for mean differences and autoregression in men. D: Controlling for mean differences and autoregression in women.

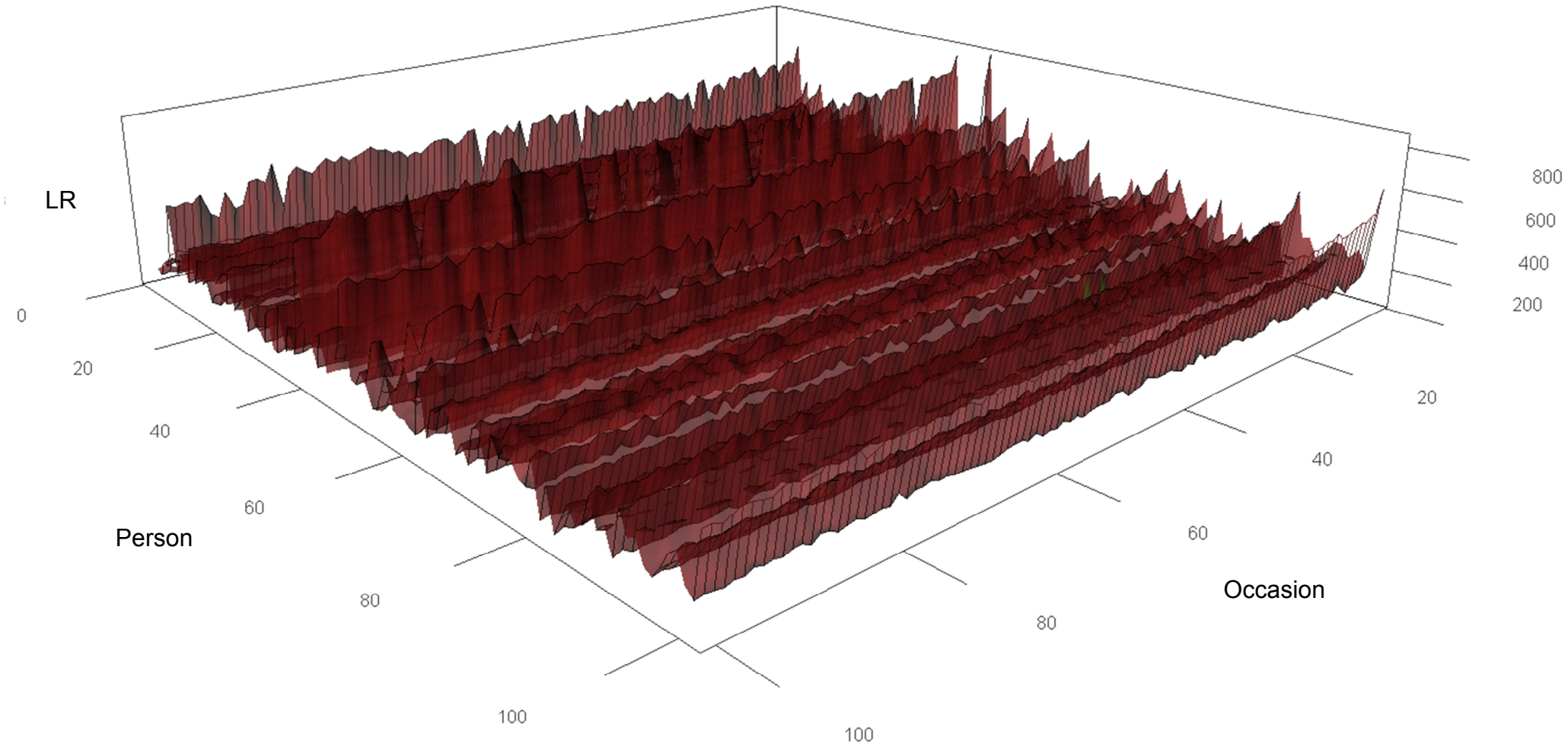
Figure 7. Chi-square distribution of two selected individuals (ID25 and ID48; both are women and marked by black arrows in Figure 6). The WP structure of ID25 is similar to the BP structure at almost all 100 occasions (the observed chi-square distribution is similar to the expected chi-square distribution under the H_0 of structural equivalence). In contrast, the WP structure of ID48 deviates clearly from the BP structure at all occasions (the distribution is shifted to the right). The individual factor structure of ID25 is illustrated in the top part of the figure.

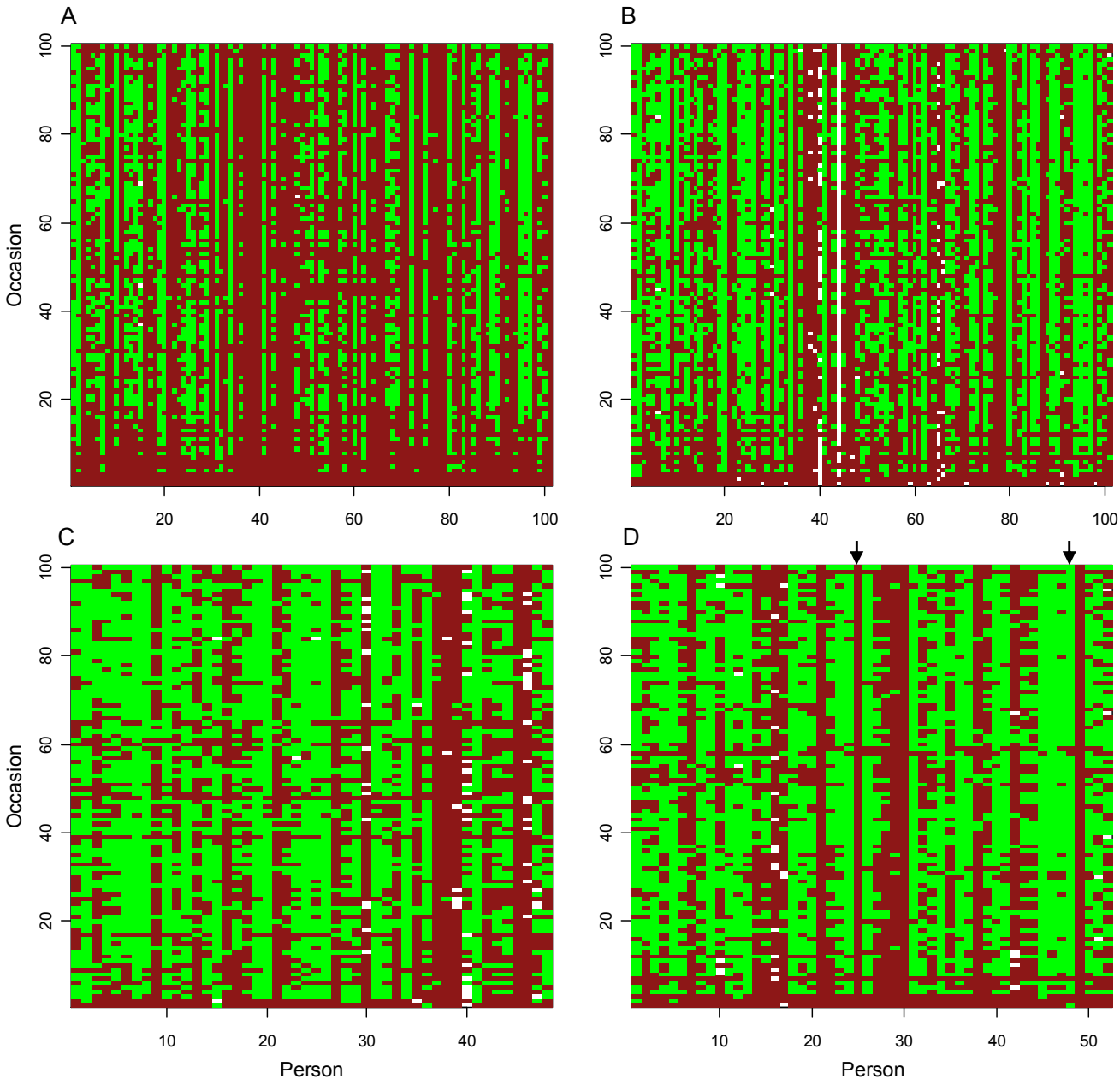

A
One occasion
One person
i = 1
N
t = 1
T
Anxiety
(1.0)
1.0
0.8
0.8
Reported nervousness
Trembling
Sweating (GSR)
ε1
ε2
ε3
B
0
C
0.8 // 0
D
Avg. variance = 1.0
Anxiety
(0.5)
Anxiety
(1.5)
Women
Men
Woman
Man

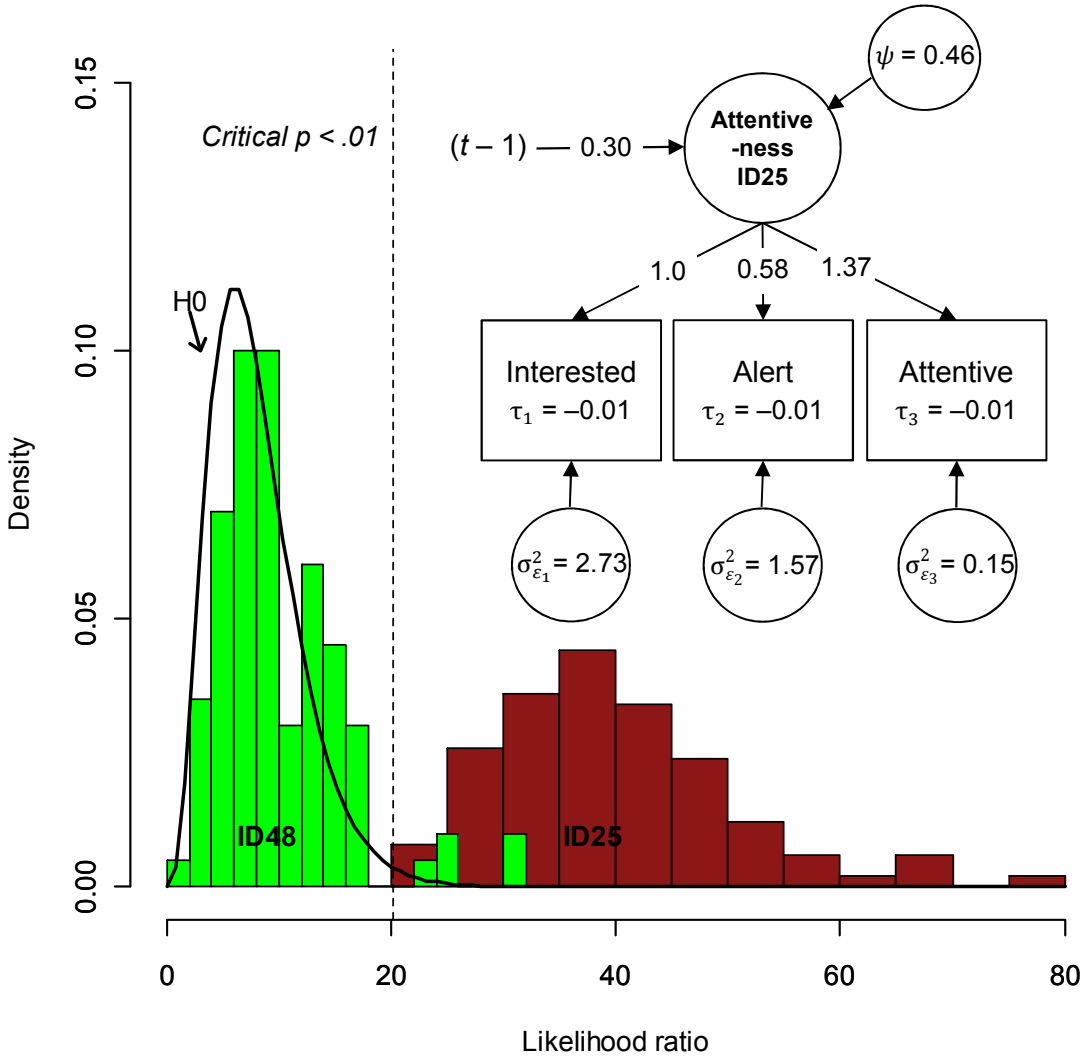

LR
0
20
40
60
80
100
Person
ID60
20
40
60
80
Occasion
5
10
15
20
25
30


A
B
C
D
Occasion
Person
20
40
60
80
100


Average Attentiveness
0
1
2
3
4
5
6
7
0
20
40
60
80
100
Occasion
Women
Men


LR
800
600
400
200
0
20
40
60
80
100
Person
20
40
60
80
100
Occasion


A
B
C
D
Occasion
Person
20
40
60
80
100
10
30
50


Critical p < .01
(t – 1) — 0.30
Attentive -ness ID25
ψ = 0.46
1.0
0.58
1.37
Interested
τ1 = –0.01
Alert
τ2 = –0.01
Attentive
τ3 = –0.01
σ²ε1 = 2.73
σ²ε2 = 1.57
σ²ε3 = 0.15
H0
ID48
ID25
Density
0.00
0.05
0.10
0.15
0
20
40
60
80
Likelihood ratio