Development and Psychopathology **22** (2010), 287–294
© Cambridge University Press, 2010
doi:10.1017/S0954579410000064

AUTHORS' RESPONSE

Statistically evaluating person-oriented principles revisited

SONYA K. STERBA AND DANIEL J. BAUER
The University of North Carolina at Chapel Hill

We are grateful for the four thought-provoking and informative Commentaries provided by expert scholars in person-oriented and idiographic research. The Commentaries of Peter Molenaar; Eun Young Mun, Marsha Bates, and Evgeny Vaschillo; Nicholas Ialongo; and Alexander von Eye offer alternative perspectives on the extent of person-oriented theories' testability using person-oriented methods in developmental psychopathology and the usefulness of this endeavor. For example, Mun et al. advocated person-oriented methods for exploratory data analysis rather than testing person-oriented principles. In contrast, Ialongo emphasized using such methods to test one principle, whereas Molenaar emphasized testing all principles, with additional suggestions for how methods reviewed can do so. Von Eye also emphasized testing all existing principles, in addition to testing several new principles. This Authors' Response focuses on comparing and contrasting these perspectives with each other and our own. It is our hope that this reply helps to move this dialog from the confines of this special section to the developmental psychopathology research community at large.

Reply to Mun, Bates, and Vaschillo

Mun, Bates, and Vaschillo (2010 [this issue]) draw attention to two overarching issues integral to the scientific operationalization of the person-oriented approach. First, they advocate methodological pluralism within the person-oriented approach. Second, they advocate exploratory data analysis as an alternative to testing person-oriented principles. We discuss each point in turn.

Mun et al. (2010) usefully remind us of situations in which variable-oriented methods can be used in service of assessing person-oriented theories. One example of this point that we find compelling is the application of idiographic methods prior to a more aggregated variable-oriented analyses (a point echoed by von Eye and Molenaar that we will return to later). Mun et al. also raise the important point that the person-oriented approach should not be thought of simply as a theoretical framework and a set of like-minded statistical methods, but should also be reflected in evaluating practical significance and conceptualizing research design. Although we agree with this point in principle, we were less convinced by some of the suggested implementations.

For implementing a person orientation in evaluating practical significance, Mun et al. (2010) suggest taking a median split on two continuous variables rather than reporting the association between the two continuous variables. This suggestion implies an overly narrow interpretation of a person-oriented approach (equating it with categorization), and is problematic for a variety of reasons (see MacCallum, Zhang, Preacher, & Rucker, 2002). For implementing a person-orientation in research design, Mun et al. suggest conceptualizing mediators and moderators, and including them in, say, variable-oriented analyses, rather than just specifying linear main effects models. This suggestion is not dissimilar from current standard practices in intervention research (e.g., Aiken & West, 1991; Baron & Kenny, 1986; Cohen, Cohen, West, & Aiken, 2003; MacKinnon, 2008). With respect to moderation in particular, we had likewise made the similar point in our Keynote Article that incorporating moderators into a latent growth curve model (LGM) moves toward a more person-centered perspective in that it allows conditional testability of the *complex-interactions principle.* With respect to mediation, although testing a causal path in and of itself does not move an analysis closer to a person-oriented perspective, certainly allowing for moderated mediation (e.g., Baron, & Kenny, 1986; Bauer, Preacher, & Gil, 2006; Edwards & Lambert, 2007; James & Brett, 1984; MacKinnon, 2008; Muller, Judd, & Yzerbyt, 2005; Preacher, Rucker, & Hayes, 2007) would do so, in allowing for testing some degree of individual differences in the causal pathway.

Mun et al. (2010) also considered whether model-based person-oriented methods should be used to test hypotheses about person-oriented theories, and raised several objections to doing so. For example, they mention instances in which testing hypotheses for model-based person-oriented methods requires a priori information. Furthermore, they mention that hypotheses about person-oriented theory could be specified and tested in a variety of ways using such methods, and they are concerned that this introduces some arbitrariness. As another example, Mun et al. state that it is impossible to test a hypothesis about person-oriented theory if our data does not allow it. As a final example, Mun

Address correspondence and reprint requests to: Sonya K. Sterba, L. L. Thurstone Psychometric Laboratory, Department of Psychology, The University of North Carolina at Chapel Hill, Campus Box 3270, Chapel Hill, NC 27599-3270; E-mail: ssterba@email.unc.edu.

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press

et al. ask how one person-oriented latent variable model can be tested if there are other equivalently well-fitting models.

Although their set of critiques is entirely valid, these critiques apply to hypothesis testing in general, not simply hypothesis testing using person-oriented methods. First, all scientists draw on a priori information from literature reviews in order to design a study and formulate a hypothesis to test, regardless of whether testing involves variable-oriented methods (e.g., linear regression) or person-oriented methods. Second, there is a certain level of arbitrariness involved in model specification and in picking which parameter to test, at which alpha level, in any analysis. For this reason, we had emphasized that we were providing examples, not an exhaustive list, of how person-oriented principles could be tested. The point of our Keynote Article was to illuminate when testing of certain principles was limited or precluded altogether by the assumptions of the method, versus where testing was possible, even in a conditional or partial form. Third, data limitations (e.g., sparse dependent variable, much missing data) always limit the extent to which we can test our theories, regardless of whether we are using person- or variable-oriented methods. If our sample size is too small to support estimation of a large complex model, we cannot test such a model, even if it were theoretically most reasonable. On this point, Cortina and Dunlap (1997) have commented "we see no point in holding NHST [null hypothesis significance testing] or any other procedure responsible for the shortcomings of the data to which they are applied" (p. 164). Fourth, with respect to the issue of equivalent models, this issue also applies to any (person- or variable-oriented) structural equation model. It is approached often by a recognition and even statistical evaluation of competing models (MacCallum, Wegener, Uchino, & Fabrigar, 1993).

After finding hypothesis testing in person-oriented methods arbitrary (but, as we have suggested, no less so than in other areas of statistics), Mun et al. (2010) instead recommend admitting that we can never know the true generating model and so ought to revert to an exploratory data analysis approach. That is, we should "ask, instead, whether data-driven, hypothesis-generating analysis has brought any new useful understanding into light." Yet how should we decide if a data-driven analysis has brought anything useful to light? Mun et al. (2010) acknowledges that "this criterion may be difficult to evaluate immediately and an affirmative answer may be warranted for some but not all such studies." Herein lies several problems. It is certainly the case that graphical exploratory analyses and discovery-oriented analyses have many important uses (for a review, see Tukey, 1977). One example is in revealing unanticipated nonlinear relationships. However, the literature review for our original Keynote Article identified many cases in which researchers were not exclusively interested in generating hypotheses, but were actually interested in evaluating existing person-oriented theory. In such cases, without hypothesis testing we lack a way to objectively determine the extent to which the data corroborate predictions of person-oriented theories, within the context of a single study (Cortina & Dunlap, 1997; Wainer, 1999). More generally, in order for person-oriented theory to be refined, or aspects of it found invalid and discarded, it must be subject to formal empirical testing (e.g., Popper, 1972). The person-oriented principles we listed in our Keynote have been widely cited for over a decade with comparatively little explicit, formal testing.

Second, without a hypothesis-testing backbone to help decide whether to accept or reject theories, we are doubtful that many psychologists can be found who *cannot* supply a cogent substantive interpretation to accompany the results of a data driven classification procedure (Mlodinow, 2008). Objective empirical testing ensures that there is a separation between theoretical and empirical evidence, which in turn, minimizes the tendency to blur boundaries between method and theory, theory and hypothesis, and hypothesis and modeling assumption in the person-oriented literature. Our Keynote suggested positing formal hypotheses when testing person-oriented theories with person-oriented methods in order to minimize these complications.

In sum, Mun et al.'s questions regarding the need for testing person-oriented theoretical principles are likely often posed by developmental psychopathology researchers. A response to such questions has not been previously articulated in detail, and so ours may be of use when analysis decisions are made in practice. With this rationale for testing person-oriented principles in place, we now turn to the topic of the remaining commentaries: how and under what conditions this can be done.

Reply to Ialongo

Ialongo (2010 [this issue]) focuses on longitudinal classification and hybrid-classification models, which have garnered a great deal of interest in recent years for offering an opportunity to parse aggregate populations. His Commentary does a nice job of reviewing the current state of the art in terms of class enumeration, model selection, and diagnostics in service of discerning whether latent subpopulations exist in models for continuous longitudinal change.

Specifically, Ialongo's Commentary (2010) reviews recommendations of Krueter and Muthén (2008), among others, regarding how to tell whether results from a hybrid-classification method—particularly growth mixture modeling (GMM)—supports the *pattern-summary principle*. Over and above having theoretical expectation of heterogeneous population subgroups, these recommendations involve four steps: (a) using model selection indices such as the Bayesian information criteria (BIC) or the Bootstrapped Likelihood Ratio Test (BLRT) to discern whether there is statistical evidence that one or more classes exist, (b) including predictors of class membership and distal outcomes of class membership to discern whether there are substantively meaningful differences among latent classes, (c) checking whether residual and outlier diagnostics are most consistent with the selected model, and (d) considering whether low sample size might be leading to underselection of classes. We discuss each step in turn, with the aid of a running illustrative simulation. In this illustrative simulation, 100 samples of $N = 500$ were generated from the conventional (i.e., *single class*)

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press

LGM described in Appendix A. This data-generating LGM features an average linear trajectory of change with continuous individual differences in intercepts and slopes, a modest nonlinear effect of a person-level predictor on intercept and slopes, a distal outcome that is predicted by intercepts and slopes, and seven modestly nonnormal repeated measures (average skew = 0.30, average kurtosis = 1.14).

Let us consider the first putative source for class validation evidence: class enumeration via model selection indices. These model selection indices are assessing whether a mixture distribution represents a better fit to the marginal distribution of the repeated measures than a unitary distribution. They are not testing whether the mixture components (classes) themselves correspond with true population subgroups, that is, the *pattern-summary principle*. For example, several classes are needed just to account for skew in repeated measures or unmodeled nonlinearity, even if a population is homogeneous (Bauer, 2007; Bauer & Curran, 2003a, 2003b). For this reason, our Keynote emphasized the need for researchers to be clear that class enumeration by model selection indices in step (a) differs from testing the pattern-summary principle directly. Furthermore, note that recent studies on the performance of popular indices for class enumeration (BIC and BLRT and/or another variant: the Lo–Mendel–Rubin LRT) have shown them to work when the models are correctly specified in all respects and the pattern summary principle is *true*, as a method for identifying the number of patterns (Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2008). That is, these studies were performed assuming that classes truly exist; the indices just had to pick the correct number. This is not the situation we are confronted with when testing the pattern-summary principle in practice. In practice, the pattern-summary principle may be false, and under this circumstance these indices can still prefer multiple classes for reasons articulated earlier.

To demonstrate this point more concretely, we fit a one-class LGM with a person-level predictor of growth (described in Appendix B) and a two-class GMM with a person-level predictor of growth and class (described in Appendix C) to our 100 simulated samples. Despite the fact that there were no discrete classes in the population from which our samples were generated, BIC preferred the two-class model in 93% of samples; the Lo–Mendel–Rubin LRT preferred the two-class model in 90% of samples, and the BLRT preferred the two-class model in 100% of samples. Again, this occurred because of the modest nonnormality of repeated measures and because of the modest unmodeled nonlinear effect of a covariate, not because there were discrete classes in the population.

Let us now consider the second putative source of class validation evidence: finding that classes predict distal outcomes or are predicted by covariates in interesting, potentially class-varying ways. Such findings do not constitute external validation that the classes correspond to true population subgroups for the following reasons. Inclusion of predictors of class membership will indeed aid in the recovery of classes, *if true subpopulations exist* (by implying greater separation between classes conditional on predictors; Lubke & Muthén, 2007). However, class membership can be significantly predicted by covariates even if there are no true subpopulations. To illustrate this phenomenon, we revisit our running example. When the two-class GMM from Appendix C was fitted to the 100 samples from our running illustration, 82% of samples showed statistically significant class-varying effects of the covariate on class membership and 100% of samples showed significant class-varying effects of the covariate on intercepts and slopes, even though no classes existed in the population. It can be seen from this example that class-varying covariate effects can just be in service of approximating conditional relationships of predictors and time in the aggregate population. They do not themselves provide evidence that classes correspond to true population subgroups.

Likewise, class-varying effects of a distal outcome on class membership can just be in service of approximating effects of growth factors on the distal outcome. To illustrate this phenomenon, we revisit our running simulation. When a distal outcome of class membership is included in the fitted two-class GMM (Appendix D), 93% of samples had statistically significant class differences in means on the distal outcome, despite the fact that no classes existed in the population. In accordance with this point, Kreuter and Muthén (2008, p. 29) themselves emphasized that "one should note that being able to predict a distal outcome from trajectory class memberships does not necessarily constitute evidence for multiple classes."

Next, we consider the third putative source of class validation evidence: checking whether residual and outlier diagnostics support the selected model. Because GMMs are more complex, flexible models than conventional LGMs (where classes = 1), they can approximate distributions of many forms to an arbitrarily close degree as the number of classes increases (Ferguson, 1983). The standardized residuals will be smaller for a GMM than a conventional LGM if one of the following occurs: (a) the GMM is overfitting the data, because residuals do not penalize for parsimony; (b) the GMM is legitimately reproducing the population distribution better, but only because the correct distribution is unknown and potentially complex, not because classes are real; or (c) the GMM is legitimately reproducing the population distribution better because classes are real. Hence, finding that the standardized residuals are smaller for a mixture model cannot be taken as evidence of true population subgroups, that is, that the pattern-summary principle holds. Because our illustrative simulation involved continuous repeated measures, comparing standardized residuals across models in terms of observed versus model implied response pattern frequencies would not be as straightforward as with Kreuter and Muthén's (2008) discrete repeated measures. In our illustrative simulation we compared example plots of Cook's Distance scores for the one-class and two-class models in Appendices B and C, for 10% of samples. In some samples, the distributions of Cook's Distance scores were similar across models, and in other samples the two-class model had several more influential cases (e.g., <10). However, it is unclear whether, in practice, we would be able to tell whether such influential cases corresponded with the need for fewer classes.

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press

The last putative source of class validation evidence, sample size considerations, may be motivated by the fact that BIC, for example, will not underestimate the true number of classes asymptotically (Leroux, 1992) but may underestimate the true number at sample sizes, for example, $<1{,}000$ or 2,000, when the model is otherwise correctly specified (e.g., Nylund et al., 2007; Tofighi & Enders, 2008). In our running illustration, modest model misspecification already resulted in class *over*extraction at $N = 500$, so we did not further increase the sample size.

In sum, the pursuit of class validation evidence from multiple sources is generally to be encouraged. However, with regard to the four specific putative sources reviewed by Ialongo (2010), it is important to keep in mind that the first source allows us to *conditionally* test the pattern-summary principle under the assumption that extracted classes correspond to true population subgroups, as described in our original Keynote's table 2. In contrast, the second through fourth sources are providing neither an unconditional nor conditional test of the pattern-summary principle, although they are certainly useful for other purposes. The second putative source in particular can be thought of as addressing the following question: if extracted classes do correspond to true population subgroups, are there substantively interesting relationships between class membership and other variables?

Reply to Molenaar

Molenaar (2010 [this issue]) provided a clear and instructive overview of the dynamic factor analysis model (DFA), beginning with background and motivation for its use in developmental psychopathology and concluding with a relevant application to patient-specific control of a disease process. In the middle sections of his article, Molenaar describes some new methodological developments involving DFA and, like Mun et al. (2010), emphasizes a bottom-up approach that begins with single-subject analyses and then proceeds to comparisons across subjects. However, unlike Mun et al., Molenaar endorses hypothesis testing as a vehicle for formally evaluating all person-oriented principles.

In particular, Molenaar (2010) agreed with our assessment that person-oriented principles of *individual specificity, pattern summary*, and *pattern parsimony*, and *holism* are testable with DFA. Regarding the remaining principles (*interindividual differences in intraindividual change* and *complex interactions*), although we labeled DFA's hypothesis testing capabilities for these principles differently than did Molenaar, at a more detailed level there is considerable agreement between our Keynote and Molenaar's Commentary. There are also a few points of divergence, and it is useful to clarify those here.

Regarding the interindividual differences in intraindividual change principle, Molenaar (2010) agreed with our assessment that intraindividual change in latent process factor means can be incorporated into the DFA (see example of DFA's structural model for mean trend in our online appendix at http://www.unc.edu/~ssterba/). Molenaar also agreed with us that intraindividual change in (co)variances of latent process factors or residuals is not readily feasible using available software (although this is a topic of his ongoing research). For the latter reason, we had labeled the interindividual differences in intraindividual change principle as only partially testable (with respect to means but not [co]variances).

However, Molenaar (2010) additionally pointed out that DFA could also allow intraindividual change in other parameters—measurement parameters (factor loadings) and other structural parameters (e.g., auto regressions of lag-0 factors on lag-1 factors)—that we did not mention (Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009). In the formulation of the DFA we presented (called the shock DFA, one of several formulations in current use; Browne & Nesselroade, 2005), there were no structural autoregressions, so such parameters could not be made time variant. However, there were factor loadings that could be made time variant. Although we are pleased with the advances in these models that broaden testability, relaxing these loadings would amount to violations of measurement invariance; that is, the meaning of the latent process factor would be time dependent, which can erode interpretability. For example, what we termed a positive emotion latent process factor under time-invariant loadings in our original article might correspond, under time-varying loadings, to an excitation factor at early time points (high loadings of enthusiastic, excited, and inspired and lower loadings of other symptoms) and to a motivation factor at later time points (high loadings of determined, attentive, and active and lower loadings of other symptoms). Consequently, it has traditionally been desirable to impose measurement invariance on factor loadings across time (as is done in related variable-oriented methods such as longitudinal factor models) and focus on whether there are interindividual differences in intraindividual change in means (testable with DFA) and variances (currently difficult to test with DFA) of latent process factors. Further discussion of this point is deferred to a later section, where it is considered as a special case of von Eye's broader concept of dimensional identity.

Regarding the complex-interactions principle, Molenaar (2010) agreed with our suggestion that pooling time series data across persons would need to be a prerequisite to investigating Person $\times$ Context interactions with DFA, and agreed with our rationale for why the single-subject DFA is not conducive to testing such interactions involving time-invariant or time-variant contextual predictors. As no procedures for implementing the former suggestion had yet been described in the DFA literature, we stopped there. Yet, Molenaar usefully described an extension designed to accommodate interactions involving *time-variant* predictors of process factors, when a DFA is fit to several persons' data. Specifically, Molenaar takes one version of a DFA (called a process DFA[1]) and ex-

1. The "process DFA" presented and expanded by Molenaar differs from the "shock DFA" (which we presented in our figure 1g and in our online equations appendix) with respect to accommodating serial dependencies. Specifically, while the process DFA allows the occasion t process factor(s) to directly influence occasion $t + 1$ process factors, the shock DFA allows the lagged occasion t measured variables to load both on the occasion $t + 1$ process factors, as well as their own occasion t process factors (for a review, see Browne & Nesselroade, 2005). That is, serial dependencies

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press

pands it to include a lagged time-varying covariate (e.g., *work stress*), which predicts process factor scores (e.g., positive and negative affect factor scores). (Work stress would be referred to as external input in Molenaar's equation 1b). For example, at measurement occasion t where an individual's work stress is high, their occasion $t + 1$ negative affect factor score may be high and their occasion $t + 1$ positive affect factor score may be low. Molenaar suggested that, after pooling across several persons' time series, it would be possible to test a Person × Context interaction by discerning whether the effect of the lagged time-varying contextual predictor (work stress) on the affect factors does or does not differ across persons. We have a minor caveat to add to this statement. Because Molenaar's procedure allows the effect of the time-varying predictor work stress to vary across *persons*, rather than across *fixed levels* of a person-level predictor (e.g., boys vs. girls, inpatients vs. outpatients), we view it as more similar to allowing a "random slope" for work stress (individual specificity principle) than a conventional Person × Context interaction (complex interactions principle) involving work stress. However, suppose we were to slightly modify Molenaar's procedure in order to, after pooling across multiple persons' time series, equate the slopes of work stress for all persons at a given level of a grouping variable (e.g., equate the slopes among boys and equate slopes among girls). We would be then be able to test a complex interaction directly (i.e., gender × work stress in predicting affect).

In sum we believe Molenaar has nicely illustrated many of the advantages of idiographic methods for evaluating person-centered theory. Although we believe there are still some existing limitations in this approach, we also regard these methods as offering the most promise for future development.

Reply to von Eye

Von Eye (2010 [this issue]) provided an informative review of person-oriented theory, as well as its links to differential psychology. He also provided a relevant empirical example of how person-oriented theoretical principles could be tested using a nonparametric method not included in our original review—configural frequency analysis (e.g., von Eye, 2002). Like Molenaar, von Eye endorsed the objective of testing all person-oriented principles, but he also proposed reformulating one principle before testing it (*pattern parsimony*) and advocated adding a new principle to be tested (*dimensional identity*).[2] We discuss each proposed change in turn.

Von Eye proposes reformulating the pattern parsimony principle in the following manner. Instead of asking whether individual differences can be summarized with a small number of patterns of variables, researchers ask whether the *frequency* of particular patterns is higher or lower than would be expected by chance. The idea of this reformulated principle is interesting. To the extent that holistic patterns of behavior exist and characterize distinct groups of individuals, we should be able to observe that certain combinations of attributes occur more (or less) frequently than would be expected if these combinations were purely random. However, there is a trade-off involved: patterns are redefined as relative rather than absolute entities. To test this reformulated principle, chance frequencies are dictated by a baseline model (often a main effects model). When the frequency of a pattern is higher than would be expected, given a baseline model, it is called a type. When the frequency of a pattern is lower than would be expected given a baseline model it is called an antitype. Hence, instead of testing whether the best-fitting number of patterns is less than some predefined small number, researchers would test whether the expected pattern frequencies (i.e., expected cell counts within cross-tabulations of categorical variables) differ from observed pattern frequencies (i.e., observed cell counts within cross-tabulations of categorical variables). Whereas the original pattern summary principle would require counting patterns that are prevalent in absolute terms (e.g., 30% of persons), the reformulated principle would ignore such patterns unless they are more frequently or less frequently observed than expected under the baseline model.

One implication of this reformulation is that it tailors the definition of the pattern-parsimony principle to be testable using the output provided by techniques, such as configural frequency analysis (von Eye, 1990, 2002), that analyze contingency tables of a limited number of categorical variables. At the same time, another implication of this reformulation is that it renders the definition of the pattern parsimony principle untestable using methods reviewed by Ialongo (e.g., growth mixture models), and using variables that are continuous (unless they are first discretized; but see MacCallum et al., 2002). Researchers need to decide whether they theoretically conceptualize patterns as absolute or relative entities, so they can choose an appropriate method that matches their definition.

Von Eye also endorses a concept from von Eye and Bergman (2003), *dimensional identity*, as another principle to be tested. von Eye urges us to address this concept with respect to the model-based person-oriented methods in our original review. We find that sometimes this concept is described as a methodological assumption (e.g., in von Eye's configural frequency analysis example it is assumed for subsets of individuals) and other times is described as a research hypothesis. To clarify this matter, we first address when dimensional identity is a methodological assumption of model-based person-oriented methods and later briefly address how and to what extent it is testable using such methods.

Dimensional identity broadly equates to the assumption of invariance for certain parameters, across persons and/or time. It can alternatively be thought of as the assumption of statistical exchangeability of units (persons and/or time points). If every person and event is absolutely unique, a statistical model cannot be fit; an inferential analysis is impossible because there can be no generalization to a broader population of persons, or time points within person. The trick is to decide where the imposition of this

are either accounted for in the structural model portion (in the process DFA) or in the measurement model portion (in the shock DFA).

2. von Eye briefly mentioned other potential new principles featured in Bogat (2009) that were geared toward community psychology, but did not focus on these.

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press

assumption is most justified. Invariance, or dimensional identity, is imposed across all persons in the population in latent growth curve modeling (for number of factors and typically factor loadings, but not factor scores). Invariance is imposed across all persons in the same class in latent class growth modeling (for factor scores, factor loadings, number of factors, but not posterior probabilities of class membership). Invariance is imposed across all persons in the same class in growth mixture modeling (for factor loadings, number of factors, but not factor scores or posterior probabilities of class membership). In addition, these methods all implicitly assume that the nature of the repeated measure retains the same interpretation across time for all persons. For the latent Markov model, dimensional identity is required across persons within sequence for transition probabilities, initial latent status probabilities, and conditional response probabilities (as well as across time within sequence for at least some conditional response probabilities). The same is required within sequences per latent chain in the mixed latent Markov model.[3] Finally, in single-subject dynamic factor analysis models, invariance is only imposed across time (for factor variances and covariances at least but typically also factor loadings), not persons. Hence, models in our original review differ markedly in the parameters for which they impose dimensional identity, or invariance, and for whom. The degree of dimensional identity imposed by models in our review directly relates to our Keynote's discussion of these models' risk for ecological versus atomistic fallacies. Specifically, as greater dimensional identity is imposed by a model, risk of ecological fallacy increases if results are used to generalize to individuals. As less dimensional identity is imposed, risk of atomistic fallacy only increases if results are used to generalize to the population at large.

Dimensional identity, or parameter invariance across say persons or time, can only be assessed to a limited extent within a given model. To investigate, it would require one of two approaches. The first approach involves comparing the fit of alternative models for n persons' data, which impose more lenient versus more strict dimensional identity. (Restrictive versions of variable-oriented models are at the latter extreme.) The second approach involves building up from an idiographic model (single-subject approach) to a more aggregated model for $m < n$ persons, who exhibit parameter invariance across key dimensions, such as time and/or persons. The first approach was endorsed by Ialongo and the second approach was endorsed by von Eye, Mun et al., and Molenaar.

In sum, von Eye reviewed how additional nonparametric methods could be used to test person-oriented principles and urged an expansion of the theoretical scope of our original Keynote Article through consideration of revised principles and new principles. This expansion should be particularly helpful for researchers who are acquainted with these other aspects of person-oriented theory and methods and who are looking for a bridge from both to the content of our original Keynote.

Conclusions

It is useful in concluding to step back and consider why this exchange began and where we are now. The theoretical principles and methods labeled person-oriented have over the past decade garnered increasing interest in developmental psychopathology. The hypothesis testing recommendations in our original Keynote Article were a reaction to the person-oriented research community's admission of two problems in current practice: the lines between theory and method had been blurred, and the theory had often been reduced to methods for categorization.

The Commentaries from Mun et al., Ialongo, Molenaar, and von Eye expressed a range of perspectives regarding how many and which person-oriented principles can and should be tested. Their Commentaries stimulated a more thorough discussion of key issues of practical relevance in person-oriented research, including foundational rationales for and against testing person-oriented theoretical principles, how existing resources can be brought to bear on statistically testing certain principles using growth mixture models, how new methodological developments can be used to expand the testability of principles in dynamic factor models, and what the substantive and methodological implications are of reformulating or adopting new person-oriented theoretical principles.

The Commentaries illustrated that the arsenal of methods deemed person-oriented is continually evolving, as is the canon of person-oriented theory. Ideally, these methodological advances help us enrich theory without theory becoming restricted or bent toward a particular method. Yet, the onus is on researchers to be vigilant in keeping the two philosophically separated enough when evaluating person-oriented theory so that problems associated with blurring their boundary are avoided. We continue to maintain that hypothesis testing is one vehicle for achieving this separation; articulating hypotheses and testing them prevents researchers from mistakenly treating model assumptions as confirmation of theory. Further, testing multiple person-oriented principles rather than just the pattern-summary principle prevents researchers from reducing person-oriented theory to a matter of categorizing people or not. We believe that finding creative ways to test relevant person-oriented theoretical principles using available methods, and understanding the limitations of these tests, is an important task of person-oriented research.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality & Social Psychology*, *51*, 1173–1182.

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, *24*, 757–786.

3. Additional restrictions to the mixed latent Markov model are necessary for identification and/or estimability depending on the number of time points, number of chains, and number of sequences per chain.

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press

Bauer, D. J., & Curran, P. J. (2003a). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*, 338–363.
Bauer, D. J., & Curran, P. J. (2003b). Over-extracting latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthén (2003), and Cudeck and Henly (2003). *Psychological Methods*, *8*, 384–393.
Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*, 142–163.
Bogat, G. A. (2009). Is it necessary to discuss person-oriented research in community psychology? *American Journal of Community Psychology*, *43*, 22–34.
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-Olivares & J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 415–452). Mahwah, NJ: Erlbaum.
Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161–172.
Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*, 1–22.
Ferguson, T. S. (1983). Bayesian density estimation via mixtures of normal distributions. In M. H. Rizvi, J. S. Rustagi, & D. Siegmund (Eds.), *Recent advances in statistics* (pp. 287–302). New York: Academic Press.
Ialongo, N. (2010). Steps substantive researchers can take to build a scientifically strong case for the existence of trajectory groups. *Development and Psychophathology*, *22*, 273–275.
James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, *69*, 307–321.
Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology*, *24*, 1–31.
Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics*, *20*, 1350–1360.
Lubke, G., & Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling*, *14*, 26–47.
MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*, 185–199.
MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.
MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Erlbaum.
Mlodinow, L. (2008). *The Drunkard's Walk: How randomness rules our lives* (pp. 169–191). New York: Random House.
Molenaar, P. C. M. (2010). Testing all six person-oriented principles in dynamic factor analysis. *Development and Psychopathology*, *22*, 255–259.
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*, 260–271.
Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, *89*, 852–863.
Mun, E. Y., Bates, M. A., & Vaschillo, E. (2010). Closing the gap between person-oriented theory and methods. *Development and Psychopathology*, *22*, 261–271.
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte-Carlo simulation study. *Structural Equation Modeling*, *14*, 535–569.
Popper, K. (1972). *Objective knowledge: An evolutionary approach*. Oxford: Clarendon Press.
Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.
Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison–Wesley.
von Eye, A. (2010). Developing the person-oriented approach: Theory and methods of analysis. *Development and Psychopathology*, *22*, 277–285.
von Eye, A. & Bergman, L. R. (2003). Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. *Development and Psychopathology*, *15*, 553–580.
von Eye, A. (1990). *Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classification*. New York: Cambridge University Press.
von Eye, A. (2002). *Configural frequency analysis—Methods, models, and applications*. Mahwah, NJ: Erlbaum.
Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212–213.

Appendix A

Data generating model: single-class LGM with covariate and distal outcome

$$
\begin{aligned}
y_{ti} &= \eta_{0i} + \eta_{1i}\lambda_t + \varepsilon_{ti},\\
\eta_{0i} &= -1 + 0.3x_i + 0.15x_i^2 + \varsigma_{0i},\\
\eta_{1i} &= 0.5 - 0.3x_i + 0.15x_i^2 + \varsigma_{1i},\\
d_i &= 3\eta_{0i} - 0.3\eta_{1i} + \varsigma_{2i},
\end{aligned}
$$

where

$$
\boldsymbol{\varsigma}_i \sim N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}\begin{bmatrix}0.5 & & \\ -0.1 & 0.4 & \\ 0 & 0 & 1\end{bmatrix}\right)
$$

and i is the individual, t is the time, d_i is the distal outcome, x_i is the person-level covariate, λ_t is the time score, η_{0i} and η_{1i} are individual-specific intercepts and slopes, ε_{ti} is the time-specific residual, and $\boldsymbol{\varsigma}_i$ are the person-specific residuals. The nonnormality of ε_{ti} is manipulated to produce modestly nonnormal repeated measures (average skew = 0.30, average kurtosis = 1.14).

Appendix B

Fitted single-class linear LGM with covariate

$$
\begin{aligned}
y_{ti} &= \eta_{0i} + \eta_{1i}\lambda_t + \varepsilon_{ti},\\
\eta_{0i} &= \alpha_{00} + \gamma_{01}x_i + \varsigma_{0i},\\
\eta_{1i} &= \alpha_{10} + \gamma_{11}x_i + \varsigma_{1i},
\end{aligned}
$$

where $\boldsymbol{\varsigma}_i \sim N(0, \boldsymbol{\Phi})$, $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Theta})$, α_{00} is the mean intercept where $x_i = 0$, and α_{10} is the mean slope where $x_i = 0$.

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press

Appendix C

Fitted two-class linear GMM with covariate

$$
\begin{aligned}
y_{tik} &= \eta_{0ik} + \eta_{1ik}\lambda_t + \varepsilon_{tik},\\
\eta_{0ik} &= \alpha_{00k} + \gamma_{01k}x_i + \varsigma_{0ik},\\
\eta_{1ik} &= \alpha_{10k} + \gamma_{11k}x_i + \varsigma_{1ik},\\
P(c_{ik} = 1|x_i) &= \frac{\exp(\nu_k + \delta_{1k}x_i)}{\sum_{k=1}^{2}\exp(\nu_k + \delta_{1k}x_i)},
\end{aligned}
$$

where $\boldsymbol{\varsigma}_{ik} \sim N(0, \boldsymbol{\Phi})$ and $\boldsymbol{\varepsilon}_{ik} \sim N(0, \boldsymbol{\Theta})$. The latent classification variable $\mathbf{c}_i = (c_{i1}, c_{i2})'$ has values $k = 1, 2$, and $c_{ik} = 1$ if individual i is in class k.

Appendix D

Fitted two-class linear GMM with covariate and distal outcome

$$
\begin{aligned}
y_{tik} &= \eta_{0ik} + \eta_{1ik}\lambda_t + \varepsilon_{tik},\\
\eta_{0ik} &= \alpha_{00k} + \gamma_{01k}x_i + \varsigma_{0ik},\\
\eta_{1ik} &= \alpha_{10k} + \gamma_{11k}x_i + \varsigma_{1ik},\\
P(c_{ik} = 1|x_i) &= \frac{\exp(\nu_k + \delta_{1k}x_i)}{\sum_{k=1}^{2}\exp(\nu_k + \delta_{1k}x_i)},\\
d_{ik} &= \beta_1 c_{i1} + \beta_2 c_{i2} + \varsigma_{2ik},
\end{aligned}
$$

where $\boldsymbol{\varsigma}_{ik} \sim N(0, \boldsymbol{\Phi})$ and $\boldsymbol{\varepsilon}_{ik} \sim N(0, \boldsymbol{\Theta})$. See the authors' online appendix (http://www.unc.edu/∼ssterba/) for details on the notation. Note that in the online appendix, $P(c_{ik} = 1|x_i) = (\pi_k)_i$.

https://doi.org/10.1017/S0954579410000064 Published online by Cambridge University Press